LREC 2026

Fourth Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2026) @ LREC 2026

Workshop Proceedings

Editors
Marco Passarotti and Rachele Sprugnoli

11 May 2026

Proceedings of the Fourth Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2026) @ LREC 2026

©ELRA Language Resources Association (ELRA), 2026
These proceedings are licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

ISBN 978-2-493814-58-6

Preface

These proceedings include the papers accepted for presentation at the Fourth Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2026).[1] The workshop was held on May 11 2026 in Palma, Mallorca (Spain), co-located with LREC 2026.[2]

The workshop wants to provide a venue to discuss research works on a wide range of topics concerning the building, analysis, exploitation and distribution of collections of digitized texts written in historical and ancient languages as well as of their lexica, with a specific focus on the development and application of Language Technologies (LTs) for such purposes. The topics of the workshop are strictly bound to the peculiar characteristics of textual and lexical data for historical and ancient languages, which set them apart from modern languages, with a significant impact on the use and development of LTs for their processing and study. The workshop welcomed contributions addressing a broad range of topics. These included the handling of spelling variation and the detection and correction of OCR errors, as well as the creation and annotation of linguistic resources, both lexical and textual. Further areas of interest were deciphering, and the morphological, syntactic, and semantic analysis of textual data, alongside the adaptation of existing tools to account for diachronic, diatopic, and diastratic variation. The workshop also encouraged work on the use of LTs for teaching ancient languages, NLP-driven theoretical studies in historical linguistics, and the computational analysis of ancient literary texts. Additional relevant themes included the evaluation of NLP tools specifically designed for historical and ancient languages, the application of Large Language Models to the automatic analysis of ancient texts, and the role of digital infrastructures (such as CLARIN) in supporting research on language resources for historical and ancient languages.

Overall, the languages tackled in the papers published in these proceedings are almost 30. This is not an exhaustive list but it gives an idea of the linguistic richness covered: Ancient Chinese, Ancient Greek, Ancient Hebrew, Arabic, Coptic, Czech, Dutch, Estonian, French, German, Icelandic, Italian, Japanese, Judeo-Arabic, Latin, Maltese, Medieval Greek, Old English, Old High German, Old Norse, Old Saxon, Phoenician, Rapa Nui, Serbian, Swedish.

Like for LREC 2026, the submission process was double-blind. Each paper was reviewed but three independent reviewers from a program committee made of 40 scholars from 17 countries. In total, we received 39 submissions (against the 32 of the previous edition). After the reviewing process, we accepted 29 submissions, leading to an acceptance rate of 74.3%.

LT4HALA 2026 was also the venue of the fourth edition of EvaLatin, the campaign devoted to the evaluation of NLP tools for Latin.[3] EvaLatin was started in 2020 (co-located with the first edition of LT4HALA) considering the important role played by textual data and linguistic metadata in the study of historical and ancient languages, with a special focus on Latin due to its prominence among such languages, both for the size and for the degree of diversity of its texts. Running evaluation campaigns in such a scenario is essential to understand the level of accuracy of the NLP tools used to build and analyze resources featuring texts that show those peculiar characteristics mentioned above. The fourth edition of EvaLatin focused on two shared tasks, i.e. Dependency Parsing, and Named Entities Recognition and Classification. Organizers and participants of EvaLatin submitted a report to be published in this proceedings thus this volume includes two overviews (one for each EvaLatin task) and 6 reports from participants.

Besides EvaLatin, LT4HALA 2024 hosted also a new edition of EvaHan, the evaluation campaign of NLP tools for the Ancient Chinese language, organized by a team of scholars

[1] https://circse.github.io/LT4HALA/2026/

[2] https://lrec2026.info

[3] https://circse.github.io/LT4HALA/2026/EvaLatin

directed by Bin Li (Nanjing Normal University).[4] The new edition of EvaHan focused on OCR tasks for multimodal large language models in ancient Chinese. Like for EvaLatin, the participants of EvaHan were required to submit a short technical report which received a light review by the organizers. Overall, these proceedings include an overview of the EvaHan campaign (authored by the organizers) and 12 technical reports, corresponding to as many participants. We wish to thank the organizers of EvaHan, whose contribution broadened the scope of historical and ancient languages represented at the LT4HALA 2026 workshop and highlighted how certain NLP challenges are inherent to ancient and historical languages as such, regardless of their typological differences.

LT4HALA 2026 also included EvaCun, an evaluation campaign on cuneiform LTs organized by Adam Anderson and Farhan Afzal Abdulla (UC Berkeley, USA).[5] The proposed shared tasks focused on transliteration normalization, morphological analysis and lemmatization, Named Entity Recognition. Unfortunately, this edition of EvaCun attracted no participants, and the proceedings therefore contain no contributions on this topic. We nonetheless remain hopeful: the computational study of cuneiform languages represents one of the most fascinating and open challenges in the field of ancient language processing, and we trust that future editions will succeed in bringing together the researchers and resources needed to rise to it.

Now in its fourth edition, LT4HALA continues to grow steadily, both in the number of participants and in the range and diversity of languages and topics addressed by scholarly contributions. We welcome this expansion as an encouraging sign for the field, though perhaps not a surprising one, since the study of ancient and historical languages has always been deeply rooted in the analysis of textual evidence. Processing the large digital collections of such texts through the most advanced LTs, with the aim of bringing computational analysis to bear on them, holds genuine promise for advancing a century-long scholarly tradition. LT4HALA aspires to be a venue that supports and accompanies this computational turn.

Marco Passarotti and Rachele Sprugnoli

[4] https://circse.github.io/LT4HALA/2026/EvaHan
[5] https://circse.github.io/LT4HALA/2026/EvaCun

Organizing Committee

Workshop Organizers:

- Marco Passarotti, Università Cattolica del Sacro Cuore di Milano (Italy)
- Rachele Sprugnoli, Università Cattolica del Sacro Cuore di Milano (Italy)

Program Committee:

- Alaa Mamdouh Akef, Peking University, China
- Adam Anderson, UC Berkeley, USA
- Masayuki Asahara, National Institute for Japanese Language and Linguistics, Japan
- Monica Berti, University of Leipzig, Germany
- Federico Boschetti, Università Ca' Foscari, Istituto di Linguistica Computazionale - CNR & CLARIN Knowledge Centre for Digital and Public Textual Scholarship (DiPText-KC), Italy
- Luca Brigada Villa, University of Bergamo and University of Pavia, Italy
- Patrick J. Burns, Institute for the Study of the Ancient World / NYU, USA
- Flavio Massimiliano Cecchini, KU Leuven, Belgium
- Thibault Clérice, Inria Paris, France
- Claudia Corbetta, Università di Bergamo and Università di Pavia, Italy
- Angelo Mario Del Grosso, Istituto di Linguistica Computazionale - CNR & CLARIN Knowledge Centre for Digital and Public Textual Scholarship (DiPText-KC), Italy
- Mathieu Dehouck, Laboratoire Lattice - CNRS, France
- Margherita Fantoli, KU Leuven, Belgium
- Andrea Farina, King's College London, UK
- Federica Gamba, Charles University, Czech Republic
- Shai Gordin, Ariel University, Israel
- Marisa Hudspeth, University of Massachusetts Amherst, USA
- Alek Keersmaekers, KU Leuven, Belgium
- Federica Iurescia, Università Cattolica del Sacro Cuore di Milano, Italy
- Bin Li, Nanjing Normal University, China
- Eleonora Litta, Università Cattolica del Sacro Cuore di Milano, Italy
- Yudong Liu, Western Washington University, USA
- Francesco Mambrini, Università Cattolica del Sacro Cuore di Milano, Italy
- Barbara McGillivray, King's College London, United Kingdom
- Domenico Giuseppe Muscianisi, Università di Parma, Italy
- Vivi Nastase, University of Geneva/Idiap Research Institute, Martigny, Switzerland
- Carlos Javier Nusch, Universidad Nacional de La Plata, Argentina
- Teresa Paccosi, Royal Netherlands Academy of Arts & Sciences (KNAW), The Netherlands
- Chiara Palladino, Durham University, United Kingdom

- Alessio Palmero Aprosio, Università di Trento, Italy
- Matteo Pellegrini, University of Surrey, UK
- Eva Pettersson, Uppsala University & CLARIN Knowledge Centre on Diachronic Language Resources (DiaRes), Sweden
- Sophie Prévost, Laboratoire Lattice - CNRS, France
- Matteo Romanello, Swiss Art Research Infrastructure (SARI), University of Zurich, Switzerland
- Colin Swaelens, Ghent University, Belgium
- Muhammad Shoaib Tahir, COMSATS University Islamabad, Lahore Campus, Pakistan
- James K. Tauber, FAU Erlangen, Germany
- Tatiana Tommasi, Università Ca' Foscari, Venezia, Italy
- Tariq Yousef, University of Southern Denmark, Denmark
- Chiara Zanchi, Università di Pavia, Italy

EvaLatin 2026 Organizers:

Dependency Parsing:

- Federica Iurescia, Università Cattolica del Sacro Cuore di Milano, Italy
- Rachele Sprugnoli, Università Cattolica del Sacro Cuore di Milano, Italy
- Marco Passarotti, Università Cattolica del Sacro Cuore di Milano, Italy

Named Entity Recognition:

- Valeria Irene Boano, KU Leuven, Belgium
- Eleonora Litta, Università Cattolica del Sacro Cuore di Milano, Italy
- Matteo Romanello, Swiss Art Research Infrastructure (SARI), University of Zurich, Switzerland

EvaHan 2026 Organizers:

- Bin Li, Nanjing Normal University, P.R. China
- Bolin Chang, Nanjing Normal University, P.R. China
- Minxuan Feng, Nanjing Normal University, P.R. China
- Chao Xu, Nanjing Normal University, P.R. China
- Dongbo Wang, Nanjing Agricultural University, P.R. China

EvaCun 2026 Organizers:

- Adam Anderson, UC Berkeley, USA
- Farhan Afzal Abdulla, UC Berkeley, USA

Table of Contents

Morphological Annotation of Old Serbian in Universal Dependencies
Vladimir Polomac and Silvie Cinkova 1

Tracing Morph Origins in Czech: A Computational Approach to Morph-Level Etymology
Aleš Manuel Manuel Papáček and Zdeněk Žabokrtský 7

Uncovering Work from Words: LLM-Based Information Extraction from Historical Petitions
Ellinor Lindqvist, Eva Pettersson and Joakim Nivre 19

Extracting Volcanological Knowledge from Historical Texts: A Language-Technology Pipeline for Diachronic Geovisualization
Costanza Marini, Gianluca Casagrande, Alessio Palmero Aprosio and Claudia Principe 38

When Lexicographic Quotations Become a Corpus: To Deduplicate or Not to Deduplicate?
Manuel Favaro, Elisa Guadagnini, Eva Sassolini, Marco Biffi and Simonetta Montemagni 49

A New State-of-the-Art BERT Model for Judeo-Arabic
Elisha Rosensweig, Yitzchak Lindenbaum, Hillel Gershuni, Vered Raziel-Kretzmer, Daniel Caine and Avi Shmidman 58

BEReshiT: an Ancient Hebrew Model based on DictaBERT
Iglika Nikolova-Stoupak, Maxime Amblard and Frédérique Rey 72

Automatic Detection of Metaphorical Expressions in Classical Japanese Using WLSP-Enhanced BERT
Hang Zhu, Mitoki Ohara, Rei Kikuchi, Kanako Komiya, Masayuki Asahara and Sachi Kato 89

Domain-Aware Error Correction for Citation NER in Medieval Hebrew Responsa
Shmuel Llebeskind, Maayan Zhitomirsky-Geffet, Binyamin Katzoff, Nati Ben-Gigi and Jonathan Schler 96

From Lemmas to Links: A Lemma Bank for Ancient Greek
Colin Swaelens, Francesco Mambrini and Marco Passarotti 106

Across Generations: A Comparative Analysis of NER for Latin Inscriptions from Classical Machine Learning to LLMs
Wenhui Cui and Phillip Benjamin Ströbel 112

POS Tagging with Generative LLMs for Historical Germanic Low-Resource Languages: An Evaluation Against Fine-Tuned BERT
Irene Miani, Gregory Darwin and Sara Stymne 125

From Manuscript to Model: Developing HTR for Medieval Greek
Nicklas Sindlev Andersen, Byron MacDougall, Tariq Yousef and Aglae Pizzone 139

I, RE:Claudius 256: Towards Linking Classical Latin Person Mentions to a Domain-specific Knowledge Base
Marijke Beersmans, Evelien de Graaf, Julie Nijs, Valeria Irene Boano, Alek Keersmaekers, Mark Depauw, Tim Van de Cruys and Margherita Fantoli 152

Capturing Ancient Chinese Sense Induction with Automatic Pipelines
Guan-Yu Tseng, Chunki Lim, Chih-Han Lin, Tung-Le Pan, Yu-Chieh Wang, Lang-Ching Yeh and Shu-Kai Hsieh 163

A Computational Evaluation of Syllabic Hypotheses for Rongorongo: Evidence from N-gram Analysis
Evgeniya Korovina 177

Building a Corpus and Database for Rare and Undeciphered Scripts
Beata Megyesi, Rune Rattenborg, Benedek Láng, Michelle Waldispühl and Mihály Héder 184

A Layered Annotation Workflow for Semitic Epigraphy
Tal Bernstein, Shai Gordin and Letizia Cerqueglini 197

Overview of the Dependency Parsing Task at EvaLatin 2026
Federica Iurescia, Marco Passarotti and Rachele Sprugnoli 211

THIVLVC: Retrieval Augmented Dependency Parsing for Latin
Luc Pommeret, Thibault Wagret and Jules Deret 219

Overview of the Named Entity Recognition Task at EvaLatin 2026
Valeria Irene Boano, Eleonora Litta and Matteo Romanello 226

Transfer Learning for Named Entity Recognition of Classical Latin through LLM Prompting
Callum Chan 234

Overview of EvaHan2026: The First International Evaluation of Ancient Chinese OCR and Layout Analysis
Dongbo Wang, Dongmei Zhu, Jieqiong Li, Chang Liu, Ruifeng Wu, Fan Yang, Zhu Yue, Xin Gao, Zhao Zhixiao, Xue Zhao, Zhongzheng Wu, Liu Liu and Li Bin 244

A Multi-Stage System for Ancient Chinese OCR and Layout Understanding in the EvaHan2026 Shared Task
KeYan Liang and Meiling Liu 263

A Multi-Modal Recognition Framework for Ancient Books Integrating DoRA-DPO Text Recognition and YOLO Layout Analysis
Chaokun Zhang, Xin Wen and Tongtong Zhou 268

Beijing Normal University at EvaHan 2026: Enhancing Ancient Chinese Character Recognition and Layout Analysis via VLM Fine-Tuning and Linguistic Post-Processing
Yihuan Yin and Qian Zhao 273

A Dual-Modality Framework for Ancient Document Layout Analysis and Text Recognition
Qi Fan, Jieming Hu and Chen Ye 277

EvaHan 2026 Ancient Books Multimodal OCR and Layout Analysis System Technical Report
Chenrui Zheng 288

A Parameter-Efficient and Data-Centric Framework for Ancient Chinese Text Recognition and Layout Analysis
Yuchun Meng 294

LVLM Optimization for Ancient Chinese Book Image Analysis with Task-specific Augmentation and Instruction Tuning
Tian Xia, Yulong Liu, Yilin Wang, Yumeng Yang, Dongheng Cai, Yuyang Tan and Menghui Yang 299

Data-Centric Strategies for Ancient Chinese Text Recognition: Augmentation, Annotation Refinement, and Style Transfer in EvaHan 2026
Li Chengfei, Zhang Yunjie, Li XiaoYi, Quan Changshun, Cao Taihe and Liu Bin 305

AnandaSky: A Vision–Language Model for Line-Level Transcription of Historical Sinographic Documents
Colin Brisson, Ayoub Kahfy, Frédéric Constant and Marc Bui 311

Multimodal Ancient Document Parsing: Technical Report for EvaHan2026 Competition
Liqi He, Qiwei Li, Ziye Yang and Zuchao Li 322

Multi-Task Learning Trade-offs in Vision–Language Models for Ancient Chinese OCR: An Empirical Analysis of Parameter-Efficient Adaptation
Yuhan Shu and Huizi Zhou 330

Building Character(s): Synthetic Data and In-Context Learning Strategies for Few-Shot Ancient Chinese Recognition
Denise Atzori, Marie Bizais-Lillig, Mathias Garnier, Maxime Létoffé, Charles Planque, Tianjie Yin and Chahan Vidal-Gorène 339

The UD_Latin-PROIEL as Linked Open Data: Integrating a Latin Treebank into the LiLa Knowledge Base
Lucas Consolin Dezotti, Marco Passarotti, Federica Iurescia and Giovanni Moretti 353

Language Models for the Restoration of Latin Legal Manuscripts
Shibingfeng Zhang, Edoardo Caraffa, Annafelicia Zuffrano, Maddalena Modesti and Giovanni Colavizza 361

Evaluating Hierarchical Aggregation and LLM-Based Matching for Synset Selection in Ancient Greek
Luca Brigada Villa, Marco Passarotti, Chiara Zanchi, Riccardo Ginevra, Erica Fratellini and Eleonora Litta 368

Miktub: A Manuscript Dataset of Historical Maltese for Handwritten Text Recognition
Thomas Koppens and Claudia Borg 380

Smelling the Past: Investigating Historical Models for Olfactory Event Extraction
Teresa Paccosi and Marijn Koolen 389

Contemporizing 20-th Century Estonian
Heiki-Jaan Kaalep 400

Cost-Aware Pre-Annotation Strategies for Nested NER in Historical Latin Notarial Deeds
Charlene Ellul, Vanessa Buhagiar, Claudia Borg and Charlie Abela 407

From Lemmatization to Legal Terminology: Assessing an Hybrid Pipeline on Justinian's Digest
Paola Marongiu and Eva Sassolini 418

UppsalaNLP at EvaLatin 2026: Multilingual parsing for Latin
Sara Stymne 429

Contextual Probing for Low-Resource Named Entity Recognition in Latin
Maria Mihaela Trusca, Mark Depauw, Violet Soen, Ine de Daele, Kevin Verbruggen and Tim Van de Cruys 437

Classificatio Sine Iactu – That Is, Zero-Shot NERC in Latin
Luisa Ripoll-Alberola, Fernando Nicolás-Flores and Francisco Javier Muñoz Acebes . 443

Extending omnes flores for the EvaLatin 2026 Dependency Parsing Tasks
Hiroshi Matsuda and Masayuki Asahara 448

Pre-Editorial Normalization for Automatically Transcribed Medieval Manuscripts in Old French and Latin
Thibault Clérice, Rachel Bawden, Anthony Glaise, Ariane Pinche and David Smith . . . 453

OldBERTur: Named Entity Recognition for Medieval Icelandic
Pontus Henningsson, Eva Pettersson and Erik Lenas 469

Neural Machine Translation for Coptic-French: Strategies for Low-Resource Ancient Languages
Nasma Chaoui and Richard Khoury 482

Workshop Program

Monday, May 11, 2026

9:00–9:15 *Welcome and introduction*
Marco Passarotti and Rachele Sprugnoli

09:15–10:30 Oral Session 1

09:15–09:30 *Morphological Annotation of Old Serbian in Universal Dependencies*
Vladimir Polomac and Silvie Cinkova

09:30–09:45 *Tracing Morph Origins in Czech: A Computational Approach to Morph-Level Etymology*
Aleš Manuel Manuel Papáček and Zdeněk Žabokrtský

09:45–10:00 *Uncovering Work from Words: LLM-Based Information Extraction from Historical Petitions*
Ellinor Lindqvist, Eva Pettersson and Joakim Nivre

10:00–10:15 *Extracting Volcanological Knowledge from Historical Texts: A Language-Technology Pipeline for Diachronic Geovisualization*
Costanza Marini, Gianluca Casagrande, Alessio Palmero Aprosio and Claudia Principe

10:15–10:30 *When Lexicographic Quotations Become a Corpus: To Deduplicate or Not to Deduplicate?*
Manuel Favaro, Elisa Guadagnini, Eva Sassolini, Marco Biffi and Simonetta Montemagni

10:30–11:30 Poster Session + Coffee Break

10:30–11:30 *A New State-of-the-Art BERT Model for Judeo-Arabic*
Elisha Rosensweig, Yitzchak Lindenbaum, Hillel Gershuni, Vered Raziel-Kretzmer, Daniel Caine and Avi Shmidman

10:30–11:30 *BEReshiT: an Ancient Hebrew Model based on DictaBERT*
Iglika Nikolova-Stoupak, Maxime Amblard and Frédérique Rey

10:30–11:30 *Automatic Detection of Metaphorical Expressions in Classical Japanese Using WLSP-Enhanced BERT*
Hang Zhu, Mitoki Ohara, Rei Kikuchi, Kanako Komiya, Masayuki Asahara and Sachi Kato

10:30–11:30 *Domain-Aware Error Correction for Citation NER in Medieval Hebrew Responsa*
Shmuel LIebeskind, Maayan Zhitomirsky-Geffet, Binyamin Katzoff, Nati Ben-Gigi and Jonathan Schler

Monday, May 11, 2026 (continued)

10:30–11:30 *From Lemmas to Links: A Lemma Bank for Ancient Greek*
Colin Swaelens, Francesco Mambrini and Marco Passarotti

10:30–11:30 *Across Generations: A Comparative Analysis of NER for Latin Inscriptions from Classical Machine Learning to LLMs*
Wenhui Cui and Phillip Benjamin Ströbel

10:30–11:30 *POS Tagging with Generative LLMs for Historical Germanic Low-Resource Languages: An Evaluation Against Fine-Tuned BERT*
Irene Miani, Gregory Darwin and Sara Stymne

10:30–11:30 *From Manuscript to Model: Developing HTR for Medieval Greek*
Nicklas Sindlev Andersen, Byron MacDougall, Tariq Yousef and Aglae Pizzone

10:30–11:30 *I, RE:Claudius 256: Towards Linking Classical Latin Person Mentions to a Domain-specific Knowledge Base*
Marijke Beersmans, Evelien de Graaf, Julie Nijs, Valeria Irene Boano, Alek Keersmaekers, Mark Depauw, Tim Van de Cruys and Margherita Fantoli

10:30–11:30 *Capturing Ancient Chinese Sense Induction with Automatic Pipelines*
Guan-Yu Tseng, Chunki Lim, Chih-Han Lin, Tung-Le Pan, Yu-Chieh Wang, Lang-Ching Yeh and Shu-Kai Hsieh

11:30–12:15 Oral Session 2

11:30–11:45 *A Computational Evaluation of Syllabic Hypotheses for Rongorongo: Evidence from N-gram Analysis*
Evgeniya Korovina

11:45–12:00 *Building a Corpus and Database for Rare and Undeciphered Scripts*
Beata Megyesi, Rune Rattenborg, Benedek Láng, Michelle Waldispühl and Mihály Héder

12:00–12:15 *A Layered Annotation Workflow for Semitic Epigraphy*
Tal Bernstein, Shai Gordin and Letizia Cerqueglini

Monday, May 11, 2026 (continued)

12:15–13:00 **EvaLatin**

12:15–12:25 *Overview of the Dependency Parsing Task at EvaLatin 2026*
Federica Iurescia, Marco Passarotti and Rachele Sprugnoli

12:25–12:35 *THIVLVC: Retrieval Augmented Dependency Parsing for Latin*
Luc Pommeret, Thibault Wagret and Jules Deret

12:35–12:45 *Overview of the Named Entity Recognition Task at EvaLatin 2026*
Valeria Irene Boano, Eleonora Litta and Matteo Romanello

12:45–12:55 *Transfer Learning for Named Entity Recognition of Classical Latin through LLM Prompting*
Callum Chan

14:00–15:15 **EvaHan**

14:00–14:02 *Opening Remarks*
Li Bin

14:02–14:07 *Invited talk*
Feng Zhiwei

14:07–14:22 *Overview of EvaHan2026: The First International Evaluation of Ancient Chinese OCR and Layout Analysis*
Dongbo Wang, Dongmei Zhu, Jieqiong Li, Chang Liu, Ruifeng Wu, Fan Yang, Zhu Yue, Xin Gao, Zhao Zhixiao, Xue Zhao, Zhongzheng Wu, Liu Liu and Li Bin

14:23–14:26 *A Multi-Stage System for Ancient Chinese OCR and Layout Understanding in the EvaHan2026 Shared Task*
KeYan Liang and Meiling Liu

14:27–14:31 *A Multi-Modal Recognition Framework for Ancient Books Integrating DoRA-DPO Text Recognition and YOLO Layout Analysis*
Chaokun Zhang, Xin Wen and Tongtong Zhou

14:32–14:35 *Beijing Normal University at EvaHan 2026: Enhancing Ancient Chinese Character Recognition and Layout Analysis via VLM Fine-Tuning and Linguistic Post-Processing*
Yihuan Yin and Qian Zhao

Monday, May 11, 2026 (continued)

14:36–14:39 *A Dual-Modality Framework for Ancient Document Layout Analysis and Text Recognition*
Qi Fan, Jieming Hu and Chen Ye

14:40–14:43 *EvaHan 2026 Ancient Books Multimodal OCR and Layout Analysis System Technical Report*
Chenrui Zheng

14:44–14:47 *A Parameter-Efficient and Data-Centric Framework for Ancient Chinese Text Recognition and Layout Analysis*
Yuchun Meng

14:48–14:51 *LVLM Optimization for Ancient Chinese Book Image Analysis with Task-specific Augmentation and Instruction Tuning*
Tian Xia, Yulong Liu, Yilin Wang, Yumeng Yang, Dongheng Cai, Yuyang Tan and Menghui Yang

14:52–14:55 *Data-Centric Strategies for Ancient Chinese Text Recognition: Augmentation, Annotation Refinement, and Style Transfer in EvaHan 2026*
Li Chengfei, Zhang Yunjie, Li XiaoYi, Quan Changshun, Cao Taihe and Liu Bin

14:56–14:59 *AnandaSky: A Vision–Language Model for Line-Level Transcription of Historical Sinographic Documents*
Colin Brisson, Ayoub Kahfy, Frédéric Constant and Marc Bui

15:00–15:03 *Multimodal Ancient Document Parsing: Technical Report for EvaHan2026 Competition*
Liqi He, Qiwei Li, Ziye Yang and Zuchao Li

15:04–15:07 *Unified Vision–Language Framework for End-to-End Ancient Chinese OCR and Layout Analysis: Our Submission to EvaHan 2026*
Yuhan Shu and Huizi Zhou

15:08–15:11 *Building Character(s): Synthetic Data and In-Context Learning Strategies for Few-Shot Ancient Chinese Recognition*
Denise Atzori, Marie Bizais-Lillig, Mathias Garnier, Maxime Létoffé, Charles Planque, Tianjie Yin and Chahan Vidal-Gorène

15:11–15:15 *Closing remarks for EvaHan 2026*
Li Bin

Monday, May 11, 2026 (continued)

15:15–16:00 **Oral Session 3**

15:15–15:30 *The UD_Latin-PROIEL as Linked Open Data: Integrating a Latin Treebank into the LiLa Knowledge Base*
Lucas Consolin Dezotti, Marco Passarotti, Federica Iurescia and Giovanni Moretti

15:30–15:45 *Language Models for the Restoration of Latin Legal Manuscripts*
Shibingfeng Zhang, Edoardo Caraffa, Annafelicia Zuffrano, Maddalena Modesti and Giovanni Colavizza

15:45–16:00 *Evaluating Hierarchical Aggregation and LLM-Based Matching for Synset Selection in Ancient Greek*
Luca Brigada Villa, Marco Passarotti, Chiara Zanchi, Riccardo Ginevra, Erica Fratellini and Eleonora Litta

16:00–17:00 **Poster Session + Coffee Break**

16:00–17:00 *Miktub: A Manuscript Dataset of Historical Maltese for Handwritten Text Recognition*
Thomas Koppens and Claudia Borg

16:00–17:00 *Smelling the Past: Investigating Historical Models for Olfactory Event Extraction*
Teresa Paccosi and Marijn Koolen

16:00–17:00 *Contemporizing 20-th Century Estonian*
Heiki-Jaan Kaalep

16:00–17:00 *Cost-Aware Pre-Annotation Strategies for Nested NER in Historical Latin Notarial Deeds*
Charlene Ellul, Vanessa Buhagiar, Claudia Borg and Charlie Abela

16:00–17:00 *From Lemmatization to Legal Terminology: Assessing an Hybrid Pipeline on Justinian's Digest*
Paola Marongiu and Eva Sassolini

16:00–17:00 *UppsalaNLP at EvaLatin 2026: Multilingual parsing for Latin*
Sara Stymne

16:00–17:00 *Contextual Probing for Low-Resource Named Entity Recognition in Latin*
Maria Mihaela Trusca, Mark Depauw, Violet Soen, Ine de Daele, Kevin Verbruggen and Tim Van de Cruys

Monday, May 11, 2026 (continued)

16:00–17:00 *Classificatio Sine Iactu – That Is, Zero-Shot NERC in Latin*
Luisa Ripoll-Alberola, Fernando Nicolás-Flores and Francisco Javier Muñoz Acebes

16:00–17:00 *Extending omnes flores for the EvaLatin 2026 Dependency Parsing Tasks*
Hiroshi Matsuda and Masayuki Asahara

17:00–17:45 Oral Session 4

17:00–17:15 *Pre-Editorial Normalization for Automatically Transcribed Medieval Manuscripts in Old French and Latin*
Thibault Clérice, Rachel Bawden, Anthony Glaise, Ariane Pinche and David Smith

17:15–17:30 *OldBERTur: Named Entity Recognition for Medieval Icelandic*
Pontus Henningsson, Eva Pettersson and Erik Lenas

17:30–17:45 *Neural Machine Translation for Coptic-French: Strategies for Low-Resource Ancient Languages*
Nasma Chaoui and Richard Khoury

17:45–18:00 Closing

Morphological Annotation of Old Serbian in Universal Dependencies

Vladimir Polomac[1], Silvie Cinková[2]

University of Kragujevac, Faculty of Philology and Arts[1], Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics[2]

v.polomac@filum.kg.ac.rs[1], cinkova@ufal.mff.cuni.cz[2]

Abstract

We report on the morphological tagging of Old Serbian in the Universal Dependencies framework. To facilitate the manual annotation, we pre-processed the data with the Old Church Slavonic 2.12 UDPipe model. The decision was based on the known similarity of these two languages as well as on the declared performance of this model compared to other models for historical varieties of Slavic languages. With over 3,000 manually annotated tokens, we evaluated the performance of the relevant pre-trained UDPipe2 models of historical Slavic languages. Besides, we also trained and evaluated custom models with UDPipe1 containing the annotated Old Serbian data. We have found that: (1) for this particular domain and amount of training data, the most suitable model is UD Old East Slavic – Birchbark 2.12, although its declared performance is much lower than that of Old Church Slavonic; (2) even 3,000 tokens of Old Serbian increase the performance of UDPipe1 models almost to the level of the Birchbark 2.12 model. The dataset is publicly available at https://doi.org/10.5281/zenodo.19317842.

Keywords: Old Serbian, Old Church Slavonic, Universal Dependencies, PoS Tagging, UDPipe, Serbian Medieval Charters, tagger evaluation.

1. Introduction

Corpus searches, as well as diverse text-mining methods, benefit from linguistic markup (lemmatization, part-of-speech tagging, and syntactic relations). Universal Dependencies (McDonald et al., 2013; Zeman, 2018; de Marneffe et al., 2021) is a largely theory-agnostic cross-lingual scheme providing principles of tokenization and sentence splitting, as well as a set of morphological tags and syntactic labels with definitions, all well-grounded in the language typology. A cross-lingual annotation standard paved the way for powerful language-agnostic tools, e.g., UDPipe (Straka et. al., 2016; Straka 2018), SpaCy (Honnibal et al., 2020), or Stanza (Qi et al., 2020), which can process diverse languages (including their specific varieties and domains) by plugging in individual language models trained on dedicated data sets.

The goal of our project is to contribute to the Universal Dependencies GitHub repository a valid UD-annotated data set of the oldest preserved Serbian texts from the 12th and 13th century. This paper reports on its first steps – the annotation pre-processing and lemmatization decisions.

The paper is organized as follows: Section 2 provides information about the Old Serbian language, its relationship to Proto-Slavic and Old Church Slavonic, as well as the criteria for selecting the manuscripts used in the study. Section 3 elaborates on the principles of text lemmatization in more detail, while Section 4 covers the principles of morphosyntactic annotation. Section 5 presents the initial results of tagging, automatic lemmatization, and morphosyntactic annotation using UDPipe. The final chapters briefly summarize the results (section 6) and point to the prospects for further work (section 7).

2. Old Serbian: Data Selection and Preprocessing

Serbian belongs to the group of South Slavic languages that developed as distinct languages from the common Proto-Slavic language in the second half of the first millennium AD, following the migration to the Balkan Peninsula. The early history of the Serbian language can be reconstructed based on comparisons with other Slavic languages, especially Old Church Slavonic, which emerged in the 9th century for liturgical purposes among the Western Slavs and later spread throughout the entire area of Pax Slavia Orthodoxa. The earliest preserved texts in the Serbian language date from the late 12th century: the Miroslav's Gospel was written in Serbian Church Slavonic, a variety of Old Church Slavonic that incorporated features of the 12th-century Serbian vernacular; the Charter of Ban Kulin (1189), was written in the Serbian vernacular with minimal elements of Serbian Church Slavonic. This means that in the earliest Serbian texts from the late 12th century and the first half of 13th century, features of the Old Serbian vernacular and Serbian Church Slavonic are mixed. Since these are different varieties of the same language, rather than two different languages (e.g., Czech and Latin in the medieval Bohemia), it is necessary to develop a unified

Proceedings of LT4HALA 2026 @ LREC 2026, pages 1–6.
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

methodological framework for the lemmatization and annotation of both varieties.

For the purposes of this research, we selected a corpus of Serbian medieval charters from the late 12th and early 13th centuries. This selection is justified by the following criteria: a) they are the oldest preserved Serbian texts, b) they are texts with the most elements of Old Serbian vernacular. The list of texts manually annotated[1] in the CoNLL-U Editor (Heinecke, 2019) is given in Table 1.

ID	Title	Tokens
№1	The Signature of Grand Prince Stefan Nemanja and Prince Miroslav on the Latin-written charter to Dubrovnik (1186)	16
№2	The Charter of Ban Kulin to Dubrovnik (1189)	208
№3	The Signature of Prince Miroslav on the Latin written charter to Dubrovnik (1190)	4
№4	The Charter of Grand Prince Stefan Nemanja (Monk Simeon) to the Hilandar Monastery (1198–1199)	898
№5	The Charter of Grand Prince Stefan Nemanjić (Saint Sava) to the Hilandar Monastery (1207–1208)	1519
№6	The Charter of Grand Prince Stefan Nemanjić to Dubrovnik (1214–1217)	220
№7	The Charter of King Radoslav to Dubrovnik (1234)	489
№8	The Charter of Prince Andrej to Dubrovnik (1214–1235)	177
	TOTAL	3531

Table 1: Initial document collection (manually annotated)

The table shows that these texts are of varying lengths. Documents №1 and №3 are only fragments with a small number of tokens. On the other hand, there are more extensive documents (e.g., №4 and №5) with several dozen sentences and several hundred tokens. The documents addressed to Dubrovnik (№1, №2, №3, №6, №7, №8) are related to trade and therefore are written in Old Serbian with minimal presence of Serbian Church Slavonic features. Conversely, the documents addressed to the Hilandar Monastery on Mount Athos (№4, №5) predominantly feature Serbian Church Slavonic, with Old Serbian vernacular present to a lesser extent.

The potentially relevant UDPipe models were old_church_slavonic-ud-2.12-2307017 (PROIEL), old_east_slavic-birchbark-ud-2.12-230717 (Birchbark), and old_east_slavic-torot-ud-2.12-230717 (TOROT), considering the similarity of language and text domains to our data, and their declared performance. The domains and vernaculars on which these models were trained were very similar to our data, but with corpus-specific transcription conventions, none of which quite matched our data. Hence, we opted for the Old Church Slavonic model due to its best declared performance (see Table 2).

UDPipe2 Model	UPOS	Lemma
old_church_slavonic-proiel-ud-2.12-230717 (PROIEL)	**96.39**	**90.18**
old_east_slavic-birchbark-ud-2.12-230717 (Birchbark)	88.50	65.58
old_east_slavic-torot-ud-2.12-230717 (TOROT)	95.40	88.09

Table 2: Declared performance of selected UDPipe2 models on UPOS and lemma (F1 score)

3. Lemmatization Principles

Lemmatization of historical language varieties is complex due to lack of knowledge, semantic homonymy, and diachronic/orthographic lemma variation as well as diglossia, with variants often co-occurring within the same text (Pichhadze 2017). Semantic homonymy affects even very frequent words, e.g., *съ* represents both the preposition *with* and the determiner *this*.

When lemmatizing the texts from the corpus, we adhered to three main principles: a) the principle of orthographic normalization, b) reconstruction of the presumed linguistic state from the late 12th and early 13th centuries, and c) the principle of preserving variant lemma forms that reflect specific phonetic features of Old Serbian and Serbian Church Slavonic.

The principle of orthographic normalization implies that lemmas are given in the modern Cyrillic script, with the addition of several specific letters for phonemes present in Old Serbian and

[1] The annotation of the texts was carried out solely by the first author of the paper, and therefore the evaluation of interannotator agreement was not applicable.

Serbian Church Slavonic vocalism at the end of the 12th and beginning of the 13th centuries: the letters <ь>, <ъ>, and <ы> for the semivowel [ə], the vowel *yat* [ě], and the vowel *yeri* [y].

The principles of lemmatization were developed not only for the specific needs of this work but also for the lemmatization of the future electronic corpus of Old Serbian and Serbian Church Slavonic as a whole. Therefore, the second principle involved reconstructing the lemma always according to the presumed state of the phonological system common to Old Serbian and Serbian Church Slavonic at the end of the 12th and beginning of the 13th centuries, regardless of later linguistic development. Since both linguistic varieties are present in the charter texts, the third principle involved retaining variant lemma forms that reflect their specific phonetic features: for example, Old Serbian *noć* and Serbian Church Slavonic *nošt* for *night* or Old Serbian *ja* and *jaz* and Serbian Church Slavonic *az* for the first-person singular pronoun. These lemmas will be linked through a ModernLemma attribute in the MISC column.

4. Universal PoS and Morphosyntactic Features

Old Serbian and Serbian Church Slavonic are grammatically much closer to Old Church Slavonic than to modern Serbian. Therefore, when adapting the tag set for the annotation of parts of speech and morphosyntactic features, we started from the tag set for Old Church Slavonic developed within the PROIEL project (Eckhoff 2018). Unlike the Old Church Slavonic tag set, which does not include PART, PUNCT, and SYM, our tag set encompasses all universal PoS tags. The category of determiner (DET), which is not common in traditional grammatical descriptions of the Serbian language and its historical varieties, has been introduced as a tag for pronominal adjectives to ensure consistency with Old Church Slavonic and modern Slavic languages (Zeman, 2015). The SYM tag is used to mark the cross that appears in the texts either in place of a ruler's signature or as a symbol of divine invocation at the beginning of charters. The issue of tagging participles, which are characteristic of other Slavic languages as well (Zeman 2016: 145), has been resolved by consistently tagging them as VERB, even though they also have nominal features such as Case, Gender, Number, or Variant. We deviated from this principle only in a few instances where the verb lemma could not be reconstructed for adjectives derived from participles, but even in such cases, the feature VerbForm=Part was included with the adjective.

Partial adaptations of the Old Church Slavonic set were also implemented when marking morphosyntactic features and values. In the category of nouns, one can find Polarity=Neg for nouns with lexical negation. A small number of indeclinable nouns have the feature InflClass=Ind. Proper nouns, in addition to the features Case, Gender, and Number, also have the feature NameType, which can take the values NameType=Geo (geographical name, toponym), NameType=Giv (given name), NameType=Sur (surname), NameType=Pat (patronym), and NameType=Nat (ethnonym). In the category of first and second-person personal pronouns, reflexive pronouns for all persons (*sebe*, *se*), and nominal pronouns (*kto* and *što*), the Gender feature is omitted. Adjectival pronouns are always tagged as DET, with the feature PronType, which can have the values PronType=Dem, PronType=Rel, PronType=Int, PronType=Ind, PronType=Neg, and PronType=Tot. To the morphosyntactic features of adjectives that are marked in Old Church Slavonic (Case, Degree, Gender, Number, and Variant), we have added Poss=Yes for possessive adjectives, as well as Polarity=Neg for adjectives with lexical negation. In the category of numerals, we consistently marked the features NumForm (with values NumType=Word for numbers written in words and NumType=Cyrill for numbers written with letters in numerical value) and NumType (with values Card, Ord, and Sets). The most significant change in the category of verbs involved the method of marking aorist and imperfect. For aorist, we used VerbForm=Past as in Old Czech, and for imperfect, VerbForm=Imp (Zeman et al., 2023)[2].

5. Tagging Results

Among the three pre-trained UDPipe2 models in consideration (Table 2), Old Church Slavonic has best declared performance, and therefore we used it for the pre-processing.

Our initial manual annotation batch consisted of 3,531 tokens (eight complete documents). We evaluated the three pretrained models on this batch (henceforth *3k*), with results as listed in Table 3 and 4, using gold tokenization.

Although we expected some decrease of model performance on *3k*, it being a different vernacular, the margin was surprising: in UPOS, the Old Church Slavonic model dropped by 32.05. Similarly, TOROT dropped by 15.62 points, while Birchbark by mere 6.6 points, suddenly ranking first among the pre-trained UDPipe2 models.

In lemmatization, Old Church Slavonic ranked again worst, dropping by almost 60 points.

[2] For marking these forms, we cannot use the Aspect feature because it would conflict with the lexical aspect that has developed in Slavic languages (Zeman et al., 2023).

Birchbark and TOROT outperformed it by ten points and with almost identical results.

UDPipe2 Model	UPOS	UPOS diff
old_church_slavonic-proiel-ud-2.12-230717 (PROIEL)	64.34	**-32.05**
old_east_slavic-birchbark-ud-2.12-230717 (Birchbark)	**81.90**	-6.6
old_east_slavic-torot-ud-2.12-230717 (TOROT)	79.78	-15.62

Table 3: Performance of selected UDPipe2 models on UPOS (F1 score), on the *3k* Old Serbian dataset, with differences from the declared performance

UDPipe2 Model	Lemma	Lemma diff
old_church_slavonic-proiel-ud-2.12-230717 (PROIEL)	30.95	**-59.23**
old_east_slavic-birchbark-ud-2.12-230717 (Birchbark)	**40.83**	-24.75
old_east_slavic-torot-ud-2.12-230717 (TOROT)	40.81	-47.28

Table 4: Performance of selected UDPipe2 models on lemma (F1 score), on the *3k* Old Serbian dataset, with differences from the declared performance

Facing such a performance drop in the pretrained UDPipe2 models, we trained several UDPipe1 models to include *3k* in different combinations with Old Church Slavonic, Birchbark, and TOROT train data.[3] We used the default parameters of the udpipe_train function in the R library udpipe (Wijfels, 2023). We measured the performance of the UDPipe1 models by a ten-fold cross-evaluation, where the train data contained all train data of the UD selected treebanks and 9/10 of *3k* and the test data contained 1/10 of *3k* (Figure 1).

6. Discussion

The experimental results show that none of our UDPipe1 models reliably outperforms the pre-trained old_east_slavic-birchbark-ud-2.12-230717 UDPipe2 model in UPOS. The greatest advantage of Birchbark is possibly the domain similarity: practical written communication, administrative notes, business and church records written in a vernacular. When training UDPipe1 models, no combination with 3k outperforms a model trained on 3k alone. We have also included a cross-validated result for a UDPipe1 model trained just on the Birchbark 2.12 training data to assess the performance difference between the UDPipe1 and UDPipe2 models trained on the same data and applied to 3k. The performance of 46.29 was poor (note that the UDPipe2 Birchbark model reached 81.9). Adding *3k* raised the performance of the Birchbark based UDPipe1 model to 78.54, which is anyway less than 3k alone, yet with a somewhat smaller dispersion. The pretrained UDPipe1 models old_church_slavonic-proiel-ud-2.5-191206 and old_russian-torot-ud-2.5-191206 also performed on *3k* much worse than their UDPipe2 counterparts.

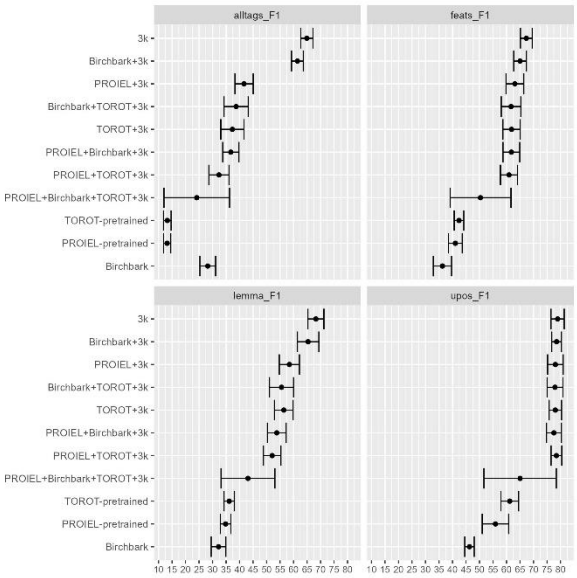


Figure 1: Performance of UDPipe1-trained models (average F1 in a ten-fold evaluation)

We analyzed the errors made by the pre-trained models for individual parts of speech. The most striking finding was that Old Church Slavonic 2.12 fails to identify punctuation, classifying colons as ADJ and periods as NOUN, due to a specific lemmatization, which interprets them in words and the parts of speech are then parts of speech of those words. However, it was not its only issue on the Old Serbian domain. When we compared the proportions of correct guesses except punctuation marks (once for the entire dataset), the models ranked the same, and the difference from TOROT 2.12 (second) remained apparent, although it had decreased.

All three models were above 90% in CCONJ and NOUN. Birchbark and TOROT also reached this mark in ADP (Old Church Slavonic below 80%) and PUNCT. Birchbark fails to detect INTJ (which are rare) and almost fails to detect ADV and AUX (but all models perform poorly here); it is also

[3] Unfortunately, only the older UDPipe1 tool has a user-friendly API to train custom models, whereas doing so with UDPipe2 requires a significant programming expertise and processing capacity.

noticeably worst at PRON. For more details see Figure 2.

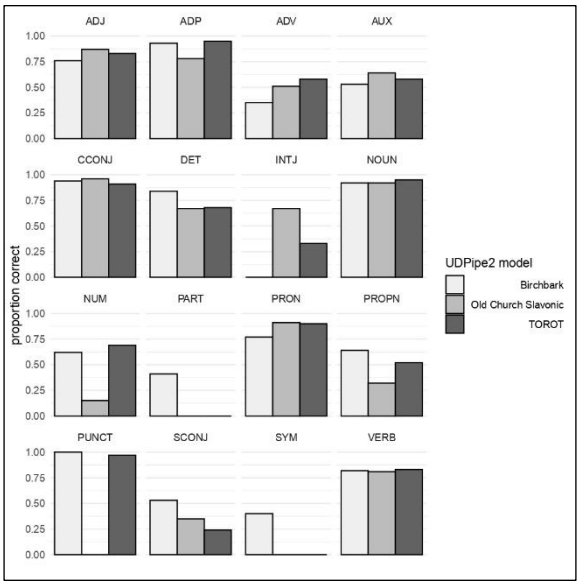


Figure 2: Proportion of correct guesses by UPOS and model

Regarding lemmatization, the performance of Old Church Slavonic and Old East Slavic models is sub-optimal (as partly expected), which can be explained by different principles of orthographic normalization of lemmas. More precisely, they assign the Old Serbian tokens to a correct lemma according to the conventions of their train data, which differ from those for Old Serbian. A good example is numerals written in letters (in Old Church Slavonic, numbers can be represented by letters with the corresponding index in the Cyrillic alphabet, e.g. *а* would stand for 1, *в* for 2, etc.). They pose no problem for UPOS tagging: only the lemmatization is completely different. On the other hand, the orthography of many pronouns is so different that they do not even get recognized as pronouns.

Another difference (among many) is that the traditional normalization of Old Serbian does not systematically use **ъ** and **ь** as endings. For illustration, we can mention the token ***богъ*** (*God*), which we lemmatize as ***бог***, while the Old Church Slavonic and Old East Slavic models lemmatize it as ***богъ***.

7. Conclusion and Future Work

From our experiments with three pre-trained UDPipe2 models and a selection of custom trained UDPipe1 models we conclude that our next annotation batch of Old Serbian texts will benefit from pre-processing with the old_east_slavic-birchbark-ud-2.12-230717 model – at least regarding UPOS. We intend to run the same experiment with the Stanza (Qi et al., 2020) models, whose declared performance on the three relevant languages is higher than UDPipe2's.

We will repeat this series of experiments incrementally with each new annotation batch of 5,000 tokens, including the evaluation of lemma and the universal features. As soon as the manually annotated corpus reaches 20,000 tokens, we will gradually add syntactic dependencies, using the most suitable model for their preprocessing, and finally contribute it to the UD github repository to make it eligible for the regular model training workflow of the UDPipe2 developers. The current dataset is available in the Zenodo repository (Polomac, 2026).

8. Acknowledgments

The work was carried out within the research grant *Universal Dependencies for Old Serbian and Serbian Church Slavonic: Creating a Training Dataset for Lemmatization and Morphosyntactic Annotation Using UDPipe*, supported by the Computational Literary Studies Infrastructure (CLS INFRA) project. This output was created with the support of the Ministry of Education, Youth and Sports and the Operational Program Johannes Amos Comenius within the project Open Science II (reg. No. CZ.02.01.01/00/24_030/0015041). The work described herein has also been using data, tools, and services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062).

9. Bibliographical References

de Marneffe, M.-C., Manning, C. D., Nivre, J., Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 47(2): 255-308.

Eckhoff, H., Bech, K., Bouma, G., Eide, K., Haug, D., Haugen, O. E. and Jøhndal M. (2018). The PROIEL treebank family: a standard for early attestations of Indo-European languages. *Language Resources and Evaluation*, 52: 29–65. URL: https://doi.org/10.1007/s10579-017-9388-5.

Heinecke, J. (2019). CoNLLU-Editor: a fully graphical editor for universal dependencies treebank files. In *Proceedings of the Third Workshop on Universal Dependencies*, pages 87–93, Paris, France. Association for Computational Linguistics. URL: https://aclanthology.org/W19-8010.

McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Bertomeu Castelló, N. and Lee, J. (2013) Universal Dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–

97, Sofia, Bulgaria. Association for Computational Linguistics. URL: https://aclanthology.org/P13-2017.

Pichhadze, A. (2017). Razmetka cerkovnoslavjanskih i drevnerusskih tekstov: problemy lemmatizacii. *Filologija*, 68: 143–155.

Polomac, V. (2026). UD_Old_Serbian-DiaC [Data set]. Zenodo. https://doi.org/10.5281/zenodo.19317842

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Stroudsburg, USA. *Association for Computational Linguistics.* URL: https://aclanthology.org/2020.acl-demos.14/

Straka, M., Hajič, J. and Straková, J. (2016). UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297, Poreč, Slovenia. European Language Resources Association. URL: https://aclanthology.org/L16-1680.

Zeman, D. (2015). Slavic languages in universal dependencies. In K. Gajdošová and Adriána Žáková (eds.), *Natural Language Processing, Corpus Linguistics, E-learning*, RAM-Verlag, Lüdenscheid: RAM-Verlag, pp. 151–163.

Zeman, D. (2016). Universal Annotation of Slavic Verb Forms. *The Prague Bulletin of Mathematical Linguistics*, 105 (2016): 143–193.

Zeman, D. (2018). *The World of Tokens, Tags and Trees*. Studies in Computational and Theoretical Linguistics, vol. 19, first ed. ÚFAL MFF UK, Praha.

Zeman, D., Kosek, P., Březina, M. and Pergler, J. (2023). Morphosyntactic Annotation in Universal Dependencies for Old Czech. *Jazykovedný časopis/Journal of Linguistics*, 74: 214–222.

10. Language Resource References

Honnibal, M., Montani, I., Van Landeghem, S. and Boyd A (2020). spaCy: Industrial-strength NaturalLanguage Processing in Python. doi:10.5281/zenodo.1212303.

Straka, M. (2018): UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In *Proceedings of CoNLL 2018: The SIGNLL Conference on Computational Natural Language Learning*, pp. 197-207. Association for Computational Linguistics, Stroudsburg, PA, USA.

Wijffels, J. (2023). udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit. URL: https://CRAN.R-project.org/package=udpipe

Tracing Morph Origins in Czech: A Computational Approach to Morph-Level Etymology

Aleš Manuel Papáček, Zdeněk Žabokrtský
Charles University, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics (ÚFAL)
Malostranské náměstí 25, 118 00 Prague, Czech Republic
papacek@ufal.mff.cuni.cz, zabokrtsky@ufal.mff.cuni.cz

Abstract

Modern languages remain connected to ancient ones in multiple ways, including through etymology; for instance, Latin is among the most influential sources of borrowings in (modern) Czech, whether transmitted directly or mediated through other languages. This work focuses on predicting the etymological origin of individual morphs in Czech words. Given morphologically segmented Czech sentences, the task is to determine for each morph whether it is native or borrowed, and if borrowed, to identify the languages through which it entered Czech. Although some linguists have examined etymology at the level of individual morphs rather than whole words (Arkadiev et al., 2015), to our knowledge, no computational work has yet addressed this level of analysis. We created a manually annotated dataset of 300 Czech sentences comprising around 10,000 morphs with morph-level etymology labels, and trained supervised models using character-based and structural features. Our best lightweight system is a feed-forward neural network with a single hidden layer, trained on data augmented with entries from an etymological dictionary, reaching 96.2 % F1 on the test set. We also developed and tested several prompting variants for large language models; the best model `Claude-Opus-4.5`, achieved 97.8 % F1. We release the code, prompts, and dataset as open source at https://github.com/ampapacek/MorphemeOrigin.

Keywords: Etymology, Language contact, Borrowing, Morphological segmentation, LLM

1. Introduction

Borrowing is one of the main forces driving language change (Grant, 2020). English, for example, owes much of its vocabulary to French and, through it, to Latin. Latin itself borrowed extensively from Greek (Durkin, 2014). What gets borrowed is not always uniform at the word level: borrowed roots often combine with native morphology, and native roots can take borrowed affixes.

Modern European languages therefore contain many hybrid words whose morphs come from different sources. For instance, *antiviruses* combines the Greek-derived prefix *anti-* with the Latin root *virus* and the English inflectional ending *-es*.

Morphs are the building blocks of words, yet most existing etymological resources stop at the word level. Studying morph origins offers a much finer view of language history and contact. Linguistic research has explored morph borrowing in works such as Arkadiev et al. (2015) and Seifart (2015). However, so far, existing computational approaches have focused on word-level etymology rather than morph-level analysis.

Automatically predicting morph origins is difficult: the space of possible morphs is large, and many morph surface forms are short and ambiguous, meaning the same form can occur in different words with different etymological origins. This is also hard to model computationally because much of the relevant knowledge is implicit in linguistic expertise, and existing etymological resources are mostly word-level rather than fine-grained at the level of word subparts.

We frame morph-level etymology prediction as a supervised learning task. Given a segmented sentence, the goal is to decide for each morph whether it is native or borrowed and, if borrowed, which languages are involved in its borrowing path.

To enable this, we manually annotated almost 10,000 morphs in 300 sentences from the SIGMORPHON 2022 shared task (Batsuren et al., 2022) with morph-level etymological labels.

We compare several baselines, including a majority-label baseline, memorization of morph–origin pairs, and rule-based methods using the Czech Etymological Lexicon (Rejzek et al., 2025). For the supervised approach, we extract character-based and structural features (character n-grams, morph type, and position within the word) and train standard classifiers: logistic regression (LR), support vector machines (SVM), and multilayer perceptrons (MLP). We also test morph and word embeddings and a semi-supervised self-training setup.

In addition, we evaluate prompt-only predictions from several state-of-the-art LLMs under different in-context training settings.

Although the general ideas are language-independent, our experiments focus on Czech, where suitable etymological resources are available, primarily the Czech Etymological Dictionary (Rejzek, 2019). We believe that the proposed

Proceedings of LT4HALA 2026 @ LREC 2026, pages 7–18.
11 May 2026. *©ELRA Language Resources Association (ELRA), 2026*

methodology should in principle be transferable to other languages, given the availability of morph-level etymological annotations.

The rest of the paper is organized as follows. Section 2 introduces key concepts from morphology and etymology that provide the necessary background. Section 3 reviews related research. Section 4 describes the data and annotation process and Section 5 outlines the design of our machine learning approaches. Section 6 explains the evaluation setup and Section 7 reports and analyses the experimental results.

In addition, the appendix A presents an etymological analysis of a diachronic Czech legal corpus covering more than 100 years. It investigates how the ratio of native to borrowed morphs evolves over time and which source languages are most strongly represented among the borrowed morphs.

2. Background Theory

One of the central tasks of morphology is to investigate the internal structure of words, often by segmenting them into the smallest meaning-bearing units, known as **morphs**.[1] For example, the English word *disrespectfulness* can be decomposed into *dis-* + *re-* + *spect* + *-ful* + *-ness*. Each morph carries lexical or grammatical meaning.

2.1. Basic Terminology

The distinction between *morpheme* and *morph* is notoriously vague (Haspelmath, 2020). Haspelmath observes that the term *morpheme* is used in several conflicting senses, ranging from concrete minimal forms to abstract grammatical units. When we work with segmented word forms, we refer to concrete realizations such as *-s* or *-es*, rather than abstract categories like *plural*. In that case, we use the term *morph* in the sense defined by Haspelmath (2020): a morph is a minimal linguistic form.

The **root** is the core morph of a word and carries its main lexical meaning. Morphs that must attach to a root are called **affixes**, which modify either lexical meaning or grammatical properties. **Prefixes** attach before the root, while **suffixes** follow it. An **ending** is a specific type of suffix that marks grammatical features such as tense, person, or number. An **interfix** appears between roots.

[1] In current theoretical debates, word-based approaches to morphology are gaining prominence (Blevins, 2016). These approaches treat words as the primary units of morphological analysis, focusing on relations between whole words, and regard morphs as secondary abstractions derived from them. However, we believe that modeling the origins of individual morphs, as pursued in this paper, is compatible with viewing morphs as secondary abstractions.

Affixes are commonly classified as either **inflectional** or **derivational**. **Inflection** modifies a word's grammatical form without changing its lexical meaning, whereas **derivation** creates a new lexeme, often with a different meaning or part of speech (Aronoff and Fudeman, 2011).

2.2. Etymology

Etymology studies how words originate and evolve over time, and whether they are inherited or borrowed from other languages. Morph-level etymology examines this at the level of individual morphs, offering finer-grained insights, especially for hybrid words whose parts come from different sources.

In much of the language contact literature, the common assumption is that affixes are usually borrowed only indirectly. That is, words containing a given affix are first borrowed as wholes from another language, and only later are these words reanalysed, and the recurring affix generalized to native stems.

For example, the suffix *-ism* (from Greek, via Latin and French) entered English through borrowed words such as *baptism*, *criticism*, and *organism*, and was subsequently extended to native bases in formations like *snobbism* or *veganism*.

There are, however, also cases where affixes have been borrowed directly, without the mediation of complex loanwords. Although such cases are rare (Seifart, 2015), examining etymology at the level of individual morphs can still provide valuable insights, even when most affixes have entered a language through indirect borrowing.

Borrowing can also occur through a chain of intermediary languages. For example, the English word *school* entered Middle English as *scole* from Old English *scol*, from Proto-West Germanic **skolu*, from Late Latin *schola*, from Ancient Greek *skhole*, and probably from Proto-Indo-European (Oxford English Dictionary, 2025).

We should note that our assumption that historical pathways of words can be captured rigorously as plain sequences of languages, is an operational simplification. The etymology of a word is not always straightforward; due to various reasons, it can be difficult – or even impossible – to determine the exact trajectory by which it entered the language or through which intermediate languages it passed.

3. Related Work

To our knowledge, no computational linguistics research specifically addresses identifying the etymological origin of individual morphs on a larger scale. While there is some research on affix borrowing and a few tools exist for determining the etymology of entire words, no approach combines these

perspectives at the morph-level.

3.1. Linguistic Research on Morph Borrowing

Linguistic research has examined how morphology spreads across languages through language contact and how such processes can be distinguished from ordinary inheritance. Johanson and Robbeets (2013) draw a key distinction between shared affixes that are *copies*—replicated through contact—and those that are *cognates* inherited from a common ancestor, framing borrowing as an active process of *code-copying*. Seifart (2015) further distinguishes between *direct* borrowing, based on speakers' knowledge of the donor language, and *indirect* borrowing, where affixes spread through complex loanwords that later become productive in the recipient language.

Gardani (2008) shows that even inflectional morphs—traditionally viewed as resistant to borrowing—can transfer when language contact is intense and structural compatibility is high. Broader perspectives are provided by Arkadiev et al. (2015), which focuses on the borrowing of affixes and shows that, although less common than lexical borrowing, derivational affixes and some basic grammatical endings can also be transferred between languages. Together, these works form a theoretical basis for our approach to studying morph-level etymology computationally.

3.2. Computational Etymology

Building on this linguistic background, recent computational studies have begun to model etymological relationships automatically, though still mostly at the word level. Wu and Yarowsky (2020) predict both the type of etymological relationship (e.g., borrowing or inheritance) and the parent language from which a word originated. Their approach shows that computational modelling of etymology is both feasible and informative, though it does not address the internal morphological structure of words.

The work by Kovácsová (2025) explores methods for training models to automatically identify loanwords across multiple languages. Our research extends this line of inquiry to the morph-level, providing a finer-grained view of how native and borrowed elements combine within words. It also advances beyond binary or single-language classification by reconstructing the complete sequence of etymological origins.

3.3. Resources with Morph-Level Etymological Information

Despite the growing interest in computational etymology, few resources provide morph-level annotation. Wiktionary (2025) includes occasional etymologies for affixes and bound morphs, but coverage is sparse and inconsistent. Similarly, the *Czech Etymological Dictionary* (Rejzek, 2019) contains etymological information for several dozen affixes but focuses primarily on whole words.

Another relevant source is the open dataset by Goldberg (2019), which lists morphs with basic linguistic information. Around 1,500 entries include etymological notes—mostly for technical affixes such as *acet-*, *agri-*, *exo-*, or *-ant*—derived mainly from Greek or Latin. However, the dataset lacks detailed source references and provides only minimal etymological depth.

Overall, these resources offer isolated examples of morph-level etymology but lack the consistency and structure required for systematic computational modelling. Our work aims to fill this gap by developing a comprehensive, linguistically grounded dataset of Czech morphs with etymological labels.

4. Data

This section describes the linguistic resources, annotation process, and datasets used for training and evaluating our models.

4.1. Czech Etymological Dictionary

The latest Czech Etymological Dictionary (Rejzek, 2019) contains over 11,000 main entries and around 21,000 derived forms.

Its digital version, the Czech Etymological Lexicon (**CzEtyL**) (Rejzek et al., 2025) reformats this information into a structured, machine-readable format. Each entry specifies whether a word is native or borrowed and, for borrowed words, lists the sequence of languages through which it passed before entering Czech.

Example entry:

```
architekt   deu,lat,ell Loan
```

The word *architekt* ("architect") comes from Greek, entering Czech via Latin and German.

We extended CzEtyL using derivational relations from DeriNet (Olbrich et al., 2025), a large Czech lexical network linking derived words with their bases. For each CzEtyL entry, we added all lexemes from the same DeriNet tree and assigned them the same etymological label, expanding the coverage from about 10,500 to over 500,000 annotated words.

4.2. Our Annotation

We used the Czech part of the morpheme segmentation dataset from (SIGMORPHON, 2022), which

provides sentences segmented into morphs. Each morph was automatically classified as *root*, *derivational*, or *inflectional affix* using the model of John (2024), and then manually annotated with its sequence of language origins based mainly on Rejzek (2019) and Wiktionary (2025).

Example: *Šetřete peníze netelefonujte faxujte.*
Translation: *Save money, don't call, fax instead.*

- **Šetřete** (*save*)
  - *Šetř* — `R` — `ces`
  - *e* — `D` — `ces`
  - *te* — `I` — `ces`
- **peníze** (*money*)
  - *peníz* — `R` — `deu,lat`
  - *e* — `I` — `ces`
- **natelefonujte** (*don't call*)
  - *ne* — `D` — `ces`
  - *tele* — `R` — `ell`
  - *fon* — `R` — `ell`
  - *uj* — `D` — `ces`
  - *te* — `I` — `ces`
- **faxujte** (*fax*)
  - *fax* — `R` — `eng,lat`
  - *uj* — `D` — `ces`
  - *te* — `I` — `ces`

The annotation distinguishes three morph types: roots (`R`), derivational (`D`), and inflectional affixes (`I`). Language origins use ISO 639-3 codes such as `ces` (Czech), `deu` (German), `lat` (Latin), and `ell` (Greek).

The dataset was annotated by two native Czech speakers: a university student with a Bachelor's background in NLP and etymology, and an IT university student without prior formal linguistic training. Both annotators received detailed guidelines and a short calibration on example sentences, as well as access to resources such as the Czech Etymological Dictionary and Wiktionary (2025). Although neither is a professional linguist in this field, their work provides a useful reference point for human performance on this task.

Cohen's kappa was 0.82, and the exact match rate reached 95.9 %, indicating strong agreement. Remaining differences mostly stemmed from varying annotation granularity (e.g., `gmh` vs. `deu`) or inconsistencies in etymological sources.

4.3. Training and Evaluation Data

The dataset was then split into training, development, and test sets. Table 1 shows their sizes in terms of sentences, words, and morphs.

Dataset	Sentences	Words	Morphs
Train	200	2,774	7,016
Dev	50	599	1,460
Test	50	609	1,485

Table 1: Size of the annotated dataset used for training, development, and testing

We drew the training set from the Czech SIGMORPHON (2022) Shared Task training data (200 sentences). We derived the development and test sets from the 500-sentence development file by selecting every 10th sentence for testing and an every 10th sentence for development (offset by 1), leaving the remaining sentences unused for possible future work. Dataset size is limited by the cost of manual annotation, which requires linguistic expertise or substantial background research.

The training data contains about 7,200 annotated morphs, but most fall into just a few common etymological classes (Table 2). The etymology column lists language codes using the ISO 639-3 standard. As expected, the majority of morphs are of Czech (native) origin. Since the training set covers only 200 sentences, it should be seen as a small sample and not a fully representative picture of Czech in general.

Rank	Etymology	Count	% of total
1	ces	6,229	87.7
2	lat	280	3.9
3	ell	100	1.4
4	deu,lat	84	1.2
5	eng,lat	41	0.6
6	lat,ell	38	0.5
7	ita,deu	38	0.5
8	eng	36	0.5
9	deu,lat,ell	29	0.4
10	fra,lat	26	0.4
–	**Rest**	299	4.2

Table 2: Top 10 most frequent etymological sequences in the training data.

5. Proposed approaches to Morph-Level Etymology Prediction

5.1. Baselines

To evaluate performance, we define several baselines to establish a lower bound on expected results and assess how much the model improves beyond simple heuristics. These include always predicting Czech, a memorization approach, methods using *CzEtyL*, and a large language model baseline.

Always-native This baseline predicts all morphs as native (ces). It serves as a simple reference point and performs well when native morphs dominate the dataset.

Memorization We store all training morphs and predict their most frequent etymology label. Unseen morphs default to the majority class (ces). This baseline fails on ambiguous forms whose origin depends on context. For example, the suffix *-um* in *muzeum* ('museum') is of Latin origin, while the *um* in *rozum* ('reason, mind') is native Czech, derived from the verb *umět* ('to be able to').

Word lemmatization Each word is lemmatized with *MorphoDiTa* (Straková et al., 2014) and the lemma is looked up in the extended *CzEtyL* (4.1). The retrieved etymology is assigned to the root morph(s). Non-root morphs are matched against a *CzEtyL*-derived list of borrowed affixes; unseen affixes and roots default to ces, and inflectional endings are always predicted as ces.

5.2. Feature-based Classification

To predict the etymological origin of each morph, we extract features from annotated data that capture the characteristics of the morph and its context. These features serve as input to the classification model. We use character n-grams, morphological classification, position within the word, and embeddings, among others. They serve as input to the classification models used in our experiments.

We evaluate the following standard classifiers,[2] Logistic Regression (LR), Support Vector Machines (SVM), and Multi-Layer Perceptrons (MLP).

Character n-grams Character n-grams capture language-specific letter patterns and help the model associate certain combinations with particular etymological origins. We use 1-grams and 2-grams, which are appropriate given the average morph length of about 2.2 characters.

Morph Types Each morph is categorized as *Root*, *Derivational affix*, or *Inflectional affix*, encoded as a one-hot vector. Positional classification is also used, labelling morphs as *Root*, *Prefix*, *Suffix*, or *Interfix* depending on their position relative to the root(s) in a word. This classification is obtained using a model from John (2024).

Vowel Start and End This feature encodes whether a morph starts or ends with a vowel, using two binary values. Some languages favour vowel-initial or vowel-final patterns, so even such a simple feature can capture useful language-specific tendencies.

It becomes especially informative when combined with morph type. Many Proto-Indo-European roots, for example, follow a consonant-vowel-consonant (CVC) structure (Gamkrelidze and Ivanov, 1995). Affixes, by contrast, often attach more smoothly when starting or ending with a vowel. As a result, prefixes frequently end in vowels, while suffixes often begin with them.

Impact of Embeddings We use FastText embeddings (Bojanowski et al., 2017), which represent each word as the sum of its character n-gram embeddings. This method enables generating vectors even for morphs that are not stand-alone words. We use Czech FastText embeddings trained on Wikipedia and Common Crawl data (Grave et al., 2018).

We experiment with two variants: using embeddings directly for individual morphs, and using embeddings of the entire words in which the morphs occur.

5.2.1. Representation of Target Classes

We model target classes using two strategies:

- **Whole-sequence classification:** Treats each language sequence as a single label, predicting only sequences seen in training.
- **Multi-label classification:** Treats each language independently, allowing prediction of unseen combinations by training a binary classifier per language.

5.2.2. Data Augmentation

Beyond the original training set, we experiment with expanding the training material using two additional sources: data from the CzEtyL lexicon and automatically self-labeled data produced by our own model.

Use of Data from CzEtyL We extracted a dictionary of roots and affixes from *CzEtyL* with their probable etymological origins to serve as additional training data for the supervised models. The extracted root and affix dictionaries together contain roughly 12,000 morphs, covering about 400 distinct language sequences and 67 individual languages. These dictionaries are not perfect, as the root dictionary was constructed by assigning each root the most frequent etymology found among the words containing it, which can occasionally introduce errors.

[2]We use the implementations from https://scikit-learn.org.

Self-Training Self-training is a semi-supervised approach where a model is first trained on a manually annotated dataset, then used to label a larger unlabelled corpus. The newly labelled examples are added to the training set to improve the model. This is useful when additional annotated data is costly or difficult to obtain.

In this work, we applied self-training using a portion of the Prague Dependency Treebank (PDT) (Hajič et al., 2020). The selected subset contains about 88,000 sentences and approximately one million running words. Before processing, the text was normalized by lowercasing and removing punctuation, numbers, and special characters.

5.3. Using a Large Language Model

We evaluated several LLMs in a prompt-only setup for morph-level etymology annotation, using the same instruction format across models. The models were asked to label Czech morphs with ISO 639-3 origin codes, including any intermediary donor languages, while preserving the input structure and morph-type tags. We tested `GPT-5.2`, `GPT-5-mini`, `GPT-4o-mini`, `Claude-Opus-4.5`, `Gemini-3-Pro`, and `Deepseek-Chat-v3.1`.

We tested multiple prompt variants and selected the one that produced the most consistent, automatically parseable outputs. The final instruction prompt defines the required output format, annotation rules, and the meaning of ISO 639-3 language codes and morph-type tags.

All models were queried via the OpenRouter API[3] in batches of roughly 350 morphs (rounded to full sentences, typically $\sim$10 sentences). Each batch was processed in a fresh chat/session, i.e., no context beyond the instruction prompt and the current batch was present. We varied the amount of in-context training material from zero-shot up to the full training set (0, 50, 200, 2,000, and $\sim$7,000 training morphs), and report results under these settings.

The full prompt, API scripts, and data are released with the code (link hidden for anonymous review). We used temperature $= 0$ and otherwise default API settings to support reproducibility.

6. Evaluation Methodology

The task is to predict, for each morph, the sequence of languages through which it entered Czech, including any intermediate stages of borrowing. Although this sequence represents a historical progression, we treat and evaluate it as an unordered set for modelling purposes. The chronological order can later be reconstructed using relatively reliable rules based on known contact periods between the respective languages.

We use the F1-score to evaluate prediction quality. It is computed for each morph occurrence and then averaged across all morphs. To account for class imbalance between native and borrowed morphs, we also compute separate F1-scores for these two groups (**Bor.**, **Nat.**). It provides a clearer view of model performance. Borrowed morphs are generally harder to classify due to diverse sources and complex borrowing paths.

Many morphs appear multiple times in the dataset, especially inflectional endings. In the training set, there are 7,205 total morph instances, with 972 distinct morphs. The top 10 morphs account for 1,965 occurrences (27 %); the top 20 cover 2,864 (40 %); and the top 50 make up 3,997 occurrences—over half the dataset.

To reduce bias from frequent morphs, we also report an additional metric (**Dst.**): for each distinct morph, we first compute its average F1-score across all its occurrences and then take the mean of these per-morph scores to obtain the final value.

We also report the relative error reduction (**RER**) of the F1 score over the all-native baseline to better illustrate each model's improvement.

7. Results

We use the development set for model selection and report final numbers on the held-out test set. For the supervised models, we tuned hyperparameters with a grid search over feature and classifier settings. For prompt-only LLMs, we compared a small set of prompting configurations, varying the prompt itself, in-context training size and batch size.

7.1. Learning models: base results

Table 3 reports development-set performance of the supervised classifiers. All models use *single-label* prediction, treating each full etymological sequence as one class. The SVM uses an RBF kernel, and in the MLP model names the number indicates the hidden-layer size (number of neurons).

Both logistic regression and SVM struggled on borrowed morphs, although SVM achieved the best score on native ones. All MLP variants performed better overall, and results were very similar across hidden-layer sizes. In our experiments, adding extra hidden layers did not yield a consistent improvement.

7.2. Embeddings and Multi-label Setup

We evaluated several ways of integrating embeddings into the model, using representations of the

[3]We enabled OpenRouter's "do not train on my data" setting for all requests.

Model	F1	RER	Nat.	Bor.	Dst.
LogReg	94.0	39.7	98.7	51.4	86.2
SVM	94.7	46.5	**99.5**	50.7	88.1
MLP30	95.5	54.5	98.9	64.1	90.1
MLP100	**95.6**	**55.5**	99.0	**64.4**	**90.4**
MLP300	95.5	54.3	99.3	60.5	90.2

Table 3: Performance on the development set. RER = relative error reduction over the All-native baseline; Nat. / Bor. = F1 on native / borrowed morphs; Dst. = average F1 over distinct morphs.

morph itself, the word in which it appears, and their combination. The experiments were run across multiple model configurations. Overall, results were very similar across these settings, suggesting that embeddings did not lead to a substantial improvement. Table 4 reports representative results for MLP30 on the development set.

We also tested a multi-label setup, where each language is predicted independently. Instead of selecting a single etymological sequence, the model predicts which languages belong to the etymological path of a given morph. This increases flexibility and allows combinations not seen in training. In practice, the setup trains one small classifier per language, effectively increasing total capacity, which is particularly useful when additional training data is available.

In Table 4, we compare multi-label models trained on the base data with models extended by roots and affixes from *CzEtyL*. In the model names, **M** denotes the multi-label setup and **C** marks the inclusion of *CzEtyL* data. Multi-label prediction alone did not improve performance on the base training set, but it brought clear gains when combined with *CzEtyL*-derived data. Increasing the hidden-layer size provided no additional benefit.

7.3. Prompt-only LLM results on the development set

Table 5 summarizes prompt-only development-set results with the full in-context training set. A consistent pattern across models is that native-morph performance is near ceiling, so most differences come from how well the models handle borrowed morphs.

`GPT-4o-mini` stays close to the all-native baseline because it rarely identifies borrowed morphs correctly, even when given many in-context examples. A likely reason is that this smaller general-purpose model has limited capacity for the fine-grained decisions needed for short, often ambiguous morph strings, so it tends to fall back to the dominant native label.

Most models benefit from additional in-context training material. For example, `GPT-5.2` improves from 94.7 % F1 in the zero-shot setting to 95.5 % (50), 96.0 % (200), 96.3 % (2,000), and 96.8 % with the full training set (∼7,000 morphs).

In contrast, `Gemini-3-Pro` is already strong in the zero-shot setting (95.9 % F1) and shows no clear improvement with added context, reaching 95.8 % F1 with the full training set.

Setting	F1	RER	Nat.	Bor.	Dst.
Effect of embeddings - MLP30 model					
No Embedd.	95.5	54.5	**98.9**	64.1	90.1
Word	95.0	49.5	98.7	61.2	89.2
Morph	95.1	50.8	**98.9**	60.4	89.4
Word+Morph	95.3	52.5	98.7	64.2	89.4
Multi-label MLPs					
Without CzEtyL data					
MLP30 M	95.3	52.8	98.8	64.1	89.7
MLP100 M	95.1	51.1	98.8	61.6	89.3
MLP300 M	95.0	49.8	**98.9**	59.6	89.1
With CzEtyL data					
MLP30 MC	**95.8**	**58.2**	98.8	69.5	**90.9**
MLP100 MC	95.6	56.1	**98.9**	66.0	90.7
MLP300 MC	**95.8**	57.3	98.4	**71.6**	90.6

Table 4: Development-set results: (i) effect of embeddings for MLP30, and (ii) multi-label MLP variants with and without additional CzEtyL-derived training data.

Model	F1	RER	Nat.	Bor.	Dst.
GPT-4o-mini	90.4	3.2	**99.9**	3.9	80.1
GPT-5-mini	96.4	63.6	99.6	67.0	92.3
GPT-5.2	96.8	67.4	99.1	**75.7**	93.5
Cld.-Opus4.5	**97.0**	**69.4**	99.4	74.9	**93.6**
Gemini-3pro	95.8	58.2	99.1	66.2	91.1
DeepSk-v3.1	95.6	55.8	99.5	60.6	90.8

Table 5: Development-set prompt-only LLM results with the full in-context training set (∼7,000 morphs).

7.4. Results on Test Set

We now evaluate all models on the test set, including the baselines, learning-based approaches, and the prompt-only LLM setup with the full training set provided in-context. Table 6 summarizes the results.

7.5. Discussion of Results

Prompt-only LLMs With the full training set provided in-context, the prompt-only LLMs reach strong performance overall (Table 6). `Claude-Opus-4.5` is the top performer in this setting, closely followed by `GPT-5.2`. `GPT-5-mini` and the other non-OpenAI models (`Gemini-3-Pro`, `Deepseek-Chat-v3.1`) form a second tier with

Model	F1	RER	Nat.	Bor.	Dst.
Baselines					
All native	89.9	–	**100.0**	0.0	77.1
Memor.	94.4	44.9	99.3	51.5	86.3
Lemmatiz.	94.8	48.1	98.6	60.7	90.0
Learning models					
MLP30	95.2	51.6	98.6	65.2	88.3
MLP100 M	95.8	57.4	99.1	66.2	90.1
MLP300 MC	96.2	62.0	98.9	72.7	90.8
MLP30MC	96.5	65.6	99.6	69.6	91.5
Self-train	96.1	60.8	98.6	72.8	91.0
Prompt-only LLMs (full train set in-context)					
Cld.-Opus4.5	**97.8**	**77.7**	99.7	**80.4**	**94.7**
DeepSk-v3.1	96.5	65.6	99.6	69.0	91.9
Gemini-3pro	96.5	65.8	99.7	68.4	91.8
GPT-4o-mini	90.9	9.4	99.9	10.1	79.4
GPT-5-mini	96.4	64.0	99.8	66.0	92.5
GPT-5.2	97.6	76.6	99.8	78.6	94.4

Table 6: Final results on the test set. Prompt-only LLMs use the full training set in-context.

similar results. In contrast, `GPT-4o-mini` lags far behind, and the gap is driven mainly by weak performance on borrowed morphs rather than on native ones, which are near ceiling for all models.

Baselines The trivial *All native* model achieved an F1 score of 89.9 %, which also reflects the proportion of native morphs in the dataset. The lemmatization approach, which draws on etymological data from CzEtyL, cut this error by nearly half. It outperformed the simpler memorization baseline.

Learning models The trends from the development set carried over to the test set. Using *CzEtyL*-derived training data and the multi-label formulation both helped, with the best supervised variants reaching around 96.5 % F1 (e.g., MLP30MC: 96.5).

In the self-training setup (5.2.2), we used the best development-set model (MLP30MC) to label the large corpus and then retrained on these pseudo-labels; the retrained model achieved similar results (96.1 % F1), correcting some morphs but introducing new errors, so overall it brought no clear improvement over the fully supervised approach.

Performance by Morph Type Since borrowing affects morph types differently, we also break down performance by *root*, *derivational affix*, and *inflectional affix*. This follows the standard intuition behind borrowability scales: lexical material and derivational morphology tend to transfer more readily than inflectional endings, which are often more resistant in ordinary contact situations. In Czech in particular, inflectional endings are overwhelmingly native, so strong scores on inflectional morphs are expected.

Model	F1	Root	Deriv.	Infl.
All native	89.9	79.6	93.1	**100.0**
Memor.	94.4	87.7	97.0	**100.0**
Lemmatiz.	94.8	91.9	94.3	**100.0**
MLP30MC	96.2	92.1	97.6	**100.0**
GPT-4o-mini	90.7	81.3	93.5	**100.0**
GPT-5-mini	95.0	89.9	96.7	99.7
GPT-5.2	97.5	94.6	98.5	**100.0**
C-Opus4.5	**97.8**	**95.1**	**98.8**	**100.0**

Table 7: F1 by morph type on the test set. For LLMs we used the full training set in-context.

8. Conclusion and Future Work

We addressed the task of predicting the etymological origin of morphs in Czech, determining whether each morph is native or borrowed and, if borrowed, which languages are involved in its borrowing path. Because most resources focus on word-level etymology, there are currently no widely used datasets providing morph-level etymological annotation for segmented words. To fill this gap, we manually annotated ∼10,000 morphs (9,961) in 300 Czech segmented sentences (3,982 words) from the Shared Task on Morpheme Segmentation by SIGMORPHON (2022). The resulting dataset is publicly available at `https://github.com/ampapacek/MorphemeOrigin`.

We compared three families of approaches: simple rule-based baselines, supervised classifiers, and prompt-only LLM predictions. The supervised models use character n-grams together with structural cues (morph type/position and simple vowel-shape features). Among supervised systems, the strongest learning-based model was a multi-layer perceptron with a single hidden layer of 30 neurons, trained in a multi-label setup and augmented with *CzEtyL*-derived entries. It substantially outperformed the rule-based baselines. While it does not match the very best prompt-only LLM setting, it remains attractive as a lightweight, deterministic, and inexpensive alternative that can be run locally without external API access or heavy computational resources. Semi-supervised self-training on the PDT corpus yielded no clear gains.

Overall, the best results come from prompt-only LLMs when the full training set is provided in-context. In this setting, `Claude-Opus-4.5` reaches 97.8 % F1. Across LLMs, adding more in-context training morphs generally improves performance.

Future work A straightforward next step is to enlarge the annotated dataset, since more examples expose models to a broader range of morphs and borrowing patterns. We also observed that scaling supervision helps: increasing in-context training improves LLM performance, and adding more labeled data benefits the supervised models as well. The annotation could be made more fine-grained, for example by distinguishing historical stages of Latin, German, or Greek, or by marking whether a morph is pan-Slavic or specific to Czech.

For learning models, richer shape-based features may help. For example, the vowel-start/end feature was useful, and encoding the full consonant–vowel (CV) pattern of each morph could capture additional regularities. On the model side, more structured architectures that predict language sequences directly, as well as ensembles combining complementary systems, could improve robustness. Another direction is to revisit semi-supervised methods such as self-training, which did not yield clear improvements in our current setup. Finally, it would be interesting to adapt pretrained models to this task (e.g., by fine-tuning) and to more systematically explore training and inference settings, including hyperparameters and prompting choices.

Although this study focused on Czech, the methodology is not language-specific and could be applied to other morphologically rich languages.

9. Acknowledgements

This work has been supported by the project TQ12000040 (CZDEMOS4AI) financed by the Technology Agency of the Czech Republic (www.tacr.cz) within the Sigma 3 Program, by the Charles University Research Centre program No. 24/SSH/009, and by the Czech Science Foundation project No. 26-21822S. This work has been using data and tools provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062). We also thank anonymous reviewers for their useful comments.

10. Bibliographical References

Peter Arkadiev, Francesco Gardani, and Nino Amiridze. 2015. *Borrowed Morphology: An Overview*, pages 1–23. De Gruyter Mouton.

Mark Aronoff and Kirsten Fudeman. 2011. *What is Morphology?* Fundamentals of Linguistics. Wiley.

Khuyagbaatar Batsuren, Gábor Bella, Aryaman Arora, Viktor Martinovic, Kyle Gorman, Zdeněk Žabokrtský, Amarsanaa Ganbold, Šárka Dohnalová, Magda Ševčíková, Kateřina Pelegrinová, Fausto Giunchiglia, Ryan Cotterell, and Ekaterina Vylomova. 2022. The SIGMORPHON 2022 Shared Task on Morpheme Segmentation. In *19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 103–116, Stroudsburg, PA, USA. Association for Computational Linguistics.

James P Blevins. 2016. *Word and paradigm morphology*. Oxford University Press.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Philip Durkin. 2014. *Borrowed Words: A History of Loanwords in English*. Oxford University Press.

Thomas V. Gamkrelidze and Vjaceslav V. Ivanov. 1995. Chapter Four: The Structure of the Indo-European Root. In *Indo-European and the Indo-Europeans: A Reconstruction and Historical Analysis of a Proto-Language and Proto-Culture*, pages 185–187. De Gruyter Mouton, Berlin, Boston.

Francesco Gardani. 2008. *Borrowing of Inflectional Morphemes in Language Contact*. Phd thesis, University of Vienna.

Anthony P. Grant. 2020. *The Oxford Handbook of Language Contact*. Oxford University Press.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Martin Haspelmath. 2020. The morph as a minimal linguistic form. *Morphology*, 30(2):117–134.

Lars Johanson and Martine Robbeets. 2013. Copies versus cognates in bound morphology. *STUF – Language Typology and Universals*, 66(2):130–135.

Vojtěch John. 2024. Morph classifier. Master's thesis, Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics (ÚFAL), Prague, Czech Republic.

Viktória Kovácsová. 2025. Automatic detection of lexical borrowings. Master's thesis, Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics (ÚFAL).

Oxford English Dictionary. 2025. school, n. https://www.oed.com/dictionary/school_n1. Oxford University Press, accessed 13 Oct 2025.

Jiří Rejzek. 2019. *Český etymologický slovník*. 3rd edition. LEDA.

Frank Seifart. 2015. Direct and indirect affix borrowing. *Language*, 91(3):511–532.

Wiktionary. 2025. Wiktionary: The free dictionary. https://www.wiktionary.org/. Accessed: 2025-10-04.

Winston Wu and David Yarowsky. 2020. Computational etymology and word emergence. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3252–3259, Marseille, France. European Language Resources Association.

11. Language Resource References

Colin Goldberg. 2019. Morphemes dataset. https://github.com/colingoldberg/morphemes. Accessed: 2025-10-13.

Jan Hajič, Eduard Bejček, Jaroslava Hlaváčová, Marie Mikulová, Milan Straka, Jan Štěpánek, and Barbora Štěpánková. 2020. Prague Dependency Treebank – Consolidated 1.0. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 5208–5218, Marseille, France.

Michal Olbrich, Viktória Brezinová, Šárka Dohnalová, Vojtěch John, Lukáš Kyjánek, Aleš Papáček, Emil Svoboda, Magda Ševčíková, Jonáš Vidra, and Zdeněk Žabokrtský. 2025. DeriNet 2.3. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Jiří Rejzek, Aleš Papáček, Viktória Brezinová, and Zdeněk Žabokrtský. 2025. Czech Etymological Lexicon 1.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

SIGMORPHON. 2022. SIGMORPHON 2022 Shared Task on Morpheme Segmentation. https://github.com/sigmorphon/2022SegmentationST. Accessed: 2025-10-04.

Jana Straková, Milan Straka, and Jan Hajič. 2014. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18, Baltimore, Maryland. Association for Computational Linguistics.

A. Etymological Analysis of a Diachronic Czech Legal Corpus

In this appendix, we apply the best lightweight supervised system described in the main paper to a diachronic corpus of Czech legal texts and examine how the predicted proportions of native and borrowed morphs change over time. The goal of this analysis is not to provide a definitive historical account, but rather to illustrate one possible downstream use of the proposed morph-level etymology framework on a long real-world time series.

The corpus was constructed from publicly available Czech legislative texts obtained through the e-Sbírka API. We downloaded texts year by year and constructed one sampled file for each year by randomly sampling sentences up to approximately 1,000 words. The resulting dataset contains 109,257 words. The corpus was created within the CZDEMOS4AI project. Because it uses a fixed yearly word budget and includes only the texts available at the time of collection, it should be interpreted as a sample of legal language in each year rather than as a complete picture of legal language over the full period.

We then automatically segmented the corpus and applied the classifier from the main paper to the resulting morphs. For this analysis, we used the same MLP configuration as in our best lightweight system: a single hidden layer of size 30, multi-label prediction, and extended training data derived from the etymological dictionary. On the original test set in the main paper, this model achieves 96.2% F1.

Several caveats should be kept in mind. First, the legal corpus belongs to a different domain than the data used for evaluation in the main paper, so the model's accuracy on legal texts may be lower. Second, both the morphological segmentation and classification were produced automatically, which means that some errors may already arise in the preprocessing stage. The figures and observations presented below should therefore be interpreted as exploratory rather than definitive.

In Table 8, we summarize the main predicted origin labels, including the most frequent borrowed source languages. The values show what percentage of morphs is assigned to each label. Native is the dominant label overall, while Latin is the dominant borrowed source in every available year of

Origin	Mean	Median	Max	Std. dev.
Native	88.12	88.22	93.03	2.69
Latin	5.58	5.29	9.97	1.49
English	1.67	1.62	3.21	0.41
German	1.41	1.31	2.54	0.47
French	1.14	1.08	2.79	0.44
Greek	1.10	1.00	3.73	0.55
Polish	0.33	0.30	1.61	0.19
Italian	0.22	0.17	0.88	0.18

Table 8: Summary of the main predicted origin labels across the years 1918–2026. Shares are computed with respect to all predicted labels within each year, and standard deviation is computed across yearly shares. Labels are sorted by mean yearly share. All values are in (%).

the sample. Other prominent borrowed sources include English, German, French, and Greek, but all of them remain clearly below Latin in the aggregate picture.

Across all sampled years, the mean predicted share of borrowed morphs is 8.96%. In the earliest available year, 1918, it is 6.03%, while the maximum in the current sample is observed in 2015, where it reaches 15.25%. Figure 1 shows the five-year weighted smoothing[4] of the overall borrowed share together with fitted linear trend lines for the overall borrowed share and for the three most prominent borrowed source languages. Overall, the fitted trend line suggests a slight upward trend in the proportion of borrowed morphs over time, corresponding to roughly 2.5 percentage points per century.

The code, data, and generated outputs for this appendix are available at: https://github.com/ampapacek/MorphemeOrigin/tree/main/legal_corpus_analysis.

[4] Each year's value is replaced by a weighted average over a five-year window centered on that year, using weights 0.1, 0.2, 0.4, 0.2, and 0.1.

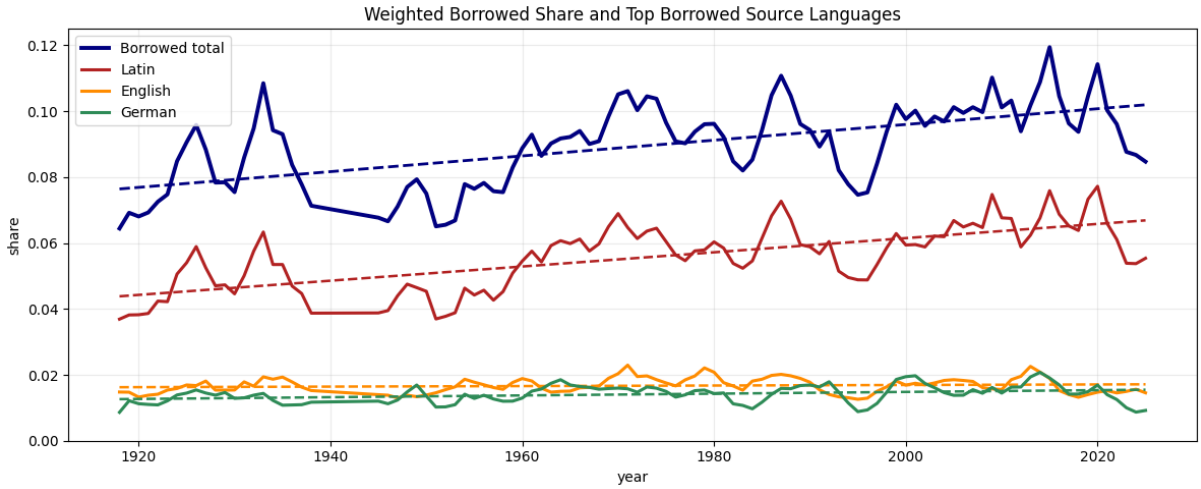


Figure 1: Five-year weighted smoothing of the overall predicted borrowed-morph share together with fitted linear trend lines for the overall borrowed share and for the three most prominent borrowed source languages: Latin, English, and German.

Uncovering Work from Words: LLM-Based Information Extraction from Historical Petitions

Ellinor Lindqvist, Eva Pettersson, Joakim Nivre
Uppsala University
Dept. of Linguistics and Philology
firstname.lastname@lingfil.uu.se

Abstract

We investigate the extraction and normalisation of phrases describing work from 18th-century Swedish petitions using four LLMs: GPT-4o, Llama-3 70B/8B, and Mixtral-8x7B. Performance is evaluated across four configurations: isolated extraction, isolated normalisation, a staged pipeline, and a combined multitasking setup, using both full and filtered texts (with formal greetings and closing sections removed). While exact phrase matching remains low ($F1 < .10$), token-level and semantic similarity scores suggest that models consistently locate relevant topical regions. Semantic similarity scores must however be interpreted with caution, since they are often only marginally higher than an average baseline. Results reveal a "multitasking paradox": combined extraction and normalisation improves phrase location for high-parameter models but degrades normalisation precision. Furthermore, normalisation benefits from the context of a staged pipeline compared to isolated tasks, while text filtering has only marginal effects. Despite a tendency towards over-prediction, qualitative analysis suggests that models can detect plausible work-related expressions missed by human annotators. These findings illustrate the challenges of historical extraction and suggest that hybrid human–machine workflows are a promising approach for enhancing coverage and interpretability in cultural heritage research.

Keywords: Information Extraction, Petitions, Large Language Models, Historical Swedish, Historical NLP, Digital Humanities, Text Normalisation

1. Introduction

The expanding field of digital humanities increasingly relies on advanced computational methods to analyse digitised historical sources. Although such sources are becoming more accessible, they remain challenging for automated processing due to linguistic variation, inconsistent orthography, and specialised historical contexts. Large language models (LLMs) show strong capabilities for complex language understanding, yet their suitability for domain-specific information extraction in structured historical texts, such as petitions, is still underexplored. In contrast to traditional encoder-based transformer models (e.g., BERT), which typically require extensive task-specific fine-tuning on annotated data, LLMs provide a more flexible few-shot alternative. This is especially beneficial in historical research, where large-scale labeled datasets are often scarce, enabling complex information extraction without the need for substantial model training.

This paper evaluates the ability of LLMs to extract and normalise phrases describing working activities in historical Swedish petitions. Extraction identifies where work is mentioned, while normalisation produces modernised and standardised forms of historically variable expressions. These normalised forms can support historians and digital humanists to analyse work and occupations across many documents and time periods. Automatically identifying information about how individuals supported themselves offers valuable insights into economic life and social relations in the past. We draw on annotated petitions from the Gender and Work (GaW) project at Uppsala University,[1] using a subset of this resource to test extraction and normalisation methods and to propose robust evaluation approaches for natural language processing (NLP) applied to historical texts.

Furthermore, early modern European petitions followed a classical rhetorical structure, offering cues that may assist automated analysis. As a secondary objective, we therefore examine whether rhetorical segmentation, for example removing the opening section (*Salutatio*) and the closing section (*Conclusio*), can enhance extraction performance.

Hence, our research objectives are:

1. Evaluate LLM performance in extracting phrases that describe work from Swedish historical petitions under different task settings.
2. Assess the impact of rhetorical segmentation by comparing full-text input with filtered versions.

This study contributes to our understanding of how LLMs can support historical research, demonstrating their potential and limitations for fine-grained information extraction in low-resource, structurally diverse corpora. We provide insights for digital humanities and NLP research on applying LLMs to complex historical documents and enabling more scalable analysis of historical texts.

[1] https://gaw.hist.uu.se/petitions/

Proceedings of LT4HALA 2026 @ LREC 2026, pages 19–37.
11 May 2026. *©ELRA Language Resources Association (ELRA), 2026*

2. Background

Digital methods have become increasingly central to the study of historical sources, enabling large-scale discovery and analysis of cultural heritage materials that would be difficult to achieve manually (Wilson, 2022; Burrows, 2023; Wijfjes, 2017). Yet historical texts pose specific challenges for NLP, including substantial orthographic variation, diachronic linguistic change, and limited standardised resources (Piotrowski, 2012; Bollmann, 2019). Annotated training data for domain-specific tasks are also scarce and expensive to produce, limiting the feasibility of supervised adaptation or fine-tuning in many archival contexts (Krusic, 2024; Hiltmann et al., 2025).

The emergence of LLMs has opened new possibilities in digital humanities. For example, LLMs have been applied to detect irony in 19th-century Spanish newspapers (Cohen et al., 2025) or to transcribe handwritten archival documents more quickly and accurately than traditional HTR systems (Humphries et al., 2025). While these models offer strong zero- and few-shot capabilities, especially valuable in low-resource historical settings, they also introduce the risk of hallucination and domain-inaccurate outputs. For instance, Tudor et al. (2025) show that LLMs may misclassify or invent entities in zero-shot named entity recognition on historical text, whereas Hiltmann et al. (2025) demonstrate that domain-aware prompting can reduce such errors and better align model predictions with historical conventions.

Some historical documents, such as petitions, also exhibit structured rhetorical organisation that can support automated analysis. In much of premodern Europe, petitions followed a classical rhetorical structure, typically comprising five or six sections (Hansson, 1988; Sokoll, 2006; Israelsson, 2016):

1. *Salutatio*: Formal salutation to the addressee.
2. *Exordium*: Brief appeal to authority, goodwill, or capacity to help.
3. *Narratio*: Background and supporting reasoning, often including (4. *Argumentatio*).
5. *Petitio*: Explicit request or plea being made.
6. *Conclusio*: Closing courtesies, often including a signature.

Recent work on rhetorical annotation of 18th-century Swedish petitions (Lindqvist et al., 2025) shows that LLMs can reliably identify the *Salutatio*, *Petitio*, and *Conclusio*, with F-scores often in the high 90s. Such automated segmentation supports focused analysis of specific parts of historical documents, which could potentially make information extraction tasks more precise.

However, it remains uncertain whether rhetorical segmentation provides measurable benefits for downstream extraction tasks. In this work, our primary focus is on assessing how effectively LLMs can identify and normalise phrases describing work in historical petitions, while also exploring segmentation as a secondary factor that may influence performance.

3. Data

3.1. The Gender and Work Project

This study uses material from the Gender and Work (GaW) project at Uppsala University,[2] which investigates how people supported themselves in Early Modern Sweden (c. 1550–1800). The project gathers evidence of work-related activities from diverse archival sources, such as court records, account books, and personal letters, and stores the information in a research database. Each entry consists of a transcribed source document and metadata describing who performed an activity, what was done, when and where it occurred, and in what context, along with archival source information and, in many cases, images of the original documents to support historical transparency and traceability.

A central part of the project's method is the verb-oriented approach (Fiebranz et al., 2011), in which historians identify and extract expressions that describe work-related activities. These activities are often, but not exclusively, realised as verb phrases (e.g., "sell clothes", "fish herring").

Because work-related activities are context-dependent and not always expressed through standard verb-phrase structures (e.g., a verb and its object), the resulting annotations sometimes include fragmentary expressions, prepositional phrases, or nominal structures. This reflects the historical realities of expression rather than prescriptive grammar. To standardise interpretation for analysis, GaW annotators manually reformulate these expressions into modern Swedish infinitive forms (Ågren et al., 2023). Table 1 provides examples of this process, illustrating the shift from archaic, historical phrases to normalised, modern verb phrases, often using an infinitive verb followed by a direct object.

For computational processing, however, this creates a non-trivial modelling challenge. Work activities must be detected not only at the lexical level, but also semantically and contextually: for instance, the Swedish verb "köpa" (*to buy*) may indicate commercial trade in one context but personal consumption in another. Combined with substantial orthographic variation in Early Modern Swedish, this makes the GaW data both historically rich and

[2] https://www.uu.se/en/research/gender-and-work

Historical Phrase	Manual Normalisation
I lånebanquen à 4 1/2 proc:t *(In the Loan Bank at 4 1/2 percent)*	låna ut pengar till bank *(to lend money to a bank)*
Förteckning på mitt husfolck *(Register of my household members)*	lämna mantalsuppgift *(to submit a census return)*
blifwit föranlaten at söma åt andra *(been obliged to sew for others)*	sy kläder *(to sew clothes)*
Till bränuinstillwärkning åtager iag mig *(For liquor production I undertake)*	bränna brännvin *(to distill liquor)*
inbrotts stöld *(burglary theft)*	göra inbrott *(to commit burglary)*

Table 1: Examples of manual normalisation by GaW historians. English (approximate) translations are in italics.

Origin	Time Period	#Docs	#Phrases
Stockholm	1667–1734	57	340
Västmanland	1758	9	109
Örebro	1719–1800	44	84

Table 2: Test data statistics, including the number of documents and annotated phrases that describe work per collection.

valuable for research, while presenting complex challenges for automated analysis.

3.2. The Petition Dataset

This paper focuses on 18th-century Swedish petitions drawn from the GaW database (Section 3.1). The material comes from three collections: petitions to the County Administrative Boards of Västmanland and Örebro, and petitions to the Board of Trade in Stockholm. Two documents used during prompt development are excluded from evaluation. The final dataset comprises 110 petitions, with detailed statistics shown in Table 2.

4. Method

We evaluate LLM-based extraction of phrases describing work from historical petitions under two main conditions: using the full transcription and using a version where selected rhetorical sections are removed. Below, we describe the filtering procedure, extraction tasks, prompt design, and evaluation framework.

4.1. Rhetorical Filtering

To assess whether removing greetings and closing statements improves extraction performance, we create a filtered dataset where the *Salutatio* and *Conclusio* are removed. Since these sections predominantly contain formulaic courtesy expressions, we hypothesise that they rarely include work-related information.

Rhetorical segmentation is performed using Llama 3.3 70B (Touvron et al., 2023) following the methodology of Lindqvist et al. (2025). To ensure complete coverage, detected spans are projected back onto the original text using semi-soft alignment: for each annotated segment, we perform a case- and whitespace-insensitive search for the closest matching occurrence and extract the corresponding substring. This procedure successfully mapped the *Salutatio* in 96% and the *Conclusio* in 88% of the documents.

4.2. Models and Extraction Tasks

We evaluate four models selected to represent a diverse range of sizes, architectures, and licensing: two LLaMA open-weight transformers from Meta AI of different sizes, Llama 3.3 70B and Llama 3.1 8B (Touvron et al., 2023); the open-source model Mixtral 8x22B (Jiang et al., 2024) from Mistral AI; and GPT-4o from OpenAI (OpenAI, 2026). All models are accessed via APIs using identical prompts and inference settings across runs. All experiments use temperature 0 for deterministic and reproducible outputs. Four tasks are evaluated across both full and filtered datasets:

1. **Extraction:** Identify phrases describing work from the historical source text, preserving original wording.
2. **Isolated Normalisation:** Given gold phrases manually extracted by historians, normalise them to modern Swedish infinitive forms.
3. **Pipeline Normalisation:** Normalise phrases extracted by the model from the source text.
4. **Combined Extraction and Normalisation:** Simultaneously identify historical phrases and provide normalised equivalents.

You are an expert in linguistic analysis of Swedish historical texts. Your task is to analyse Swedish petitions from the 18th century and identify all phrases describing work, work performance, or work-related actions.

 # Main Instructions

1. **Read the text carefully.** The text may contain archaic spelling, free word order, and compound words.
2. **Extract only phrases describing work, work performance, or work-related actions.** This includes performing, contributing to, or participating in work, occupations, service, or duty for a livelihood, regardless of whether payment is given.
3. **Preserve the original text exactly.** Do NOT add, remove, or change words, spelling, or punctuation.
4. **Return a list of extracted phrases** in the order they appear in the text.

 # Output Format

Return only a JSON object with the key `"extracted_phrases"` and a list of phrases from the text. Return JSON exactly in this format, without extra text, comments, or code-block markers:

```
{ "extracted_phrases": ["grind some grain", "in service have been"] }
```

 # Examples

(Only short excerpts are shown in the examples. In the task, the entire text should be processed.)

Input 1: "/. . . / since at this time hops were so hard to be come by? Was answered she had bought hops from Tidö, whereto she had made a small addition of some Wormwood Sprigs, /. . . /"
Output 1: `{ "extracted_phrases": ["purchased hops"] }`

/.../

 # Identify and extract the phrases describing work in the following petition:

Figure 1: The prompt used for Task 1, translated to English, including role definition, extraction instructions, formatting constraints, and few-shot examples.

4.3. Prompting and Data Recovery

The pipeline uses few-shot prompting with twelve example snippets from the GaW corpus. All prompts follow a consistent structure: (1) role instruction positioning the model as an expert in linguistic analysis of Swedish historical texts, (2) explicit instructions emphasizing preservation of original features (archaic spelling, word order), (3) clear definition of phrases describing work as expressions describing performance of, contribution to, or participation in work for earning a living, and (4) step-by-step task instructions with examples. This shared structure is exemplified by the Task 1 prompt in Figure 1 (abbreviated to one of twelve few-shot examples), and all prompts and their English translations can be found in the Appendix. Output is requested in structured JSON.

To address API timeouts and parsing failures, we implemented a three-step recovery protocol. First, models were rerun on the specific input documents that previously yielded empty or failed outputs (up to two additional attempts). Second, for model outputs remaining unparseable, an *ad-hoc* regex-based script salvaged phrases from malformed or truncated JSON and pruned repetitive loops by collapsing phrase repetitions into single instances.

Model	#Docs	R1	+R	+Rep
GPT-4o	880	94.9%	98.7%	100%
Llama 70B	880	99.4%	99.5%	100%
Llama 8B	880	72.1%	73.1%	99.8%
Mixtral 8x22B	880	66.7%	92.0%	96.0%

Table 3: Cumulative parsing success rates across models and processing stages ($N = 880$; representing 110 documents across 4 tasks and 2 text modes). R1: after initial run; +R: after reruns (up to 2); +Rep: after ad-hoc regex repair parsing.

Task Type	Initial	Final
T1: Extraction	91.5%	100%
T2: Isolated Norm	80.1%	94.0%
T3: Pipeline Norm	86.8%	97.0%
T4: Combined Extr/Norm	75.4%	92.0%

Table 4: Parsing success rates per task, both initially and after reruns and ad-hoc parsing.

Finally, outputs from all runs were merged using a quality-priority system that prioritised valid parsed data over salvaged JSON. Documents for which no parseable output could be obtained after these steps were treated as empty outputs and assigned a score of 0.0 for all metrics to penalise model fragility.

The choice of JSON as the output format was intended to facilitate the direct integration of extracted data into downstream pipelines and to ensure clear pairing between historical and normalised phrases in combined tasks. While simpler formats (e.g., plain-text lists) might have reduced syntax errors, they might also have increased the risk of ambiguity when mapping multiple entities. Our evaluation of structural reliability (Table 3) shows that while GPT-4o and Llama 3.3 70B achieved near-perfect JSON adherence, the smaller models Llama 3.1 8B and Mixtral 8x22B were more fragile. Our recovery protocol, including reruns and ad-hoc regex repair parsing, significantly improved final parsing success, particularly for Llama 8B.

In addition, we find that rhetorical filtering has a negligible effect on the structural success of the output, where the average success rate was 83.1% averaged for all models after the first run and 98.9% after post-hoc parsing, compared to the full text version of 79.2% and 98.3%, respectively. The nature of the NLP task and requested output, however, is a significant predictor of model fragility, as can be seen in Table 4. While the more simple and straightforward extraction (Task 1) reached 91.5% initial success rate, the more complex combined extraction and normalisation task (Task 4) yielded only 75.4%. This likely reflects the increased dif-

ficulty of maintaining nested JSON syntax during simultaneous linguistic transformations.

4.4. Evaluation

To evaluate the correspondence between model-predicted and gold-standard work phrases, we apply a combination of **exact-match**, **string-level**, and **embedding-based** similarity metrics. This multi-level evaluation captures both surface-form correspondence and semantic adequacy.

Before evaluation, all phrases are lowercased and tokenised. Metrics are computed on a per-document basis and aggregated across the corpus for each prompt variant, enabling comparison of model performance under different input conditions (full text vs. rhetorically filtered) and task configurations.

Duplicate handling: Since the same work phrase may appear multiple times within a document, we preserve all instances during evaluation. For exact match metrics, if a phrase appears n times in gold and m times in predictions, we count $\min(n, m)$ as true positives. For alignment-based metrics (Levenshtein and embedding similarity), each phrase instance is matched independently, ensuring models are credited for extracting all occurrences of repeated phrases.

4.4.1. Exact Match Metrics

We compute **precision**, **recall**, and **F1-score** based on *exact string identity* at both phrase and token levels, counting true positives as $\sum \min(\text{count}_{\text{gold}}(p), \text{count}_{\text{pred}}(p))$ across unique phrase types p. Token-level metrics apply the same principle to individual tokens. Together, these provide complementary insights into fully versus partially correct outputs.

4.4.2. String-Level Similarity

To capture surface-level similarity beyond exact matches, we compute *normalised Levenshtein similarity* between gold and predicted phrases, allowing partial matches to be recognised despite small spelling or formatting differences. Levenshtein distance counts the minimum number of insertions, deletions, or substitutions to transform one string into another, converted to a similarity score in $[0, 1]$:

$$\text{Sim}_{\text{lev}}(g_i, p_j) = 1 - \frac{d(g_i, p_j)}{\max(|g_i|, |p_j|)},$$

where $d(g_i, p_j)$ is the character edit distance. A score of 1 indicates identical strings, 0 indicates maximal difference.

Since the number of predicted phrases may differ from the gold standard, we treat alignment as a matching problem and compute two complementary scores per document:

- **Gold-based (recall-oriented):** measures how well each gold phrase is recovered,

$$\text{LevSim}_{G \to P} = \frac{1}{|G|} \sum_{g_i \in G} \max_{p_j \in P} \text{Sim}_{\text{lev}}(g_i, p_j),$$

- **Prediction-based (precision-oriented):** measures how well each predicted phrase matches a gold phrase,

$$\text{LevSim}_{P \to G} = \frac{1}{|P|} \sum_{p_j \in P} \max_{g_i \in G} \text{Sim}_{\text{lev}}(g_i, p_j).$$

Phrases without a matching counterpart are assigned a similarity of 0.

4.4.3. Embedding-Based Semantic Similarity

To capture semantic similarity beyond surface form, we compute the cosine similarity between mean phrase embeddings, using optimal phrase-to-phrase alignment, allowing recognition of semantically equivalent phrases despite differing wording. Historical phrases use a domain-specific fastText model trained on diachronic Swedish text (Hengchen and Tahmasebi, 2021), specifically the incrementally trained vectors spanning the period 1740–1800 to best match our data. While transformer models are state-of-the-art for many tasks, we use static embeddings because our data consists of isolated phrases lacking sufficient context. A historical fastText model further helps align with older language and improves robustness to non-standard spelling by representing words through character-level n-grams. We here match the model's training constraints by removing non-Swedish characters and tokens of length two or smaller. Normalised phrases use a Swedish fastText model trained on Wikipedia (Grave et al., 2018). Both models use subword information for robust out-of-vocabulary handling. However, should there be phrases where no valid tokens remain after preprocessing, we assign a zero vector, yielding a similarity of 0 when compared to any phrase. To obtain a single embedding per phrase, we compute the average of all word embeddings within each phrase.

Given gold phrases G and predicted phrases P, we construct a similarity matrix S where S_{ij} is the cosine similarity between embeddings of gold phrase g_i and predicted phrase p_j:

$$S_{ij} = \frac{v(g_i) \cdot v(p_j)}{\|v(g_i)\| \, \|v(p_j)\|},$$

where -1 indicates maximally opposed meaning, 0 no semantic relationship (orthogonal vectors), and 1 perfect alignment.

As with string-level metrics, predicted and gold phrases may differ in number or order. We perform optimal phrase-to-phrase matching, assigning unmatched phrases a similarity of 0. We report two complementary perspectives analogous to the Levenshtein alignment:

- **Gold-based (recall-oriented):**

$$\text{Sim}_{G \to P} = \frac{1}{|G|} \sum_{g_i \in G} \max_{p_j \in P} S_{ij},$$

  measuring how well predicted phrases cover gold phrases.

- **Prediction-based (precision-oriented):**

$$\text{Sim}_{P \to G} = \frac{1}{|P|} \sum_{p_j \in P} \max_{g_i \in G} S_{ij},$$

  measuring how well predicted phrases correspond to gold references.

5. Results and Discussion

We evaluate four LLMs across four work-phrase extraction and normalisation tasks. Results are reported for both full original texts and filtered versions (with *Salutatio* and *Conclusio* removed). All results are reported as means ± standard deviation.

5.1. Overall Performance and Output Statistics

Identifying phrases describing work in historical petitions remains a significant challenge. As shown in Table 5, a primary driver of low precision is the tendency of LLMs to over-produce content. While the gold standard averages 4.81 phrases per document, models, particularly Llama 8B, consistently over-generate candidates (13.65 phrases). This high output volume directly explains why recall consistently exceeds precision: by over-generating candidates, models increase the likelihood of overlapping with gold spans at a high cost to precision. Furthermore, the extracted historical phrases are consistently shorter than those in the gold standard. This suggests that models sometimes fail to capture the full extent of annotated work-related spans, or that responses are being prematurely truncated during JSON generation. However, this length discrepancy does not extend to the normalisation process, since the predicted normalised phrases remain close to the gold standard in length.

To provide a reference for the embedding-based metrics, we establish a domain-internal baseline by calculating the mean similarity between randomly paired gold and predicted phrases across the test set. This provides a baseline score of 0.72 for historical phrases (used in Tasks 1, 3, and 4) and 0.51 for normalized phrases (used in Tasks 2, 3, and 4). The higher historical baseline reflects the formulaic nature and lexical density of 18th-century petitionary language, where work-related expressions frequently share domain-specific semantics. Consequently, embedding scores in the following sec-

Source	Mode	Phrases/Doc	Tokens/Phrase
Gold Hist.	-	4.81 ± 4.50	6.32 ± 4.62
Gold Norm.	-	4.81 ± 4.50	2.80 ± 0.87
Task 1: Phrase Extraction			
Llama 70B	Full	7.85 ± 5.88	3.97 ± 1.87
	Filt.	7.81 ± 5.90	4.08 ± 1.96
Llama 8B	Full	13.65 ± 9.58	4.98 ± 2.56
	Filt.	13.64 ± 9.59	4.99 ± 2.65
Mixtral	Full	5.85 ± 4.10	4.42 ± 2.16
	Filt.	6.44 ± 5.83	4.81 ± 3.12
GPT-4o	Full	6.85 ± 5.06	5.66 ± 2.20
	Filt.	6.72 ± 5.54	5.58 ± 2.17
Task 2: Isolated Normalisation (Gold Input)			
Llama 70B	Full	13.35 ± 10.31	2.69 ± 0.60
	Filt.	11.10 ± 5.87	2.71 ± 0.67
Llama 8B	Full	11.64 ± 8.55	2.95 ± 1.28
	Filt.	13.23 ± 14.20	3.21 ± 2.01
Mixtral	Full	8.26 ± 5.17	2.27 ± 0.59
	Filt.	8.76 ± 5.69	2.24 ± 0.50
GPT-4o	Full	7.04 ± 5.54	2.63 ± 0.81
	Filt.	6.35 ± 5.12	2.53 ± 0.61

Table 5: Descriptive Statistics (Mean ± SD) for Tasks 1 and 2. Predicted outputs are shown for both original full text and filtered text modes.

Metric	Mode	L70B	L8B	Mixtral	GPT
Phrase P	Full	.069±.15	.041±.13	**.099±.20**	.077±.17
	Filt.	**.098±.19**	.042±.13	.081±.17	.077±.17
Phrase R	Full	.097±.21	.089±.20	.102±.20	**.147±.30**
	Filt.	.116±.22	.090±.20	.086±.18	**.147±.30**
Phrase F1	Full	.071±.15	.049±.13	**.087±.16**	.084±.16
	Filt.	**.098±.18**	.049±.13	.073±.14	.084±.16
Token P	Full	.272±.24	.196±.19	.298±.23	**.301±.21**
	Filt.	.303±.24	.197±.19	.270±.21	**.310±.20**
Token R	Full	.364±.28	.471±.27	.351±.28	**.513±.30**
	Filt.	.393±.27	.473±.27	.349±.28	**.495±.27**
Token F1	Full	.255±.18	.224±.16	.279±.20	**.319±.18**
	Filt.	.294±.20	.227±.16	.253±.17	**.329±.16**
Lev (G)	Full	.378±.20	.398±.19	.354±.23	**.427±.26**
	Filt.	.397±.20	.397±.19	.357±.22	**.427±.26**
Lev (P)	Full	.251±.19	.168±.16	**.293±.22**	.280±.18
	Filt.	.278±.21	.168±.16	.277±.19	**.280±.18**
Emb (G)	Full	.784±.16	**.836±.13**	.707±.23	.781±.21
	Filt.	.796±.15	**.833±.13**	.722±.23	.781±.21
Emb (P)	Full	.518±.28	.354±.26	.567±.30	**.568±.28**
	Filt.	.528±.28	.354±.26	.561±.28	**.568±.28**

Table 6: Task 1 Extraction Performance: Mean ± SD for Precision (P), Recall (R), and F1-score (F1). (G) denotes Gold-oriented (recall-like) and (P) denotes Prediction-oriented (precision-like) similarities. Comparison of Full original and Filtered (Filt.) text modes across all models.

Metric	Mode	L70B	L8B	Mixtral	GPT
Task 2: Isolated Normalisation					
Phrase P	Full	.023±.06	.011±.04	.037±.12	**.077±.19**
	Filt.	.032±.09	.009±.04	.059±.18	**.085±.21**
Phrase R	Full	.050±.14	.023±.11	.039±.10	**.088±.19**
	Filt.	.047±.11	.012±.05	.066±.17	**.089±.19**
Phrase F1	Full	.027±.06	.013±.04	.033±.09	**.073±.16**
	Filt.	.035±.08	.010±.04	.049±.13	**.077±.16**
Token F1	Full	.124±.10	.096±.08	.121±.13	**.196±.16**
	Filt.	.139±.12	.091±.08	.142±.16	**.207±.16**
Lev (G)	Full	**.446±.14**	.374±.14	.342±.20	.414±.18
	Filt.	**.437±.14**	.375±.13	.364±.22	.414±.20
Emb (G)	Full	**.716±.11**	.649±.15	.559±.27	.650±.19
	Filt.	**.704±.11**	.651±.14	.575±.26	.628±.21
Task 3: Pipeline Normalisation					
Phrase P	Full	.049±.10	.026±.11	.060±.15	**.089±.19**
	Filt.	.055±.13	.025±.11	.058±.14	**.089±.18**
Phrase R	Full	.069±.15	.033±.12	.066±.15	**.082±.16**
	Filt.	.081±.18	.033±.12	.063±.13	**.086±.19**
Phrase F1	Full	.053±.10	.024±.08	.055±.12	**.080±.16**
	Filt.	.059±.12	.024±.08	.055±.11	**.081±.17**
Token F1	Full	.157±.13	.095±.10	.166±.15	**.217±.17**
	Filt.	.164±.13	.093±.10	.176±.15	**.217±.16**
Lev (G)	Full	**.447±.14**	.376±.15	.373±.19	.413±.18
	Filt.	**.447±.15**	.371±.16	.387±.18	.408±.18
Emb (G)	Full	**.705±.12**	.640±.15	.599±.23	**.631±.20**
	Filt.	**.703±.12**	.631±.16	.618±.20	.626±.20

Table 7: Task 2 and Task 3 Normalisation Performance: Mean ± SD for Precision (P), Recall (R), and F1-score (F1). (G) denotes Gold-oriented (recall-like) and (P) denotes Prediction-oriented (precision-like) similarities. Comparison of Full original and Filtered (Filt.) text modes across all models.

Metric	Mode	L70B	L8B	Mixtral	GPT
Combined Task: Extraction Component					
Phrase P	Full	.083±.17	.033±.08	.081±.19	**.086±.16**
	Filt.	.081±.17	.033±.08	.067±.13	**.088±.14**
Phrase R	Full	.100±.22	.062±.15	.090±.19	**.141±.28**
	Filt.	.103±.22	.062±.15	.081±.17	**.149±.28**
Phrase F1	Full	.083±.17	.038±.09	.078±.16	**.094±.17**
	Filt.	.081±.17	.038±.09	.067±.12	**.098±.16**
Token F1	Full	.278±.21	.216±.15	.273±.18	**.316±.19**
	Filt.	.294±.20	.216±.15	.253±.17	**.311±.19**
Lev (G)	Full	.377±.21	.366±.17	.350±.21	**.411±.24**
	Filt.	.381±.21	.366±.17	.345±.21	**.398±.25**
Emb (G)	Full	.765±.18	**.813±.16**	.734±.21	.776±.20
	Filt.	.768±.18	**.813±.16**	.716±.24	.755±.23
Combined Task: Normalisation Component					
Phrase F1	Full	.014±.05	.010±.04	.026±.07	**.027±.09**
	Filt.	.020±.07	.010±.04	.021±.06	**.032±.09**
Token F1	Full	.141±.12	.087±.09	.131±.12	**.155±.14**
	Filt.	.146±.12	.087±.09	.129±.12	**.168±.14**
Lev (G)	Full	.338±.14	.332±.15	.325±.15	**.340±.15**
	Filt.	.346±.14	.332±.15	.323±.15	**.349±.17**
Emb (G)	Full	.623±.17	.630±.18	.571±.20	**.631±.18**
	Filt.	**.629±.16**	**.629±.18**	.570±.20	.624±.20

Table 8: Task 4 Combined Extraction and Normalisation Performance: Mean ± SD for Precision (P), Recall (R), and F1-score (F1). (G) denotes Gold-oriented (recall-like) and (P) denotes Prediction-oriented (precision-like) similarities. Comparison of Full original and Filtered (Filt.) text modes across all models.

tions must be interpreted relative to these thresholds.

5.2. Task 1: Extraction of Phrases Describing Work

Task 1 measures the models' ability to identify historical work-phrase spans. As shown in Table 6, models generally struggle with exact span reproduction. However, token-level F1 consistently exceeds phrase-level F1 by a factor of three, and Embedding Similarity scores average ≈ 0.80. While these scores indicate that models consistently locate relevant topical regions, they are only slightly above the 0.72 random baseline, suggesting that while the general semantic content is captured, the models struggle to isolate the specific nuances or precise boundaries defined by human annotators.

5.3. Normalisation: In Isolation vs. Pipeline

Here, we evaluate the models' ability to normalise historical phrases describing work into modern, standardised Swedish. Tasks 2 and 3 compare the linguistic capability of normalisation in isolation (based on manually extracted gold phrases) against a more demanding pipeline setting (based on phrases extracted by the system). Descriptive statistics for these tasks are summarised in Table 7.

Interestingly, Task 2 (Isolated) proved more difficult than Task 3 (Pipeline) for all models. Exact phrase matching remained low ($F1 < .09$), but performance improved when moving from isolated phrases to the pipeline. This suggests a benefit from surrounding context cues: performing extraction and normalisation in sequence allows models to leverage broader semantic cues to resolve historical ambiguities that are absent when phrases are processed in isolation.

5.4. Task 4: Combined Extraction and Normalisation

Task 4 tests whether a joint approach functions as an implicit chain-of-thought. As shown in Table 8, for GPT, this combined task acted as a helpful guide; its extraction score improved from .084 (Task 1) to .098 (Task 4). This suggests that, for very strong models, the extra effort of normalising a phrase may help them "sharpen" their focus and identify the correct historical span.

In contrast, other models seem overwhelmed by the dual-task requirement. For Llama 70B, doing both tasks at once appears to be a distraction, causing its extraction score to drop from .098 to .081. Most importantly, the quality of the normalisation collapse for all models in this combined setup

compared to the staged pipeline. For instance, GPT's normalisation score falls from .081 in Task 3 to only .032 in Task 4. These results indicate that while combined processing may sharpen initial span detection, it compromises the lexical precision required for accurate normalisation.

5.5. Cross-Task Comparisons and Analysis of Text Filtering

A comparison of all four tasks reveals a distinct difficulty hierarchy: Task 4 (Normalisation Component) is the most difficult, followed by Task 2 (Isolated Normalisation). Paradoxically, models perform best on the Extraction Component of Task 4. This suggests that while multitasking helps models locate relevant text, the complexity of producing a dual-language output somewhat degrades their ability to perform the linguistically demanding task of normalisation.

The impact of text filtering (removing *Salutatio* and *Conclusio*) is generally marginal and inconsistent. Across most tasks, the difference between Full and Filtered modes is minor. Furthermore, the effect is not uniform across models; while filtering provides a small boost to normalisation in the stronger models (GPT and L70B) by reducing noise, it occasionally hinders extraction for others. For example, Mixtral's extraction performance actually drops in Filtered mode in both Task 1 and Task 4. This suggests that while noise reduction can slightly help with the linguistic focus required for normalisation, the models are largely robust to the structural "noise" of historical petitions, making filtering a secondary factor in overall performance.

5.6. Qualitative Error Analysis

To complement the quantitative evaluation, we conduct a targeted qualitative error analysis. We manually examine cases in which the models perform poorly, as well as targeting instances where surface-level and semantic-level evaluation metrics diverge.

A recurrent issue is the failure to produce parseable output, with Llama 8B and Mixtral exhibiting the highest failure rates (see more in Section 4.3). Llama 8B most often produces malformed outputs, likely reflecting the limitations of a smaller model handling complex formatting instructions. By contrast, Mixtral frequently returns empty outputs or echoes the input, suggesting that technical factors like API instability or system load contribute to these failures. Notably, occasional parse failures occur for all models without an apparent structural cause, illustrating the unpredictability of LLM behavior.

Another prominent error pattern in low-performing cases is systematic over-prediction. Models frequently identify more phrases than present in the gold data, consistent with higher recall than precision. In extreme cases, several models tag a wide range of isolated verbs as work-related, such as "emottaga" (*receive*) and "frågat" (*asked*), even when the surrounding context does not support such interpretations. These errors are not limited to verbs, but also instances of nouns. Longer input texts appear to amplify this tendency toward overgeneralisation. We further observe sporadic hallucinations, i.e., predictions that cannot be traced to any span in the input text. However, a larger-scale analysis would be needed to determine their frequency.

At the same time, we identify cases in which models predict phrases that are not present in the gold annotations but can plausibly describe work-related activities when considered in context. Examples include verb phrases such as "brukat Qwacksalverÿ och ögonsalvor" (*practised quackery and eye-ointments*), which are normalised to "bruka kvacksalveri" (*practice quackery*) and "tillverka/sälja/bruka ögonsalvor" (*make/sell/use eye salves*), as well as the normalised phrase "tjäna som fältskärs gesäll" (*serve as an apprentice/journeyman barber-surgeon*). There are also several cases of occupational titles being extracted, including "Rådman" (*councilman*) and "Inspector" (*inspector*), which are normalised to "arbeta/tjäna som rådman" (*work as a councilman*), "arbeta/tjäna som inspektor" (*work as an inspector*), or "utföra inspektion" (*perform inspection*). Consultation with historians suggests that, although models often overgeneralise, some outputs seem to highlight valid work activities that could enrich manual curation.

Comparisons between surface-level (Levenshtein) and semantic (embedding-based) similarity reveal frequent divergences. Low string similarity often arises when predicted spans overlap with gold spans but include additional material, or when character overlap is incidental. Because each metric aligns predictions independently, direct one-to-one interpretation is somewhat limited. Moreover, the metrics capture fundamentally different aspects of similarity, which further limits direct one-to-one interpretation. Nevertheless, some consistent patterns emerge. When a prediction preserves most of the intended meaning but diverges in surface form, embedding similarity tends to align more closely with human judgement. For example:

Gold: "stå uthe på fiskaretorget"
Prediction: "stå uthe på fiskaretorget att förtiäna något till uppehälle"
Gold translated: *stand outside in the fishermen's square*
Prediction translated: *stand outside in the fishermen's square to earn something for a living*
Levenshtein score: 0.43
Embedding score: 0.93

However, embedding similarity also exhibits behaviour that appears counterintuitive from a hu-

man perspective. High scores do not always correspond to strong semantic equivalence, and conversely, low scores can occur even when the core work-related activity is preserved. This variability indicates that embedding-based similarity can diverge from human judgement in both directions, making it difficult to interpret in isolation. As established in Section 4.1, embedding scores must be interpreted in light of the high domain-internal baselines (0.72 for historical text and 0.51 for normalised text). A score slightly above these thresholds risks overstating semantic alignment, while scores below the baseline may under-represent semantic proximity between near-synonyms. Examples from both Task 1 (the extraction task, with historical original phrases) and Task 3 (pipeline normalisation, with normalised modern phrases) include:

Gold: "sälja varor"
Prediction: "bedriva handel"
Gold translated: *sell goods*
Prediction translated: *conduct trade*
Embedding score: 0.66

Gold: "bruka stenbod"
Prediction: "bruka lokal"
Gold translated: *use a stone shop*
Prediction translated: *use premises*
Embedding score: 0.78

Gold: "taget penningar till låns"
Prediction: ""smida för sin räkning"'
Gold translated: *borrowed money*
Prediction translated: *forge for one's own account*
Embedding score: 0.80

Taken together, these observations suggest that embedding-based similarity provides a valuable signal for identifying partially correct extractions, but also carries a risk of overstating semantic alignment between gold annotations and model outputs. A combined evaluation strategy that integrates both surface-level and semantic metrics therefore appears necessary for a more comprehensive and reliable assessment of model behaviour than relying on either metric alone.

6. Conclusions

This study explores the potential of four large language models, GPT-4o, Llama-3 (70B and 8B), and Mixtral-8x7B, for extracting and normalising phrases describing work from historical Swedish petitions. Our experiments demonstrate that while LLMs struggle to reproduce human-annotated spans exactly, yielding low phrase-level precision and F1 scores, token-level and embedding-based metrics offer complementary perspectives. These results indicate that models consistently locate relevant topical regions, although the high domain-specific baselines suggest that embedding scores can reflect general semantic intent rather than precise equivalence.

The choice of task architecture emerges as a factor in performance. We find that normalisation is notably more effective when performed within a contextual pipeline (Task 3) than in isolation (Task 2), suggesting that historical normalisation relies on integrative semantic cues. Furthermore, our results reveal a "multitasking paradox" in combined setups: while simultaneous extraction and normalisation can sharpen phrase identification for high-parameter models, the increased complexity significantly degrades the lexical precision of the normalised output. Although rhetorical filtering may prove valuable for other downstream applications, where removing formulaic sections could more effectively improve signal quality, the removal of *Salutatio* and *Conclusio* sections show only marginal effects in the present tasks.

Manual error analysis highlights a systematic tendency toward over-prediction and occasional hallucinations. However, the models also identify several plausible work-related expressions absent from the human annotations, pointing to the potential value of LLMs for discovery-driven research. A more extensive manual evaluation of these candidates would be valuable for understanding how much genuinely new information can be recovered versus the amount of noise added through overgeneralisation.

Overall, our results indicate that LLMs show promise for facilitating the identification of work-related activities in historical texts, provided their outputs are subject to expert evaluation and interpreted through metrics that account for high background similarity. Future work may explore more robust prompting strategies, domain-adapted fine-tuning, and hybrid human–machine workflows that combine automated semantic detection with expert validation. While a staged pipeline currently offers the most reliable balance of accuracy, the models' ability to detect valid expressions missed by humans suggests that such workflows could significantly enhance the coverage and interpretability of cultural heritage datasets.

7. Limitations

While this study provides insights into the capabilities of LLMs for processing historical Swedish, several limitations should be noted.

First, our evaluation is limited to a specific genre of historical text, 18th-century petitions, and a single language. The linguistic challenges found here, such as non-standardised orthography and formulaic rhetorical structures, may behave differently in other historical contexts, time periods, or languages.

Second, the gold standard used for evaluation, while expert-annotated, represents only one possible interpretation of phrases describing work. As noted in our qualitative analysis, the models sometimes identify potential work activities that are not present in the gold data, suggesting that exact-match metrics like F1-score may provide an overly conservative estimate of model utility in discovery-driven research.

Third, while we made use of twelve purposefully selected, diverse examples to ground the models across a range of linguistic challenges, performance in LLMs can be highly sensitive to prompt wording and the specific selection and ordering of examples. Although our prompt template was iteratively refined, we did not perform an thorough search of the prompt space; thus, our instructions may represent a suboptimal configuration for certain model architectures. Furthermore, the instructional complexity and total prompt length required to convey these nuanced historical tasks likely strained the context-handling capabilities of smaller models, contributing to the observed output instability and malformed JSON responses.

Finally, while few-shot prompting provides a strong baseline, this study did not explore domain-specific fine-tuning or more advanced reasoning architectures, which could improve the balance between extraction and normalisation accuracy in the combined task setup.

8. Ethical Considerations

The use of large language models in historical research raises several ethical considerations. Although large models entail energy costs, our few-shot prompting approach avoids resource-intensive fine-tuning and thus has a comparatively lower environmental impact. The historical petitions used in this study are in the public domain and contain no modern personal data, and their use follows established ethical guidelines for historical sources. Finally, our approach is designed to support, not replace, human expertise. We adopt a human-in-the-loop perspective, where automated extraction serves to assist and enrich manual curation rather than provide definitive interpretations, helping to guard against the uncritical acceptance of model-generated errors as historical fact.

9. Acknowledgments

We are grateful to our colleagues in the Petitions and GaW projects: Jonas Lindström, Örjan Kardell, Jezzica Israelsson, Maria Ågren, Sofia Ling, Linda Oja, and Fredrik Wahlberg, for their insightful discussions and collaborative support. Finally, we appreciate the constructive feedback from the anonymous reviewers. This research was funded by the Swedish Research Council (grant number 2018-06159).

10. Bibliographical References

Maria Ågren, Jonas Lindström, and Sofia Ling. 2023. The principles of entering data into the Gender and Work database. Technical report, Uppsala University and Umeå University. Confirmed by the management of the GaW database 2023-12-11.

Marcel Bollmann. 2019. A large-scale comparison of historical text normalization systems. *arXiv preprint arXiv:1904.02036*.

Toby Burrows. 2023. Reproducibility, verifiability, and computational historical research. *Journal on Computing and Cultural Heritage*, 16(3):1–12.

Kevin Cohen, Laura Manrique-Gómez, and Rubén Manrique. 2025. Historical ink: Exploring large language models for irony detection in 19th-century Spanish. *arXiv preprint arXiv:2503.22585*.

Rosemarie Fiebranz, Erik Lindberg, Jonas Lindström, and Maria Ågren. 2011. Making verbs count: the research project 'gender and work' and its methodology. *Scandinavian Economic History Review*, 59(3):273–293.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Stina Hansson. 1988. *Svensk brevskrivning: teori och tillämpning*, volume 18. Göteborgs universitet.

Simon Hengchen and Nina Tahmasebi. 2021. A collection of Swedish diachronic word embedding models trained on historical newspaper data. *Journal of Open Humanities Data*, 7(2):1–7.

Torsten Hiltmann, Martin Dröge, Nicole Dresselhaus, Till Grallert, Melanie Althage, Paul Bayer, Sophie Eckenstaler, Koray Mendi, Jascha Marijn Schmitz, Philipp Schneider, et al. 2025. Ner4all or context is all you need: Using llms for low-effort, high-performance ner on historical texts. a humanities informed approach. *arXiv preprint arXiv:2502.04351*.

Mark Humphries, Lianne C Leddy, Quinn Downton, Meredith Legace, John McConnell, Isabella

Murray, and Elizabeth Spence. 2025. Unlocking the archives: Using large language models to transcribe handwritten historical documents. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, pages 1–19.

Jezzica Israelsson. 2016. In consideration of my meagre circumstances: The language of poverty as a tool for ordinary people in early modern Sweden. Master's thesis, Uppsala Universitet.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Lucija Krusic. 2024. Constructing a sentiment-annotated corpus of austrian historical newspapers: Challenges, tools, and annotator experience. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 51–62.

Ellinor Lindqvist, Eva Pettersson, and Joakim Nivre. 2025. Finding the plea: Evaluating the ability of llms to identify rhetorical structure in swedish and english historical petitions. In *Proceedings of the First on Natural Language Processing and Language Models for Digital Humanities*, pages 86–101, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

OpenAI. 2026. GPT-4o. API model accessed January 2026.

Michael Piotrowski. 2012. *Natural language processing for historical texts*, 1 edition. Synthesis lectures on human language technologies. Morgan Claypool Publishers, Cham.

Thomas Sokoll. 2006. Writing for relief: Rhetoric in English pauper letters, 1800–1834. In Andreas Gestrich, Steven King, and Lutz Raphael, editors, *Being Poor in Modern Europe: Historical Perspectives 1800–1940*, pages 91–112. Peter Lang, Bern.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Crina Tudor, Beáta Megyesi, and Robert Östling. 2025. Prompting the past: Exploring zero-shot learning for named entity recognition in historical texts using prompt-answering llms. In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, pages 216–226.

Huub Wijfjes. 2017. Digital humanities and media history: A challenge for historical newspaper research. *Digital Journalism*, 5(1):29–43.

David C. S. Wilson. 2022. Working at scale: What do computational methods mean for research using collections, models and big data? *Science Museum Group Journal*, (18).

Appendix: Prompts Used in Experiments

Prompt Task 1: Extraction

Swedish Original

```
Du är expert på lingvistisk analys av svenska historiska texter. Din uppgift är att
    analysera svenska suppliker från 1700-talet och identifiera alla fraser som
    beskriver arbete, arbetsutförande eller arbetsrelaterade handlingar.

# Huvudinstruktioner
1. **Läs texten noggrant.** Texten kan innehålla gammaldags stavning, fri ordföljd
    och sammanskrivningar.
2. **Extrahera endast fraser som beskriver arbete, arbetsutförande eller
    arbetsrelaterade handlingar.** Detta inkluderar att utföra, bidra till eller
    delta i arbete, syssla, tjänst eller plikt för försörjning, oavsett om ersä
    ttning ges.
3. **Bevara originaltexten exakt.** Lägg INTE till, ta bort eller ändra ord,
    stavning eller skiljetecken.
4. **Returnera en lista över extraherade fraser** i den ordning de förekommer i
    texten.

# Outputformat
Returnera endast ett JSON-objekt med nyckeln **"extracted_phrases"** och en lista
    med fraser från texten. Returnera JSON exakt i detta format, utan extra text,
    kommentarer, sammanfattningar, extra nycklar eller kodblock-markeringar:

{
  "extracted_phrases": ["mala någon säd", "i tiänst warit"]
}

# Exempel
(Endast korta utdrag visas i exemplen. I uppgiften ska hela texten bearbetas.)

### Input 1:
"/.../ efter som thenne tiden warit så swårt efter humble? Swarades hon hade kiöpt
    humbla ifrån Tidön, hwartil hon giordt en liten tilökning af några Malörts
    Qwistar, /.../"

### Output 1:
{
  "extracted_phrases": ["kiöpt humbla"]
}

/.../

# Identifiera och extrahera de fraser som beskriver arbete i följande supplik:
```

English Translation

```
You are an expert in the linguistic analysis of Swedish historical texts. Your task
    is to analyse Swedish petitions from the 18th century and identify all phrases
    describing work, the performance of work, or work-related actions.

# Main Instructions
1. **Read the text carefully.** The text may contain archaic spelling, free word
    order, and compound words.
2. **Extract only phrases describing work, the performance of work, or work-related
    actions.** This includes performing, contributing to, or participating in work,
    occupation, service, or duty for a livelihood, regardless of whether
    compensation is provided.
3. **Preserve the original text exactly** Do NOT add, remove, or change words,
    spellings, or punctuation marks.
4. **Return a list of extracted phrases** in the order they appear in the text.

# Output Format:
Return only a JSON object with the key **"extracted_phrases"** and a list of phrases
     from the text. Return the JSON exactly in this format, without any additional
    text, comments, summaries, extra keys, or code block markers:

{
  "extracted_phrases": ["grind some grain", "been in service"]
}

# Examples
(Only short excerpts are shown in the examples. In the actual task, the entire text
    should be processed.)

### Input 1:
"/.../ seeing as hops have been so scarce of late? She answered that she had
    purchased hops from Tidö, to which she had made a small addition of some
    Wormwood Twigs, /.../"

### Output 1:
{
  "extracted_phrases": ["purchased hops"]
}

/.../

# Identify and extract the phrases describing work in the following petition:
```

Prompt Task 2: Normalisation in Isolation

Swedish Original

```
Du är expert på lingvistisk analys av svenska historiska texter. Din uppgift är att
    normalisera fraser som beskriver arbete, arbetsutförande eller arbetsrelaterade
    handlingar i svenska suppliker från 1700-talet. Du kommer att få en lista med
    extraherade fraser som har identifierats som beskrivningar av arbete. Du ska
    endast normalisera dessa fraser.

# Huvudinstruktioner
1. **Läs varje fras noggrant.** Fraserna kan innehålla gammaldags stavning,
    grammatik och ordföljd.
2. **Normalisera varje fras till modern svenska i grundform (infinitiv).**
   - Använd modern stavning och modern grammatik.
   - Formulera varje fras som en verbfras i infinitiv, även om originalfrasen inte
     är verbal.
3. **Ändra inte innehållet.**
   Lägg inte till, ta bort eller slå ihop fraser.
4. **Bevara ordningen.** Varje normaliserad fras ska motsvara frasen på samma
    position i indata.

# Outputformat
Returnera endast ett JSON-objekt med nyckeln **"normalised_phrases"** och en lista
    med fraser i normaliserad grundform. Returnera JSON exakt i detta format, utan
    extra text, kommentarer, sammanfattningar, extra nycklar eller kodblock-
    markeringar:

{
  "normalised_phrases": ["mala säd", "vara i tjänst"]
}

# Exempel

### Input 1:
{
  "extracted_phrases": ["kiöpt humbla"]
}

### Output 1:
{
  "normalised_phrases": ["köpa humle"]
}

# Normalisera följande fraser:
```

English Translation

```
You are an expert in the linguistic analysis of Swedish historical texts. Your task
    is to analyse Swedish petitions from the 18th century and identify all phrases
    describing work, the performance of work, or work-related actions.

You are an expert in the linguistic analysis of Swedish historical texts. Your task
    is to normalise phrases describing work, the performance of work, or work-
    related actions in Swedish petitions from the 18th century. You will be provided
     with a list of extracted phrases that have been identified as descriptions of
    work. You are to only normalise these phrases.

# Main Instructions
1. **Read each phrase carefully.** The phrases may contain archaic spelling, grammar
    , and word order.
2. **Normalise each phrase into modern Swedish in the base form (infinitive)**
   - Use modern spelling and modern grammar.
   - Formulate each phrase as an infinitive verb phrase, even if the original phrase
       is not verbal.
3. **Do not change the content.**
   Do not add, remove, or merge phrases.
4. **Preserve the order.** Each normalised phrase must correspond to the phrase at
    the same position in the input data.

# Output Format:
Return only a JSON object with the key **"normalised_phrases"** and a list of
    phrases in normalised base form. Return the JSON exactly in this format, without
     any additional text, comments, summaries, extra keys, or code block markers:

{
  "normalised_phrases": ["grind grain", "to be in service"]
}

# Examples

### Input 1:
{
  "extracted_phrases": ["purchased hops"]
}

### Output 1:
{
  "normalised_phrases": ["purchase hops"]
}

/.../

# Normalise the following phrases::
```

Prompt Task 3: Normalisation Pipeline

Swedish Original

```
Du är expert på lingvistisk analys av svenska historiska texter. Din uppgift är att
    analysera svenska suppliker från 1700-talet och identifiera alla fraser som
    beskriver arbete, arbetsutförande eller arbetsrelaterade handlingar, och återge
    dessa i normaliserad grundform (infinitiv).

# Huvudinstruktioner
1. **Läs texten noggrant.** Texten kan innehålla gammaldags stavning, fri ordföljd
    och sammanskrivningar.
2. **Identifiera endast fraser som beskriver arbete, arbetsutförande eller
    arbetsrelaterade handlingar.** Detta inkluderar att utföra, bidra till eller
    delta i arbete, syssla, tjänst eller plikt för försörjning, oavsett om ersä
    ttning ges.
3. **Normalisera varje identifierad fras till modern grundform.**
    Använd modern stavning och grammatik.
    Formulera varje fras som en verbfras i infinitiv, även om originalet inte är
       verbalt.
4. **Returnera en lista över normaliserade fraser** i den ordning de förekommer i
    texten.

# Outputformat
Returnera endast ett JSON-objekt med nyckeln **"normalised_phrases"** och en lista
    med fraser i normaliserad grundform. Returnera JSON exakt i detta format, utan
    extra text, kommentarer, sammanfattningar, extra nycklar eller kodblock-
    markeringar:

{
  "normalised_phrases": ["mala säd", "vara i tjänst"]
}

# Exempel
(Endast korta utdrag visas i exemplen. I uppgiften ska hela texten bearbetas.)

### Input 1:
"/.../ efter som thenne tiden warit så swårt efter humble? Swarades hon hade kiöpt
    humbla ifrån Tidön, hwartil hon giordt en liten tilökning af några Malörts
    Qwistar, /.../"

### Output 1:
{
  "normalised_phrases": ["köpa humle"]
}

# Identifiera och normalisera (i grundform) de fraser som beskriver arbete i
  följande supplik:
```

English Translation

```
You are an expert in the linguistic analysis of Swedish historical texts. Your task
    is to analyse Swedish petitions from the 18th century and identify all phrases
    describing work, the performance of work, or work-related actions, and provide
    these in normalised base form (infinitive).

# Main Instructions
1. **Read the text carefully.** The text may contain archaic spelling, free word
    order, and compound words.
2. **Identify only phrases describing work, the performance of work, or work-related
     actions.** This includes performing, contributing to, or participating in work,
     occupation, service, or duty for a livelihood, regardless of whether
    compensation is provided.
3. **Normalise each identified phrase into modern base form (infinitive)**
   - Use modern spelling and grammar.
   - Formulate each phrase as an infinitive verb phrase, even if the original phrase
       is not verbal..
4. **Return a list of normalised phrases** in the order they appear in the text.

# Output Format:
Return only a JSON object with the key **"normalised_phrases"** and a list of
    phrases in normalised base form. Return the JSON exactly in this format, without
     any additional text, comments, summaries, extra keys, or code block markers:

{
  "normalised_phrases": ["grind grain", "to be in service"]
}

# Examples
(Only short excerpts are shown in the examples. In the actual task, the entire text
    should be processed.)

### Input 1:
"/.../ seeing as hops have been so scarce of late? She answered that she had
    purchased hops from Tidö, to which she had made a small addition of some
    Wormwood Twigs, /.../"

### Output 1:
{
  "normalised_phrases": ["purchase hops"]
}

/.../

# Identify and normalise (in base form) the phrases describing work in the following
     petition:
```

Prompt Task 4: Combined Extraction and Normalisation

Swedish Original

```
Du är expert på lingvistisk analys av svenska historiska texter. Din uppgift är att
    analysera svenska suppliker från 1700-talet och identifiera alla fraser som
    beskriver arbete, arbetsutförande eller arbetsrelaterade handlingar. Först ska
    du extrahera dessa fraser exakt som de förekommer i texten, därefter ska du
    normalisera varje extraherad fras till modern grundform (infinitiv).

# Huvudinstruktioner
1. **Läs texten noggrant.** Texten kan innehålla gammaldags stavning, fri ordföljd
    och sammanskrivningar.
2. **Extrahera endast fraser som beskriver arbete, arbetsutförande eller
    arbetsrelaterade handlingar.** Detta inkluderar att utföra, bidra till eller
    delta i arbete, syssla, tjänst eller plikt för försörjning, oavsett om ersä
    ttning ges.
3. **Bevara originaltexten exakt vid extraktion.** Lägg INTE till, ta bort eller ä
    ndra ord, stavning eller skiljetecken i de extraherade fraserna.
4. **Normalisera varje extraherad fras till modern grundform.**
   - Använd modern stavning och grammatik.
   - Formulera varje fras som en verbfras i infinitiv, även om originalfrasen inte ä
       r verbal.
5. **Bevara ordningen.** Den normaliserade frasen ska motsvara den extraherade
    frasen på samma position i listan.

# Outputformat
Returnera endast ett JSON-objekt med nycklarna **"extracted_phrases"** (fraser exakt
     ur texten) och **"normalised_phrases"** (motsvarande fraser i normaliserad
    grundform). Listorna ska ha samma längd och ordning. Returnera JSON exakt i
    detta format, utan extra text, kommentarer, sammanfattningar, extra nycklar
    eller kodblock-markeringar:

{
  "extracted_phrases": ["mala någon säd", "i tiänst warit"],
  "normalised_phrases": ["mala säd", "vara i tjänst"]
}

# Exempel
(Endast korta utdrag visas i exemplen. I uppgiften ska hela texten bearbetas.)

### Input 1:
"/.../ efter som thenne tiden warit så swårt efter humble? Swarades hon hade kiöpt
    humbla ifrån Tidön, hwartil hon giordt en liten tilökning af några Malörts
    Qwistar, /.../"

### Output 1:
{
  "extracted_phrases": ["kiöpt humbla"],
  "normalised_phrases": ["köpa humle"]
}

# Identifiera, extrahera och normalisera de fraser som beskriver arbete i följande
    supplik:
```

English Translation

```
You are an expert in the linguistic analysis of Swedish historical texts. Your task
    is to analyse Swedish petitions from the 18th century and identify all phrases
    describing work, the performance of work, or work-related actions. First, you
    must extract these phrases exactly as they appear in the text; subsequently, you
     must normalise each extracted phrase into modern base form (infinitive).

# Main Instructions
1. **Read the text carefully.** The text may contain archaic spelling, free word
    order, and compound words.
2. **Extract only phrases describing work, the performance of work, or work-related
    actions.** This includes performing, contributing to, or participating in work,
    occupation, service, or duty for a livelihood, regardless of whether
    compensation is provided.
3. **Preserve the original text exactly during extraction.** Do NOT add, remove, or
    change words, spellings, or punctuation marks in the extracted phrases.
4. **Normalise each extracted phrase into modern base form**
   - Use modern spelling and grammar.
   - Formulate each phrase as an infinitive verb phrase, even if the original phrase
       is not verbal.
5. **Preserve the order.** The normalised phrase must correspond to the extracted
    phrase at the same position in the list.

# Output Format:
Return only a JSON object with the keys **"extracted_phrases"** (phrases exactly
    from the text) and **"normalised_phrases"** (corresponding phrases in normalised
     base form) and a list of phrases in normalised base form. The lists must have
    the same length and order. Return the JSON exactly in this format, without any
    additional text, comments, summaries, extra keys, or code block markers:

{
   "extracted_phrases": ["grind some grain", "in service have been"],
   "normalised_phrases": ["grind grain", "to be in service"]
}

# Examples
(Only short excerpts are shown in the examples. In the actual task, the entire text
    should be processed.)

### Input 1:
"/.../ seeing as hops have been so scarce of late? She answered that she had
    purchased hops from Tidö, to which she had made a small addition of some
    Wormwood Twigs, /.../"

### Output 1:
{
  "extracted_phrases": ["purchased hops"],
  "normalised_phrases": ["purchase hops"]
}

/.../

# Identify, extract and normalise the phrases describing work in the following
    petition:
```

Extracting Volcanological Knowledge from Historical Texts: A Language-Technology Pipeline for Diachronic Geovisualization

Costanza Marini[*], **Gianluca Casagrande**[†], **Alessio Palmero Aprosio**[†], **Claudia Principe**[‡]

[*] University of Pavia, costanza.marini@unipv.it

[†]University of Trento, gianluca.casagrande@studenti.unitn.it, a.palmeroaprosio@unitn.it

[‡]Institute of Geosciences and Earth Resources - CNR, c.principe@igg.cnr.it

Abstract

This paper presents the first results of the CorVo project, a transdisciplinary project combining volcanology and computational linguistics to extract and structure volcanological knowledge from historical documents concerning Mount Vesuvius. We introduce the CorVo corpus, a multilingual diachronic corpus of 180 digitized texts (16th–20th centuries), selected to represent the main eruptive scenarios of the volcano. The digitization workflow integrates image pre-processing, OCR, and LLM-based post-correction to address challenges posed by degraded pages, historical typefaces, and orthographic variation. A domain-aware information extraction pipeline was developed to identify both standard toponyms and fine-grained spatial entities, which are typically overlooked by traditional NER systems. Extracted entities undergo human-in-the-loop validation and georeferencing through a dedicated annotation interface supporting multiple spatial geometries. The resulting dataset enables temporally normalized diachronic geovisualization of the textual-spatial footprint of Vesuvian eruptions across centuries.

Keywords: Historical Corpora, OCR, LLM, Named Entity Recognition, Diachronic Geovisualization

1. Introduction

Due to its long history of civilisation and high density of population, Italy is one of the few countries in the world, next to Japan (Kudo and Hoshizumi, 2006) and Indonesia (Voight et al., 2000), to possess an extensive historical record of the eruptive activity of its volcanoes (Giannini and Paolillo, 2018) – a record extensive enough to be exploited for the automatic extraction of useful information for volcanic eruption forecasting, risk assessment, and resilience planning.

Up until now, the use of historical sources by the geological community has been scarce due to the limited availability of digitized documents, the small number of research teams with both historiographic and volcanological expertise, and the time-consuming nature of manual data extraction from written sources (Principe, 1990).

Moreover, data extraction based solely on extensive reading of document collections cannot be considered fully objective or complete. It reflects the reader's bias, and goals and cannot be replicated without repeating the entire reading process (Guidoboni and Ebel, 2009). Today, the application of automated text analysis techniques to written sources can provide volcanologists with access to organised information on the historical activity of volcanoes in a time-efficient, replicable, and reliable manner, while volcanology can provide computational linguistics with a new testing ground for language technologies.

In this paper, we present the first findings of the CorVo project, which relies on a synergy of these two disciplines and makes three main contributions.

- First, it introduces CorVo corpus, a multilingual diachronic corpus of 180 digitized historical documents (16th–20th centuries) concerning the eruptive activity of Mount Vesuvius, designed to support both volcanological research and computational linguistic analysis.

- Second, it presents a domain-aware NLP pipeline that integrates OCR preprocessing, LLM-based post-correction, and structured geographic entity extraction tailored to fine-grained volcanological categories that are typically overlooked by standard NER systems.

- Third, it describes a human-in-the-loop geospatial annotation framework and a temporally normalized diachronic geovisualization system that enables systematic exploration of the textual-spatial footprint of Vesuvian eruptions across centuries.

Together, these contributions demonstrate how language technologies can operationalize historical sources for scientific risk assessment while advancing NLP research in low-resource, domain-specific contexts.

Proceedings of LT4HALA 2026 @ LREC 2026, pages 38–48.
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

2. The Project

The CorVo project[1] is a transdisciplinary project bringing together volcanology and computational linguistics to support research on volcanic activity and its cultural impact, as well as to improve volcanic risk forecasting, impact assessment, and resilience planning. The project relies on a digitised corpus of historical documents (the CorVo corpus) and state-of-the-art NLP tools to facilitate its exploration and analysis. The documents contained in the CorVo corpus describe the past activity of one of Italy's most high-risk volcanoes: the Vesuvius. By querying the corpus, Civil Protection units, volcanologists, and other stakeholders will be able to quickly obtain important information concerning past eruptions, their social impact and institutional reactions to the related events. With data on eruption precursors, the actual sequence of eruptive phenomena, deposit distribution, and damages at hand, they are going to be able to update the current eruptive scenarios of the volcano. This information is particularly valuable because it cannot be derived from field data alone and, if properly integrated, may contribute to saving lives.

At present, the CorVo corpus interface enables exploration of 180 of the project's already digitised documents, the spatial data they contain, and their curated metadata.

The CorVo corpus, its metadata and the source code interfaces described in this paper will be released open source upon acceptance. See Section 8 for further details.

3. Related Work

In NLP research, volcanology can be currently considered a low-resource domain. The only Italian volcanology-focused corpus we found in the literature contains 2,839 daily and weekly Stromboli monitoring bulletins (2015-2021) and was compiled by Berardi et al. (2022) to extract numerical parameters for time-series reconstruction. More geology-focused corpora have been built in other languages, mostly for NLP tasks such as Named Entity Recognition (NER), Information Extraction (IE) and Information Retrieval (IR). For instance, for Chinese, Fan et al. (2020) created a geological-hazard corpus of 4,548 sentences annotated for NER and knowledge-graph construction, while Chen et al. (2022) released a fine-grained geological NER corpus annotated with 21 entity labels (e.g., sedimentary rock). Ma et al. (2023) developed an ontology-based BERT model for IE on geological hazard reports, while Qiu et al. (2020) worked on spatiotemporal IE from unstructured geoscience reports. For Portuguese, Lima de Oliveira et al. (2021) created REGIS (Retrieval Evaluation for Geoscientific Information Systems) a test collection containing over 20,000 scientific documents, 34 query topics and the corresponding relevance judgments for geological information retrieval evaluation, an important field often forgotten. On the other hand, Nunes et al. (2024) introduced GeoCorpus3, a geological NER resource containing 30 fine-grained entity classes including rock types (e.g., magmatic, metamorphic and sedimentary), as well as stratigraphical elements (e.g., sedimentary basin and geotectonic unit). For English, Lawley et al. (2022) have retrained language models on geoscientific documents and obtained better performing domain-specific geoscience language models, while Padarian and Fuentes (2019) introduced the GeoVec corpus, a corpus of 280,764 English geoscience articles and Wikipedia pages for word embeddings. Together, these studies define the current landscape of volcanology-specific and broader geological textual corpora for NLP.

As one can see, NLP research has not been focusing on extracting volcanological knowledge and reconstructing past eruptions from historical sources. On the other hand, this is exactly what several Italian volcanologists and geologists have been doing since the 1980s while compiling catalogues of geological disasters to extract this information from, as in the case of both the catalogue of Italian earthquakes by Boschi et al. (1995) and the BIBV database of Italian active volcanoes (Giannini and Paolillo, 2018), albeit without the help of computational methodologies. This being said, due to the enormous amount of time required to manually collect and analyse historical documents, the extraction of volcanological information from the available sources has always been restricted to the most documented eruptions and oriented towards specific phenomena, such as an eruption precursor (Bertagnini et al., 2006) or a given territory occupied by a lava flow (Branca et al., 2013).

Thanks to an extensive document collection, which could benefit from decades of expertise in the manual consultation of the relevant sources, and the help of LLMs, the CorVo project overcomes these limitations.

4. The Corpus

The CorVo corpus contains a balanced selection of historical documents, chronicles and monographs describing the three possible eruptive scenarios Vesuvius may experience in case of reactivation: (1) Plinian-type eruption, (2) fissural eruption, and (3) mixed (explosive-effusive) eruption.

[1] https://www.corvo-project.eu/

4.1. Document Selection

The document selection process was facilitated by the BIBV database (Giannini and Paolillo, 2018), which contains the bibliographical metadata (e.g., manuscript title, author, number of pages, exact location in the Italian National Libraries, etc.) of over 3,400 documents describing the historical eruptive activity of all active Italian volcanoes. Since the CorVo project is a case study restricted to Vesuvius, 180 Vesuvius-related documents – 176 of which from BIBV – were selected, traced back to a physical Library, scanned and transformed into machine-readable text format using OCR technology. The difficulties encountered at this stage will be dealt with in section 5.

In terms of additional criteria used for the selection of these 180 documents, we decided to favour descriptions of eruptive events as contemporary as possible to the eruption year, authored by eyewitnesses or educated individuals quoting eyewitnesses. The cultural backgrounds of said authors can vary greatly and range from well-known scholars of the time to administrators and clergymen. Furthermore, since not all authors have described the same phenomena from the same geographical point of view, the author's awareness of their geolocation and perspective was an important component in the selection process. In addition to documents and chronicles concerning the three main eruptions/scenarios of the Vesuvius, a small amount of treatises on the history of the volcano have been selected to obtain more detailed geographical information on the location.

This being said, the bulk of the CorVo corpus consists of documents that have already been individually used in previous volcanological studies to compare manually extracted historical data from written sources with volcanological data from fieldwork and deposits – see Sigurdsson et al. (1985) on the Vesuvian eruption of 79 AD and Cole and Scarpati (2010) on the eruption of 1944. We see the existence of these studies as an added value, since they will allow us to validate the findings emerging from our work with something we already know well.

However, unlike these works, the effort made in the CorVo project differs in both extent and methodology. For one, previous works usually exploited only a few historical sources at a time due to the limitations of manual consultation: an obstacle erased by the use of NLP tools. Moreover, most of these documents had not been digitised: one of our achievements. Another hurdle was identifying the right physical locations where the digitization could take place, minimising the number of Libraries and movements of the digitisation company, while acknowledging the restrictions of the equipment in terms of time, space and accessibility. Only printed (i.e., non-handwritten) texts were selected for the project to minimize the difficulties in producing OCR formats. Despite this, the quality of the documents varied greatly depending on their age, state of preservation and binding. Therefore, some compromises had to be made, and only texts from which sufficiently high-quality images could be obtained were selected.

4.2. Corpus languages

In terms of languages, the CorVo corpus is a multilingual diachronic corpus skewed towards Italian, as you can see from Table 1. According to Renzi and Salvi (2010), a distinction can be made between Old and Modern Italian around 1525, the year in which Bembo published his *Prose della volgar lingua*. Therefore, we can state that the vast majority of the documents of the CorVo corpus is in Modern Italian, which optimizes the trade-off between the amount of linguistic variation in our data and potential processing problems cause by such variation.

As for the other languages in the corpus, it is well-known that the activity of Italian volcanoes has been described over time by numerous observers of different native languages. Therefore, the CorVo project includes: eleven Latin texts published between the 16th and 17th century; two documents in 17th-century Spanish written by members of the Spanish bureaucratic-governmental apparatus during the period of Spanish domination of Southern Italy (17th-18th centuries); and finally six accounts in French, four in German, and five in English by various travelers on their Italian Grand Tour (18th-19th centuries).

To facilitate access to the corpus to Protection units, volcanologists and other stakeholders interested in obtaining information concerning past Vesuvian eruptions, the documents in foreign languages were translated using TranslateGemma (Finkelstein et al., 2026), with the exception of text in Latin, translated using ChatGPT 5.2.[2] Toponym extraction was performed on original-language texts; translations are provided only for accessibility (see Section 6).

5. Optical Character Recognition

The first phase of the digitization workflow consisted of Optical Character Recognition (OCR), aimed at converting scanned page images into machine-readable text. Given the heterogeneous nature of the document collection, the OCR process required careful configuration and post-processing

[2] https://openai.com/index/introducing-gpt-5-2/

	Century (publication year)					
Lang	16th	17th	18th	19th	20th	21st
it	-	26	61	38	26	1
en	-	-	4	1	-	-
de	-	-	-	4	-	-
fr	-	-	-	5	1	-
es	-	2	-	-	-	-
la	2	9	-	-	-	-

Table 1: Number of documents, collected by year of publication and language.

to address several structural and linguistic challenges.

5.1. Image Pre-processing

Prior to text recognition, the scanned images underwent a pre-processing stage to improve OCR accuracy. Historical documents often exhibit physical degradation such as ink stains, paper discoloration, bleed-through, and scanning artifacts. To mitigate these issues, image enhancement techniques were applied, including:

- Binarization and adaptive thresholding to increase contrast between text and background;
- Noise reduction filters to remove ink spots and speckle artifacts;
- Deskewing and geometric correction to address page misalignment;

The Python library OpenCV[3] has been used to perform the above-described pre-processing tasks.

5.2. Layout Analysis and Segmentation

A layout analysis phase was conducted to identify and segment distinct structural components of each page, including main text blocks, footnotes, marginal notes, tables, and other paratextual elements. Accurate segmentation was particularly important for:

- Separating footnotes from the main body text to preserve logical structure;
- Identifying tables and maintaining row–column relationships;
- Distinguishing text regions from non-textual elements such as decorative initials or graphical separators.

These steps have been mainly carried forward using Mistral OCR API[4] and were essential to reduce recognition errors introduced by degraded page conditions.

5.3. OCR issues

Although OCR is a well-established technology, the characteristics of the document collection significantly affected transcription accuracy. Apart from physical degradation, already described in Section 5.1, handwritten marginal notes and interlinear additions further complicated the process due to their variability and overlap with printed text.

Historical typefaces, obsolete characters, and non-standard orthographic conventions reduced recognition performance, especially when combined with print deterioration. The corpus' multilingual and diachronic nature—spanning modern and historical languages—introduced additional variability in vocabulary, syntax, and character sets, increasing the likelihood of errors (Dereza et al., 2024). Complex layouts also posed challenges: footnotes and tables were prone to segmentation errors and structural misinterpretation.

To improve textual quality, Large Language Models (LLMs) were employed in post-processing (Thomas et al., 2024). Rather than applying blind correction, the model performed contextual disambiguation to distinguish OCR artifacts from historically valid spellings, enabling selective corrections that reduced noise while preserving philological authenticity.

6. Geographic Entity Extraction

A key task in the information extraction pipeline developed within the CorVo corpus project consists in the identification and structuring of geographic entities (toponyms) mentioned in historical documents related to volcanic phenomena, and in particular to the eruption of the Vesuvius. These sources, digitized through scanning and processed with Optical Character Recognition (OCR), exhibit the typical challenges of historical textual data, including recognition noise, inconsistent spelling, and fragmented layout, all of which complicate automatic geoparsing.

Traditional Named Entity Recognition (NER) systems typically focus on broad-grained geographic categories such as cities, countries, rivers, mountains, and other major geopolitical or natural entities. While this level of abstraction is sufficient for many general-purpose NLP applications, it proves inadequate in the context of historical volcanological studies, where references often concern highly specific and localized spatial entities.

[3] https://opencv.org/

[4] https://mistral.ai/news/mistral-ocr

Figure 1: Two examples of degraded pages.

In particular, historical accounts of the eruption of the Vesuvius frequently mention micro-toponyms and fine-grained spatial markers that are crucial for reconstructing the spatial dynamics of the event, patterns of destruction, and settlement structures. Standard NER taxonomies do not usually distinguish such detailed categories, nor do they explicitly model the diversity of built, infrastructural, and geomorphological entities described in early modern sources.

A previous study in the volcanological domain (Paolillo et al., 2016) identified a set of entity categories that are especially relevant for this type of historical-spatial analysis. These include architectural structures (e.g., palace, villa, convent, church, basilica, chapel, fortress/tower, hospital, clinic, university, café, restaurant, thermal baths), infrastructural elements (e.g., aqueduct, bridge, road, mills), settlement and administrative units (e.g., district, neighborhood, rione), archaeological or historical sites (e.g., ruins, epitaph), as well as fine-grained geomorphological features (e.g., canal, ravine, ridge (cognoli), ditch, plain, sinkhole, rock formations). Additional categories include entities such as masseria, atrium, rocks (scogli), and other locally specific denominations frequently attested in the corpus.

As a consequence, the extraction step cannot be framed as a generic "extract all named entities" task. Instead, the Large Language Model is prompted with a structured and domain-aware instruction that explicitly enumerates the entity types of interest. The full prompt specification, including the detailed list of categories and formatting constraints, will be reported in the Appendix of the camera-ready. This design ensures that the model's output aligns with the analytical requirements of historical volcanological research, capturing both conventional toponyms and domain-specific spatial entities that would otherwise be overlooked by standard NER pipelines.

6.1. Toponym Extraction with Large Language Models

The first stage of the process focuses on the automatic extraction of geographic named entities, aiming at identifying place names occurring in the OCR-transcribed text. In this work, preliminary toponym recognition is performed using a large

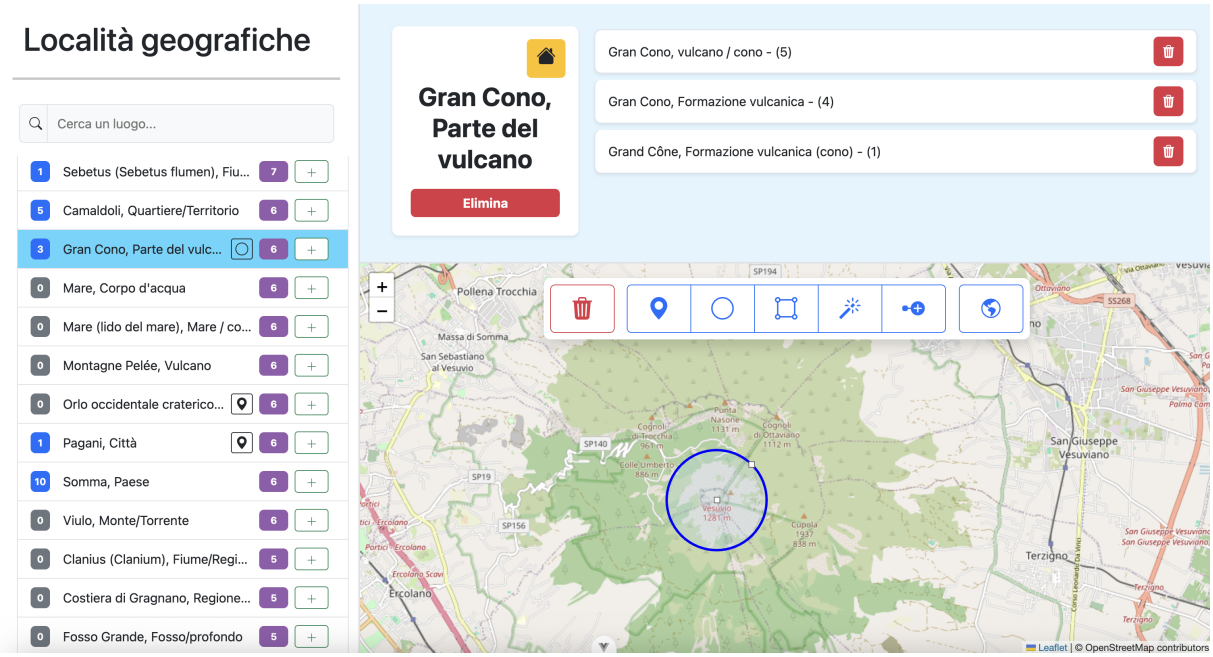


Figure 2: The annotation interface.

language model (LLM), specifically Mistral Large,[5] leveraging its ability to handle noisy textual input and to generalize over variant historical forms.

Unlike traditional Named Entity Recognition approaches, which often require substantial annotated corpora and struggle with OCR-induced errors, the LLM-based method provides a flexible alternative for bootstrapping entity candidates in low-resource and domain-specific contexts. The extracted toponyms are then collected as structured mentions, serving as input for the subsequent human-in-the-loop validation and georeferencing stage.

To illustrate the potential of extracted toponyms for volcanological interpretation, we consider a case study based on historical accounts of the 1631 Vesuvius eruption, focusing on the site of the Madonna del Soccorso in Portici. The automatic extraction of place names enables the alignment of textual evidence with known eruptive dynamics. In this case, multiple sources consistently associate the location with the passage of pyroclastic density currents (PDCs), described as fast-moving, high-temperature flows causing widespread destruction but locally diverted by morphology. Furthermore, the spatial clustering of place mentions supports the reconstruction of flow paths and depositional patterns, which can be compared with geological evidence. This example demonstrates how toponym extraction from historical narratives provides complementary data for volcanological analysis, contributing to the reconstruction of eruptive sequences, hazard mapping, and interpretation of past events beyond what can be inferred from deposits alone.

[5] https://mistral.ai/it/news/mistral-large

7. Human-in-the-loop Geospatial Annotation

Given the ambiguity and variability of historical toponyms, automatic extraction alone is insufficient to produce reliable geospatial references. Place names may refer to the same location under different spellings, may correspond to obsolete administrative divisions, or may be used metonymically. For this reason, the CorVo corpus integrates a dedicated annotation interface (see Figure 2) designed to support expert validation, entity consolidation, and geolocation.

The annotation tool provides two core functionalities:

- Toponym merging, allowing annotators to unify multiple extracted mentions that refer to the same geographic entity (e.g., alternate spellings or synonymous references).
- Geospatial grounding, enabling the association of each entity with explicit geographic coordinates.

7.1. Coordinate Selection and Georeferencing

The interface supports several mechanisms for selecting the geographic position of an entity. Annotators may:

- directly select a point on the interactive map,
- query the Google Maps geocoding service for automated suggestions,
- manually enter latitude and longitude coordinates,
- or navigate freely across the map until the correct location is found.

This flexibility allows efficient handling of both well-known locations and obscure or historically shifted place references.

7.2. Multi-geometry Spatial Annotation Modes

In addition to selecting coordinates, the platform allows annotators to represent geographic entities according to different spatial configurations. In particular, the interface supports multiple annotation geometries:

- Single-point annotation, suitable for discrete landmarks such as buildings or archaeological sites.
- Circular area annotation, defined by a center point and radius, typically used for cities or settlements with an approximate spatial extent.
- Polygonal area annotation, where users manually define the vertices of a polygon, appropriate for larger regions or administrative territories.
- Polyline annotation, representing linear geographic structures such as valleys, ravines, or volcanic flows.

By supporting different spatial representations, the system accommodates the diverse nature of geographic references in historical volcanological narratives.

The extraction process yielded 812 distinct geographical toponyms, corresponding to 437 unique locations (such as cities, regions, mountains, rivers).

The second extraction phase, targeting fine-grained spatial entities, resulted in 2,964 annotated locations, categorized as listed in Table 2.

Type	Count	Type	Count
palace	85	thermae	6
villa	88	ruins	62
epitaph	9	canyon	96
convent	95	trench	202
church	291	plain	380
basilica	2	floor	222
aqueduct	7	masseria	41
fortress/tower	87	mill	3
district	390	bridge	39
hospital	9	reef/rock	29
tavern	8	road/railway	311
university	53	chasm	449

Table 2: Number of specific entities found in the dataset.

7.3. From Annotated Toponyms to Diachronic Geovisualization.

Building upon the annotated dataset described above, we construct a diachronic geovisualization that represents how different eruptions of the Vesuvius are reflected in historical sources through the places they mention (see Figure 3). For each annotated entity, all textual occurrences across the corpus are aggregated and grouped according to predefined temporal intervals (e.g., centuries). However, a direct comparison of raw mention counts across periods would be misleading, as the number of available texts varies substantially from one century to another. A period with a larger number of documents would naturally produce higher absolute frequencies, independently of the actual prominence of specific locations.

To address this imbalance, we apply a *temporal normalization* procedure. For each time interval, the total number of geographic mentions in the corpus is computed. The frequency of each individual location within that period is then divided by the total number of mentions for the same period, producing a relative measure. Formally, if $f_{i,t}$ denotes the number of mentions of location i in period t, and T_t the total number of geographic mentions in that period, the normalized value is computed as:

$$n_{i,t} = \frac{f_{i,t}}{T_t} \tag{1}$$

This normalization yields comparable proportions rather than absolute counts, enabling a meaningful diachronic comparison across periods with heterogeneous textual coverage. The normalized values are subsequently used to generate an interactive cartographic representation in which each annotated location is positioned according to its validated coordinates and visually encoded with its temporal distribution. In the implemented visualization, each location is represented by a georeferenced marker enriched with temporal information, allowing users to perceive at a glance how

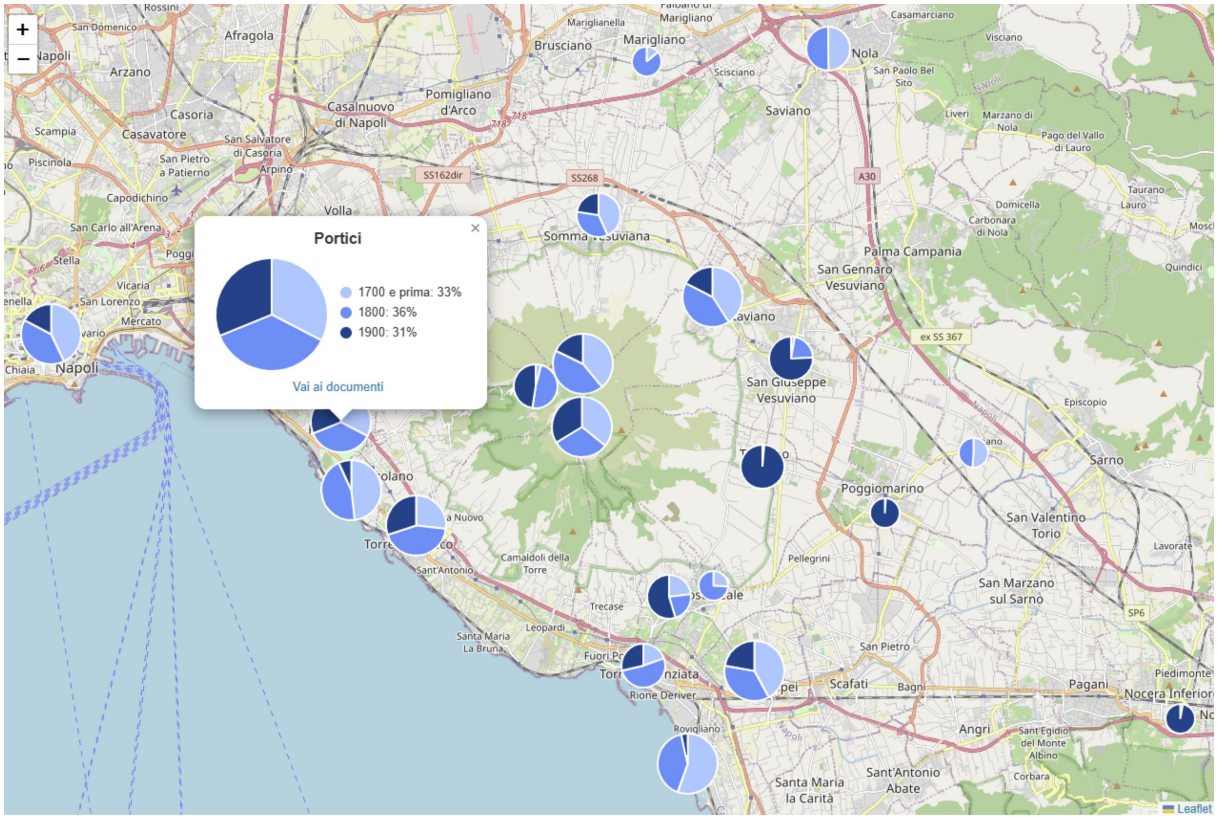


Figure 3: The diachronic visualization.

the relative prominence of that place varies across historical periods.

Through this pipeline, the manual geospatial annotations serve not only as a corrective layer for entity extraction, but as the enabling infrastructure for higher-level interpretative representations. The resulting system does not attempt to reconstruct the physical evolution of eruptions directly; rather, it reconstructs their *textual-spatial footprint* across time, offering an intuitive way to explore how different areas around the Vesuvius emerge, recede, or persist in historical narratives.

The manual annotation process was conducted by a single expert annotator with extensive domain knowledge in volcanology and long-standing fieldwork experience. One of the authors is a specialist in Vesuvian volcanology and has contributed to multiple scientific studies on the area, including the production of geological maps (Principe et al., 2013), providing further domain validation of the annotation process.

8. Release

The CorVo corpus and the software components developed within the project are released on Github[6] under open and permissive licenses to ensure transparency, reproducibility, and reuse by the research community.

8.1. Data

The released dataset includes:

- the OCR-cleaned textual transcriptions of the digitized historical documents contained in the corpus, with the exception of a limited number of recent documents whose copyright has not yet expired;
- the extracted geographic entities, including both toponyms and fine-grained domain-specific spatial entities;
- the associated categorical labels (e.g., architectural, infrastructural, geomorphological, administrative categories);
- the validated geographic information resulting from the human-in-the-loop annotation process, including coordinates and geometry types (point, circular area, polygon, polyline).

[6] https://github.com/ziorufus/corvo-project

In total the raw dataset contains around 16,000 pages and 4,160,000 words, in which 2,964 toponyms are annotated with spatial information (437 unique locations).

All data will be distributed under the Creative Commons Attribution 4.0 International license (CC BY 4.0),[7] allowing reuse, redistribution, and adaptation for both research and educational purposes, provided appropriate credit is given to the project.

8.2. Software

In addition to the dataset, the source code of the software tools developed for the project is publicly released in the same Github project. These include, both implemented with a Vue/JavaScript frontend and a Python/SQLite backend:

- the geographic annotation interface, designed to support entity consolidation and geospatial grounding;
- the diachronic geovisualization interface, enabling temporally normalized spatial exploration of the corpus.

The source code is distributed under the Apache License 2.0,[8] to allow modification, redistribution, and integration into other systems, including commercial applications, while preserving attribution and license notices.

9. Conclusion and Future Work

This project aimed to evaluate the financial, technical, and methodological feasibility of using documents describing past volcanic activity to address pressing issues in modern volcanology, such as the assessment and mitigation of volcanic risk in Italy. For this purpose, Vesuvius was selected, due to the abundance and diachronic nature of the descriptions of its activity; the three most representative eruptions of the three possible reactivation scenarios were selected, and 180 related digitized documents were collected, from which geographical entities were extracted, and a graphical interface was developed to represent them.

The use of historical linguistics with the most up-to-date technologies has been shown to be an effective tool for preserving precious linguistic data that survived from old periods. The greatest difficulties have thus far been encountered in building the corpus of documents. This difficulty could be largely overcome in the future through the ongoing digitization of their documentary collections, a process now underway at many libraries. This will enable the specialist to work on select texts to form targeted corpora, such as the one we have built in this project, without the burden of researching and digitizing the original texts.

Beyond the specific case study on Vesuvius, an important outcome of this work lies in the generality of the proposed framework. The combination of OCR preprocessing, LLM-based post-correction, domain-aware entity extraction, and human-in-the-loop geospatial annotation defines a modular pipeline that can be readily adapted to other domains and datasets. In particular, the tools and methodologies described here enable researchers to initiate similar studies from scratch on different volcanic systems (or, more broadly, on other types of geographically grounded historical phenomena) without requiring pre-existing structured resources. By reconfiguring the entity schema and adapting the annotation guidelines to a new domain, the same infrastructure can support the systematic extraction and diachronic spatial analysis of knowledge from heterogeneous historical corpora, thus opening the way to comparative and cross-regional studies.

10. Limitations

The corpus is restricted to printed documents of sufficient physical quality, excluding handwritten materials and severely degraded sources, which limits coverage of the historical record. OCR errors, although mitigated through preprocessing and LLM-based correction, may still propagate to downstream extraction tasks. The toponym extraction pipeline relies on large language models without gold-standard annotated training data, and therefore may introduce inconsistencies or omissions, particularly for rare or highly ambiguous entities. Furthermore, georeferencing depends on expert validation, which, while improving reliability, constrains scalability. Finally, the diachronic visualization reflects the textual footprint of eruptions rather than their physical evolution, and is therefore influenced by historiographical bias and uneven document distribution across centuries.

An additional limitation concerns the absence of a systematic quantitative evaluation of the NLP components in the current phase of the project. While the pipeline integrates OCR preprocessing, LLM-based post-correction, and domain-aware toponym extraction followed by expert validation, no gold-standard annotated subset has yet been constructed to compute standard metrics such as precision, recall, F1-score, or character error rate reduction. As a result, the performance of the automatic extraction stages has not been formally benchmarked against existing NER systems or al-

[7] https://creativecommons.org/licenses/by/4.0/

[8] https://www.apache.org/licenses/LICENSE-2.0

ternative approaches. Future work will address this gap by developing a manually annotated evaluation dataset, enabling rigorous assessment of extraction accuracy, error propagation across pipeline stages, and the overall reliability of the system.

11. Bibliographical References

Pietro Bembo. 1525. *Prose di M. Pietro Bembo nelle quali si ragiona della volgar lingua*. Giovanni Tacuino, Venezia. Editio princeps.

Margherita Berardi, Luigi Santamaria Amato, Francesca Cigna, Deodato Tapete, and Mario Siciliani de Cumis. 2022. Text Mining from Free Unstructured Text: An Experiment of Time Series Retrieval for Volcano Monitoring. *Applied Sciences*, 12(7).

A. Bertagnini, R. Cioni, E. Guidoboni, M. Rosi, A. Neri, and E. Boschi. 2006. Eruption early warning at Vesuvius: The A.D. 1631 lesson. *Geophysical Research Letters*.

E. Boschi, G. Ferrari, P. Gasperini, E. Guidoboni, G. Smriglio, and G. Valensise. 1995. *Catalogo dei forti terremoti in Italia dal 461 a.C. al 1980*. ING-SGA.

S. Branca, E. De Beni, and C. Proietti. 2013. The large and destructive 1669 AD eruption at Etna volcano: reconstruction of the lava flow field evolution and effusion rate trend. *Bulletin of Volcanology*, page 1–16.

S. Chen, W. Hua, X. Liu, X. Deng, X. Zeng, and J. Duan. 2022. Chinese Fine-Grained Geological Named Entity Recognition With Rules and FLAT. *Earth and Space Science*, 9(12).

P.D. Cole and C. Scarpati. 2010. The 1944 eruption of Vesuvius, Italy: combining contemporary accounts and field studies for a new volcanological reconstruction. *Geological Magazine*, 147(3):391–415.

Runyu Fan, Lizhe Wang, Jining Yan, Weijing Song, Yingqian Zhu, and Xiaodao Chen. 2020. Deep Learning-Based Named Entity Recognition and Knowledge Graph Construction for Geological Hazards. *ISPRS International Journal of Geo-Information*, 9(1).

Mara Finkelstein, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Geza Kovacs, Cole Dilanni, Colin Cherry, Eleftheria Briakou, Elizabeth Nielsen, Jiaming Luo, Kat Black, Ryan Mullins, Sweta Agrawal, Wenda Xu, Erin Kats, Stephane Jaskiewicz, Markus Freitag, and David Vilar. 2026. TranslateGemma Technical Report.

Emanuela Guidoboni and John E Ebel. 2009. *Earthquakes and tsunamis in the past: A guide to techniques in historical seismology*. Cambridge University Press.

Takashi Kudo and Hideo Hoshizumi. 2006. Construction of a new catalog of eruptive events during the last 10,000 years in Japan (V101-P029) (poster session) (abstract). In *Abstracts, Japan Geoscience Union Meeting (CD-ROM)*, volume 2006, pages V101–P029, Japan. Japan Geoscience Union. Poster abstract.

Christopher J.M. Lawley, Stefania Raimondo, Tianyi Chen, Lindsay Brin, Anton Zakharov, Daniel Kur, Jenny Hui, Glen Newton, Sari L. Burgoyne, and Geneviève Marquis. 2022. Geoscience language models and their intrinsic evaluation. *Applied Computing and Geosciences*, 14:100084.

Kai Ma, Miao Tian, Yongjian Tan, Qinjun Qiu, Zhong Xie, and Rong Huang. 2023. Ontology-Based BERT Model for Automated Information Extraction from Geological Hazard Reports. *Journal of Earth Science*, 34:1390–1405.

Rafael O. Nunes, Andre S. Spritzer, Dennis G. Balreira, Carla M. D. S. Freitas, and Joel L. Carbonera. 2024. An Evaluation of Large Language Models for Geological Named Entity Recognition. In *2024 IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 494–501.

J. Padarian and I. Fuentes. 2019. Word embeddings for application in geosciences: development, evaluation, and examples of soil-related concepts. *SOIL*, 5(2):177–187.

Annarita Paolillo, Claudia Principe, Marina Bisson, Roberto Gianardi, Daniele Giordano, and Sonia La Felice. 2016. Volcanology of the Southwestern sector of Vesuvius volcano, Italy. *Journal of Maps*, 12(sup1):425–440.

Claudia Principe. 1990. ADSVI – Archivio Documenti Storici sui Vulcani Italiani (IGG internal technical report).

Claudia Principe, Daniele Giordano, Marina Bisson, Annarita Paolillo, and Roberto Gianardi. 2013. Volcanological map of the South-Western sector of Vesuvius between Torre del Greco and Ercolaneum.

Qinjun Qiu, Zhong Xie, Liang Wu, and Liufeng Tao. 2020. Automatic spatiotemporal and semantic information extraction from unstructured

geoscience reports using text mining techniques. *Earth Science Informatics*, 13.

Lorenzo Renzi and Giampaolo Salvi. 2010. *Grammatica dell'italiano antico*. Il Mulino, Bologna. 2 vols.

H. Sigurdsson, S. Carey, and W. Cornell. 1985. The Eruption of Vesuvius in A.D. 79. *National Geographic Research*, 1(3):332–387.

Alan Thomas, Robert Gaizauskas, and Haiping Lu. 2024. Leveraging LLMs for post-OCR correction of historical newspapers. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 116–121, Torino, Italia. ELRA and ICCL.

B. Voight, E.K. Constantine, S. Siswowidjoyo, and R. Torley. 2000. Historical eruptions of Merapi Volcano, Central Java, Indonesia, 1768–1998. *Journal of Volcanology and Geothermal Research*, 100(1):69–138.

12. Language Resource References

Oksana Dereza, Deirdre Ní Chonghaile, and Nicholas Wolf. 2024. "To Have the 'Million' Readers Yet": Building a Digitally Enhanced Edition of the Bilingual Irish-English Newspaper an Gaodhal (1881-1898). In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 65–78, Torino, Italia. ELRA and ICCL.

S. Giannini and A. Paolillo. 2018. BIBV V: "Bibliografia storica dei vulcani attivi italiani". Database (IRIS CNR). Accessed via IRIS CNR Institutional Research Information System.

Lucas Lima de Oliveira, Regis Kruel Romeu, and Viviane Pereira Moreira. 2021. REGIS: A Test Collection for Geoscientific Documents in Portuguese. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2363–2368, New York, NY, USA. Association for Computing Machinery.

When Lexicographic Quotations Become a Corpus: To Deduplicate or Not to Deduplicate?

Manuel Favaro[§], Elisa Guadagnini[§], Eva Sassolini[§], Marco Biffi*^, Simonetta Montemagni[§]

[§]Istituto di Linguistica Computazionale "Antonio Zampolli" – CNR
*Università di Firenze
^Accademia della Crusca
[§]name.surname@ilc.cnr.it
*marco.biffi@unifi.it

Abstract

Historical dictionaries are increasingly reused as sources for diachronic language corpora. In this context, lexicographic quotations represent a valuable yet challenging type of data, as they are both editorially curated and diachronically representative. A major issue in their computational reuse is the presence of duplicate and near-duplicate quotations. This paper addresses quotation deduplication in corpora derived from lexicographic resources. We introduce QRD (Quotation Reuse Detection), a multi-stage pipeline designed to identify, compare, and cluster quotations based on graded similarity rather than binary matching. The approach combines string-based similarity measures, iterative threshold analysis, and clustering, enabling both quantitative and qualitative investigation of quotation reuse. Our results show that deduplication in this context cannot be reduced to the automatic elimination of redundant data. The variability observed in the quotations - ranging from OCR-related noise to substantial editorial variation - reflects both technical and structural factors and calls for a more nuanced approach. QRD supports the identification of OCR-related errors and reveals patterns of textual reuse underlying the compilation of the dictionary. We argue that quotation deduplication should be conceived primarily as a task of identification and clustering. This perspective reframes deduplication from a data-cleaning operation into an analytical methodology for historically and editorially curated textual resources.

Keywords: Historical Corpora, Text Deduplication, Data Matching Process, Historical Lexicography

1. Introduction

Historical dictionaries are among the most important resources for the study of language over time, as they provide a structured interpretation of the different layers of lexical information (semantics, register and usage labels, multiword expressions, etc.), grounded in diachronic textual data. In these resources, authorial quotations play a central role: they document actual language use, support lexicographic decisions, and offer a bridge between linguistic description and primary textual sources (Atkins and Rundell, 2008). For this reason, quotation-based resources derived from historical dictionaries are increasingly reused in digital humanities and natural language processing (NLP) research, including diachronic corpus studies (see e.g. Hoffmann, 2004; Rohdenburg, 2013).

However, the reuse of lexicographic quotations as textual corpora raises a number of data quality issues that are often overlooked. One of the most prominent among these is the presence of duplicate or near-duplicate quotations. In historical dictionaries, it is frequently the case that the same textual passage appears multiple times across different entries, possibly in slightly different forms. Such duplication can result from the reuse of canonical sources, from editorial adaptation of quotations, or from the inheritance of citation material across lexicographic traditions (Béjoint, 2010). From a computational perspective, unaddressed duplication can affect quantitative analyses, bias frequency counts, and reduce the reliability of downstream NLP tasks using this type of resources (Lee *et al*., 2022). At the same time, in this specific context, duplication should not be considered merely as noise: repeated quotations reflect the centrality of canonical sources, and therefore need to be preserved and made explicit rather than removed indiscriminately.

In this paper, we address the open issue of how to treat duplicated quotations in the creation of a corpus extracted from the examples in a historical dictionary of Italian. This work is carried out as part of ongoing efforts aimed at the digitization, organization, and computational exploitation of the *Grande Dizionario della Lingua Italiana* ('Great Dictionary of the Italian Language', henceforth GDLI) (Sassolini *et al*., in preparation). In particular, the work presented here builds on and extends Favaro *et al*. (2022), which describes the creation of a diachronic corpus of Italian based on GDLI quotations, focusing on their extraction from a structured digital representation of the dictionary and on subsequent linguistic annotation. While that work primarily addresses corpus construction and annotation, it also implicitly highlights data quality issues that become crucial when quotations are reused as corpora.

The resulting resource, hereafter referred to as the *GDLI Quotations Corpus* (GDLI-QC), shares

Proceedings of LT4HALA 2026 @ LREC 2026, pages 49–57.
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

the main features of the original lexicographic resource, such as editorial selection, diachronic breadth, and textual inconsistencies, but also exhibits properties introduced by the digitization process, including OCR-related noise. Within the corpus building process, a key issue concerns the treatment of duplicate or near-duplicate quotations. In tackling this issue, we explore why and where duplicates occur, how they can be detected, and what to do about them. This issue is analysed from a threefold perspective, namely with respect to: the identification of OCR-related errors occurring across identical or near-identical textual passages; quantitative analyses of the corpus; and, last but not least, the impact of duplicate examples for training of linguistic annotation models.

Taken together, these considerations indicate that duplication in lexicographic quotation corpora cannot be addressed through simple removal strategies, but instead requires an approach that accounts for both computational and editorial dimensions. In response to the issues outlined above, we adopt a method that does not eliminate redundancy indiscriminately, but aims to identify clusters of quotations deriving from the same underlying textual passage. This strategy allows us to improve corpus quality while preserving editorial variation, maintaining transparency with respect to lexicographic practices, and supporting scholarly inspection. On this basis, we propose a transparent multi-stage duplication processing pipeline specifically tailored to editorially curated lexicographic data.

The remainder of the paper is organized as follows: Section 2 presents the main contributions of this work; Section 3 reviews related work on text deduplication and historical language data; Section 4 describes the GDLI quotation corpus, and Section 5 the proposed methodology; Section 6 discusses the evaluation and qualitative analysis; and Section 7 concludes the paper.

2. Contributions

This paper makes the following contributions:

Contextualized problem definition - We frame the identification of duplicate lexicographic quotations as an issue of data complexity, one that requires both an analysis of the underlying lexicographic and editorial practices and an assessment of how duplication affects the transformation of a historical dictionary into a textual corpus suitable for use as an independent linguistic resource.

Task-driven motivation - We explicitly link this topic to three concrete objectives relevant for the construction and reuse of historical language resources: (i) supporting the identification of recurring OCR-related errors in digitized historical data; (ii) evaluating their effect on quantitative analyses derived from quotation-based corpora; (iii) assessing the impact of repeated quotations on training linguistic annotation models.

Methodological contribution - We propose a transparent multi-stage pipeline tailored to handle duplication of examples in editorially curated historical lexicographic data.

Resource-oriented perspective - Rather than treating deduplication as a purely eliminative step, we demonstrate how it can be used to identify clusters of quotations deriving from the same underlying textual passage, thereby enhancing corpus quality while preserving editorial variation and enabling scholarly inspection.

3. Related Work

Text deduplication can be framed as a specific instance of the more general data matching task (Christen, 2012), which aims at identifying, matching, and possibly merging items that correspond to the same underlying entity across one or more data sources, typically databases. While data matching has most commonly been applied to entities such as individuals (e.g. patients, customers, or taxpayers), it has also been extended to bibliographic records and textual data. In the latter case, the goal is to identify redundant documents or passages within a single collection or across multiple textual collections, which is the focus of the present work.

Since the early work by Broder (1997), text deduplication has been extensively studied in information retrieval and corpus construction, where it is commonly employed to remove redundant documents or passages and to improve overall data quality.

A comprehensive overview of the challenges and solutions in duplicate detection is provided by Elmagarmid *et al*. (2007). As they observe, duplicate records often lack a common key and may contain various types of errors (such as transcription mistakes, incomplete information, or inconsistent formatting) that make matching particularly challenging. To address these issues, a wide range of similarity measures has been proposed, including edit distance, token overlap, and character n-gram-based approaches, often combined with hashing or blocking techniques to ensure computational efficiency and scalability. These methods form the basis of many contemporary deduplication pipelines.

In more recent NLP research, deduplication has gained renewed importance due to its impact on language model training and evaluation. Several studies have shown that duplicate or near-duplicate texts can significantly bias empirical results and model behavior (Lee *et al*., 2022). As a consequence, deduplication is increasingly regarded as a standard pre-processing step in the creation of language resources.

Text deduplication in historical language data poses additional challenges due to orthographic variation, diachronic change, and editorial or transcriptional noise. Prior work in historical NLP emphasizes the need for normalization strategies and noise-robust representations, as well as for multi-stage pipelines that separate candidate generation from fine-grained similarity assessment (Bollmann, 2019; Piotrowski, 2012). In this context, deduplication may also serve as a tool for identifying systematic digitization errors, such as recurring OCR mistakes across repeated textual passages, a potential use that has received limited attention in previous work.

Unlike most previous approaches, which focus on entire documents or extended textual passages, the present work applies duplicate detection to GDLI quotations, that is, relatively short spans of text covering a limited number of sentences and often containing internal interruptions. This specificity introduces additional challenges, which are discussed in the remainder of the paper.

4. The GDLI Quotation Corpus

The GDLI is the most authoritative dictionary of the Italian language. Initially edited by Salvatore Battaglia and subsequently by Giorgio Barberi Squarotti, the *GDLI* comprises 21 volumes published between 1961 and 2002, amounting to more than 23,000 pages and approximately 200,000 headwords.

The lexicographic description provided by the GDLI is grounded in an exceptionally rich apparatus of quotations, ranging from the earliest extant Italian documents, such as the «Placiti capuani» of the 10th century, to contemporary usage. While retaining a strong overall literary orientation, the dictionary exhibits considerable diatextual variation: alongside lyric and narrative texts, it includes hagiographic and technical–theoretical works, as well as practical and legislative texts, in addition to material drawn from periodical publications. It should also be noted that the dictionary, published over a period of 41 years, progressively increased both the number and the range of texts excerpted, thereby significantly expanding the coverage of the Italian language. In particular, this expansion involved a broadening along the dimensions of diatopy (including dialects and regional varieties of Italian), diastraty (extending to popular and semi-educated varieties), and diaphasy (encompassing specialized and domain-specific languages). The corpus of quotes draws on passages extracted from more than 6,000 authors and nearly 14,000 distinct sources (excluding quotations taken from periodical press; cf. Biffi and Guadagnini, 2022; Biffi *et al.*, 2025). Each quotation is selected so as to constitute a syntactically complete unit and often includes a relatively extended textual span. Overall, the corpus is estimated to include more than 2,000,000 cited passages, amounting to over 40 million tokens and covering a chronological range of more than one thousand years.

GDLI quotations were automatically extracted from the TEI XML version of the dictionary, which was obtained through a semi-automatic conversion process designed at structuring the dictionary contents starting from the OCRed source. The overarching goal of this process is to obtain a detailed and articulated representation of GDLI entries by means of a sequence of iterative steps, each progressively refining and organizing the previously identified dictionary structure.

The general approach to the extraction and structuring of GDLI contents (described in Sassolini *et al.*, 2019, 2021, and Biffi *et al.*, 2020) is largely based on pattern-matching techniques. The identification criteria rely on a wide range of features, including page layout characteristics and structural information related to the different components of the lexical entry. The ultimate goal of this process is the conversion of the dictionary contents into structured macro-fields, which are subsequently mapped onto the TEI XML standard.

Quotation extraction constitutes one of the steps of this iterative workflow. The TEI XML encoding of the GDLI quotation macro-field is exemplified in Figure 1, which shows the automatically generated TEI XML representation of the entry *abiatico* ‘grandchild’. It can be noticed that for each sense the set of quotations is encoded using the `<cit>` element, which contains one or more `<bibl>/<quote>` pairs. These elements respectively encode a loosely structured bibliographic reference and the quotation text itself. The extraction process has so far been applied to volumes I–X of the dictionary (out of a total of XXI), for which the manual revision of entry segmentation has been completed.

In transforming GDLI quotations into a textual corpus, each quotation is associated with metadata concerning authorship and the headword under which it appears. The corpus currently includes all the quotations from volumes I–X, as previously noted, for a total of more than 830,000 quotations and more than 18,000,000 tokens. At this stage, quotations are extracted as independent textual units, without attempting to determine whether identical or near-identical passages occur elsewhere.

The extracted quotation set turned out to contain multiple instances of the same underlying textual passage. The most complex and interesting cases, however, involve a substantial number of quotations that differ in certain elements, while still exhibiting a clear degree of similarity. This form of duplication can be attributed to two main factors. On one hand, noise arising from OCR errors introduced near-identical or slightly corrupted strings during the digitization process.

On the other hand, many redundancies stem from editorial practices: the GDLI editors frequently reused the same authorial quotations to illustrate different lexical entries. In these instances, the text was often re-elaborated, truncated, or adapted to better suit the specific requirements of the entry, leading to a complex layer of intentional textual variation.

The resulting GDLI-QC, covering nearly half of the dictionary volumes, therefore constitutes a suitable testbed for investigating the nature and extent of quotation duplication, as well as for developing and evaluating deduplication strategies specifically tailored to historical lexicographic data.

```
<entry>
    <form type="lemma">
        <orth>Abiàtico</orth>
    </form>
    <sense level="1" n="1">
        <def>(abbiàtico, aviàtico) sm. (femm. abiòtica, abbiàtica) Dial.
        Nipote (figlio del figlio o della figlia).</def>
        <cit>
            <bibl> Comp. Antico Testamento [Tommaseo]:</bibl>
            <quote> Seppe... come 10 aveva trovato in lo bosco e comprese
            che questo era suo abiatico.</quote>
            <bibl>Vite di imperatori romani [Tommaseo] :</bibl>
            <quote> Tiberio voleva piuttosto che succedesse questo suo
            abiatico che questo Gaio.</quote>
            <bibl>Idem [Tommaseo]:</bibl>
            <quote> La figlia e le abiatiche fece usare all'esercizio
            della lana.</quote>
            <bibl>Fogazzaro, 5-18:</bibl>
            <quote> A poppa sotto la bandiera, v'era seduto don Franco
            Maironi, l'abiatico della vecchia marchesa.</quote>
        </cit>
    </sense>
    <sense level="1" n="2">
        <def>2. Antenato.</def>
        <cit>
            <bibl>Moretti, 32-116:</bibl>
            <quote> Camere... piene di tante goffaggini, oleografie,
            ritratti dei padroni di casa, d'abiatici, che non sono stati
            neppure Reggenti.</quote>
        </cit>
    </sense>
    <etym>= Lat. mediev. abiatlcus o aviatlcus (nella Lex Longobardorum,
    del sec. XXXX dal lat. class, avus (attraverso l'agg. 'avius :
    dell'avo, nipote dell'avo). È voce dell'Italia settentrionale.</etym>
</entry>
```

Figure 1: TEI XML representation of the GDLI entry *abiatico* 'grandchild'.

5. Methodology

The methodology we devised for dealing with GDLI-QC duplicate quotations consists of three different phases, respectively aimed at:

1. minimizing differences due to systematic OCR errors;
2. carrying out quotation pair comparison;
3. clustering quotations identified as potentially matching at the previous step.

The first stage involved a rigorous data-cleaning process to minimize OCR errors as much as possible. We employed a series of regular expressions (*regex*) to perform systematic normalization and correct recurring OCR errors in the quotation text. This phase was crucial to ensure that subsequent fuzzy matching would not be skewed by predictable mechanical discrepancies.

Table 1 provides a few examples of the systematic corrections implemented during this stage. As one can see, key interventions include resolving the graphic overlap between the digraph <ll> and high-density uppercase characters (e.g., U, H), as well as correcting spurious whitespace following hyphenation. These patterns facilitated a hybrid workflow of automatic and semi-automatic character substitution, significantly improving the lexical integrity of the digital resource.

Regex	Output	Correct string
([a-z])([A-ZÀÈÉÌÒÙ])([''])	"deU'adrie"; "deH'esemplare"; "aU'appianarmeli"	"dell'adrie" "dell'esemplare"; "all'appianarmeli"
(<quote>.*?)(\b[a-z]{2,})-(\s)([a-z]{2,}\b)(.*?</quote>)	"con- clusione; "; "abbas- sare"	"conclusione"; "abbassare"

Table 1: Examples of OCR systematic corrections

In the second step, quotation pairs were compared by computing string similarity using the Levenshtein distance (Levenshtein, 1966; Navarro, 2001). As a starting point, we considered the FuzzyWuzzy algorithm[1], which has been successfully employed in previous work on historical language varieties, such as BERToldo (Palmero Aprosio *et al.*, 2022). We ultimately adopted RapidFuzz, a C++ and Python reimplementation that provides substantially higher computational efficiency, enabling iterative analyses that would otherwise be impractical.

Importantly, we adapted the original tool to function not merely as a filtering mechanism for duplicate removal, but as a diagnostic instrument capable of producing fine-grained similarity information. To this end, punctuation and diacritical marks were removed prior to similarity computation. A key aspect of our approach is the iterative analysis of quotation pairs across discrete similarity intervals of 5%. This choice reflects the observation that exact duplicates (100% similarity) constitute a relatively small and easily manageable portion of the data, whereas the methodological challenges primarily arise in cases of partial similarity (near-duplicates).

Accordingly, we focused our analysis on similarity ranges between 70% and 99.99%, which proved crucial for identifying and characterizing different types of near-duplicate quotations. This subdivision allowed for a qualitative investigation going beyond the binary "duplicate/non-duplicate" dichotomy, leading to the definition of an articulated typology of duplicate examples, described in detail in Section 6. The 70% threshold was established through preliminary empirical testing on a random sample of 500

[1] https://github.com/xdrop/fuzzywuzzy

pairs. Lower thresholds (e.g., 60–69%) were found to introduce excessive noise, primarily due to short and frequent formulaic expressions.

In the final phase, we integrated the NetworkX[2] library into the pipeline to identify clusters of quotations based on configurable similarity ranges. In addition, we implemented a difference extraction function (*diff_words*) to identify and inspect divergent string segments within clustered quotations, supporting qualitative analysis of variation and error patterns. An example from Boccaccio is reported in Table 2.

Entry	Quote
Abbattere	Commossa adunque la santa dea per le costui opere, propose di riducerlo a niente, abbattendo la infiammata sua superbia, come quella degli antecessori aveva altra volta abbattuta.
Antecessore	Commossa adunque la santa dea per le costui opere, propose di riducerlo a ***mente***, abbattendo la infiammata sua superbia, come quella degli antecessori aveva altra volta abbattuta.
Degno	Commossa adunque la santa dea per le costui opere, propose di riducerlo a niente, abbattendo la infiammata sua superbia, come quella degli antecessori aveva altra volta abbattuta, ***con degno mezzo***.

Table 2: Example of a cluster of quotations characterized by minimal variations

As shown in the table, a comparison between the first two instances, cited under the entries *abbattere* and *antecessore*, reveals a near-perfect match disrupted only by the discrepancy between *niente* and *mente*: «mente» is an OCR error which, however, results in a form that is in itself admissible in Italian. The third instance differs from the previous two in that it includes an additional segment of text—which, in this specific case, justifies the inclusion of the quotation under the entry *degno*. To provide a proper context for the target adjective, the textual span was extended to include the phrase *con degno mezzo*, an addition absent in the previous two entries.

Overall, we identified a set of approximately 91,000 quotations, including both exact duplicates and near-duplicates, which represent nearly 11% of the total number of quotations in the corpus. Table 3 highlights cluster size in relation to the number of quotations involved.

Cluster Size	Number of Clusters	Involved Quotations
2	59275	118550
3	9783	29349
4	2265	9060
5	723	3615
6	240	1440
7	94	658
8	40	320
9	18	162
10	5	50
11	3	33
13	1	13
14	1	14

Table 3: Duplicates and near-duplicates cluster distribution

The pipeline implementing the three steps described above will hereafter be referred to as QRD (Quotation Reuse Detection).

6. Typology of duplicated quotations in GDLI-QC

In this section, we present a qualitative analysis of automatically clustered quotations with a similarity score ≥ 70%, including both exact (100% similarity) and near-duplicates (similarity ranging between 70% and 99.99%). A quantitative evaluation of near-duplicate identification in terms of precision and recall is not provided, as it would require manual validation of each pair. Moreover, the current corpus - limited to 10 of 21 volumes and with source mapping still ongoing - does not support yet the construction of a representative quotation sample. The analysis therefore focuses on qualitative evidence, with quantitative assessment deferred to future work on a complete and stabilized dataset.

Exact duplicates correspond to cases in which an identical string of text is cited under different headwords: for example, the quotations of Giovan Battista Marino (the foremost Italian Baroque writer) «Tu tra questi deserti, / ond'uscir mai non spero, / inculti, abbandonati, / disleal, m'abbandoni» are cited - exactly in this form - s.vv. *abbandonare*, *abbandonato*, and *deserto*[2].

As far as near-duplicate quotations are concerned, the QRD pipeline allows us to extract different types of variation, which are illustrated below. We identified several categories of near-duplicates, ranging from simple spelling differences due to OCR errors or oversights introduced directly by the editors of the printed dictionary, to quotations that share the same semantic core but result from different excerpting strategies (textual extensions and/or reductions).

[2] https://networkx.org/en/

The first type consists of cases in which quotations that are identical in the printed dictionary appear as different strings in the digitized version due to errors introduced by OCR processing. For example, the lines by Ciro da Pers (another Baroque author) «O sonno, tu ben sei fra i doni eletti / dal **ciel** concesso ai miseri mortali; / tu l'agitato sen placido assali / e tregua apporti ai combattuti affetti» cited s.vv. *agitato*, *apportare*[1]*, assalire*, and *combattuto*, and «O sonno, tu ben sei fra i doni eletti / dal **del** concesso ai miseri mortali; / tu l'agitato sen placido assali / e tregua apporti ai combattuti affetti» s.v. *affetto*[1] (in which, as can be seen, «ciel» has been erroneously OCRed as «del»). Note that this type of OCR error was not dealt with in the pre-processing stage, where the focus was on systematic OCR errors, which could be automatically corrected.

Another type of near-duplicate quotations is represented by those that already appear as such in the *GDLI* original printed version, where the difference is not simply due to an OCR error but because of a transcription error by the entry editor. For example, again citing Marino, the quote «Gli sguardi e l'orme / a le mura superbe intento gira, / e mentre queste ed altre illustri forme, / di cui son tutte effigiate, ammira, / sembra, né sa s'ei **vegghia** o pur s'ei dorme, / statua animata; imagine che spira» (s.v. *animato*) appears s.v. *ammirare* with the reading «veglia» instead of «vegghia». The cited source gives the reading «vegghia,» but both forms are admissible in Italian for the third-person singular of the verb *vegliare* ('to be awake'), and this explains the quotation error introduced s.v. *ammirare*. This type of case is also automatically extracted: the variant that distinguishes the quotes correctly detected as repeated is present in the printed original (thus raising the question of whether it would be legitimate to modify it), and, naturally, in order to determine the correct reading, one must consult the cited source.

A formally similar case, yet profoundly different in its lexicographical implications, is the series of instances in which near-duplicate quotations differ from a single reading not because of an OCR error nor because of a transcription error by the lexicographer, but because of an error in the attribution of the cited source. For example, a passage from Giovanni Villani's *Cronica* reading «Le dette sue prediche... erano molto efficaci e d'una buona loquela e di sante parole, dicendole molto dubbiose e accentive a commuovere genti» (s.v. *accentivo*) presents the same quote s.v. *accettevole*, except that the reading «accettevole» appears in place of «accentivo». Both quotes are traced back to the same source, the Moutier edition of the *Cronaca*: here the reading «accentive» is printed in the text, while the reading «accettevoli» is recorded in a note (and appears in the *Vocabolario della Crusca*, which has cited this passage since its first edition, s.v. *accettevole*) (Moutier 1845: 31, note 1)[3].

A different case is represented by identical quotes that in the dictionary are associated with two different identifying strings for the cited source: for example, the quotation «È permesso di fabbricare aceto per diluizione dell'acido acetico puro e di buon gusto purché si venda col nome di aceto artificiale» is cited s.v. *acetico* as an extract from the «Leggi sanitarie, 74-153» and s.v. *aceto* as an extract from the «Leggi industriali e commerciali, 810-153». In this instance, research into the sources would be necessary to determine whether this is a "duplicated quotation" (in which the declared source differs owing to an inconsistency in the dictionary) or a passage that appears identically in two distinct texts. At times, however, consultation of the dictionary alone is sufficient to establish that identical quotes (identified as such by the pipeline) occur in different texts and that we are therefore dealing with a case of intertextual reuse: see, for example, the celebrated incipit of Boccaccio's *Decameron*, «Umana cosa è l'avere compassione agli afflitti», which is cited directly from Boccaccio s.v. *a2* and from Torquato Tasso's reuse s.v. *compassione*. In both cases, as can be seen, the match flagged by the QRD pipeline is not sufficient to determine whether we are truly dealing with replicated quotes: the judgment of a human user is necessarily required.

So far, we have examined cases in which the text strings correspond (punctuation marks notwithstanding, as noted above). The majority of near-duplicates, however, do not present identical text strings: the same passage may be excerpted differently from one entry to another, including additional words at the beginning or end of the quote, or even within it. For example, from the same poem by Vincenzo Cardarelli are drawn the quote «Il Disinganno, / che si nutre di sottigliezze / acerrime e conclusive» (cited s.v. *acerrimo*) and the quote «Il più frettoloso figliolo / del Tempo, il Disinganno, / ...si nutre di sottigliezze / acerrime e conclusive» (cited s.vv. *disinganno* and *frettoloso*). In this case as well, the analysis of the quote strings alone is insufficient: the pipeline detects the quotation similarity, but only the human reader can discriminate between cases of duplicated quotations and those of merely similar passages; cf. «Se mi s'accostava un passo di più, ...l'infilavo addirittura, prima che avesse tempo d'accomodarmi me, il birbone», cited s.v. *accomodare* from Alessandro Manzoni's *Promessi sposi*, matched by the tool with the quote «Se mi s'accostava un passo di più..., l'infilavo addirittura, prima che avesse tempo d'accomodarmi me, il birbone» (s.v. *infilare*),

[3] I. Moutier (ed.), *Cronica di Giovanni Villani*, t. 3, Firenze, Sansone, 1845.

which is likewise drawn from Manzoni but from an earlier redaction of the *Promessi Sposi* (the so-called “Ventisettana”).

Note that the different phenomena we have presented can occur simultaneously: within clusters of duplicated quotations, we may find strings of varying length due to additional elements appearing at any point in the sentence, possible OCR errors, and readings that already diverge in the original printed text, whether drawn from the same source or from different sources.

Overall, as can be seen, the picture of the quotes identified by the pipeline as potentially duplicated quotations is extremely complex, above all because of the characteristics of the original resource: in order to be certain that one is truly dealing with replicated quotes, the judgment of a human expert would almost always be required.

7. Conclusion

The choice between a minimalist solution - where each quotation is considered only once (or a balanced number of times) - and an approach that preserves all occurrences entails both advantages and drawbacks. On the one hand, the noise generated by repeated quotations may be problematic when the corpus of quotations is used for the construction of lexicographic resources. On the other hand, from the perspective of representativeness and quantitative analysis, the recurrence of cited authors and lexical items is far from negligible.

To maximize the flexibility of the resource and broaden its potential applications, it is therefore preferable to keep both options available, allowing users to tailor corpus queries to their specific needs (cf. Biffi *et al*., 2026). In this perspective, the extreme variability observed in the data, ranging from purely technical noise to substantial structural variation in the reuse of textual material, makes the automatic and indiscriminate removal of duplicate quotations both impractical and scientifically undesirable. A “blind” deduplication strategy would entail a systematic risk of eliminating data that are essential to the integrity and interpretability of historical lexicographic resources.

By contrast, the methodology implemented in QRD goes beyond mere dataset cleaning, enabling a deeper and more systematic understanding of the resource itself. In particular, the analysis of similarity-based quotation clusters makes it possible to uncover editorial practices underlying the compilation of the GDLI, shedding light on recurrent patterns of textual reuse, selection, and adaptation over time.

More generally, our findings highlight the intrinsic complexity of quotation duplication in lexicographic data and show that it cannot be addressed through a single, univocal solution. At the same time, QRD effectively supports the identification of OCR-related errors in lexicographic quotations, encompassing both systematic patterns and isolated cases. These errors can subsequently be corrected both in the TEI XML representation of the GDLI and in the derived GDLI Quotations Corpus.

With respect to citation deduplication in corpora derived from lexicographic resources, we contend that it must be conceived primarily as a task of identification and clustering rather than as a straightforward elimination procedure. By explicitly encoding the membership of reused quotations within similarity clusters, our approach enables the adoption of task-specific strategies, ranging from the preservation of all quotation variants for philological and lexicographic analysis to the selection of representative instances for computational processing.

This flexible framework allows multiple requirements to be satisfied simultaneously: corpus quality is improved, editorial variation is preserved and made accessible for scholarly analysis, and - when necessary - linguistically informed deduplication strategies can be applied to support downstream computational tasks, such as the training or fine-tuning of linguistic annotation models.

Finally, the in-depth analysis of repeated quotations has made it possible to characterize the patterns underlying quotation reuse in the GDLI and to inform the design of clustering-based solutions for advanced dictionary consultation and exploration.

Looking ahead, several directions will be pursued to extend this work. First, the GDLI Quotations Corpus will be completed to cover all dictionary volumes, thereby enabling large-scale and fully diachronic analyses. Second, ongoing research is focusing on the systematic evaluation of the impact of duplicated quotations on the training and fine-tuning of linguistic annotation tools. Third, clustering-based methods will be further developed to support advanced functionalities in the structured digital dictionary, including the automatic detection, navigation, and exploration of reused quotations. Together, these directions aim to consolidate QRD as both a methodological framework and an operational tool for the analysis and reuse of historical lexicographic data.

In this perspective, deduplication is no longer a data-cleaning operation, but an analytical methodology for understanding and exploiting historically and editorially curated textual resources.

8. Acknowledgments

This contribution forms part of the ongoing digitization and structuring efforts concerning the GDLI, carried out since 2018 by the “Istituto di

Linguistica Computazionale" of CNR in collaboration with the "Accademia della Crusca". The current implementation of the GDLI-QC constitutes a primary objective of the "GDLIplus" project, funded under the Regione Toscana FSE+ Programme 2021–2027 (Action 4.a.5) and co-financed by the "Accademia della Crusca" (CUP B53C24004090007).

9. Bibliographical References

Atkins, B.T.S., Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford University Press.

Béjoint, H. (2010). *The Lexicography of English. From Origins to Present.* New York, Oxford University Press.

Biffi, M., Favaro, M., Guadagnini, E., Montemagni, S., Sassolini, E. (2026). Gli interventi redazionali negli esempi citati nelle voci del «Grande Dizionario Della Lingua Italiana». *Studi di Lessicografia italiana*, XLIII.

Biffi, M., Guadagnini, E., Montemagni, S., Sassolini, E. (2025). La stampa periodica citata nel GDLI: il rapporto tra voci e indice bibliografico e le prospettive per il dizionario strutturato. *Studi di Lessicografia Italiana*, XLII, pp. 267-94

Biffi, M., Guadagnini, E., Montemagni, S., Sassolini, E. (2023). Il lemmario del «GDLI»: dati quantitativi e prime osservazioni. *Studi di Lessicografia Italiana*, XL, pp. 331-51

Bollmann, M. (2019). A Large-Scale Comparison of Historical Text Normalization Systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long and Short Papers), pp. 3885–3898. Association for Computational Linguistics.

Broder, A. (1997). On the Resemblance and Containment of Documents. In *Proceedings of the International Conference on Compression and Complexity of Sequences*, Salerno, Italy, pp. 21-29.

Christen, P. (2012). *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer Science and Business Media, Berlin.

Elmagarmid, A.K., Ipeirotis, P.G. and Verykios, V.S. (2007). Duplicate Record Detection: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 1, pp. 1-16, Jan. 2007.

Favaro, M., Guadagnini, E., Sassolini, E., Biffi, M., Montemagni, S. (2022). Towards the Creation of a Diachronic Corpus for Italian: A Case Study on the GDLI Quotations. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2022)*, pp. 94–100, Marseille, France. European Language Resources Association (ELRA).

Hoffmann, S. (2004). Using the OED quotations database as a corpus: A linguistic appraisal. *ICAME Journal*, 28, pp. 17-30.

Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C. and Carlini, N. (2022). Deduplicating Training Data Makes Language Models Better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), pp. 8424–8445, Dublin, Ireland. Association for Computational Linguistics.

Levenshtein, V.I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Navarro, G. (2001). A guided tour to approximate string matching. *ACM Computing Surveys (CSUR)*, 33(1):31–88.

Palmerio Aprosio, A., Menini, S., and Tonelli, S. (2022). The "Bertoldo" Corpus: An Annotated Corpus of 17th-Century Italian. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2022)*, pages 78–82, Marseille, France. European Language Resources Association (ELRA).

Piotrowski, M. (2012). *Natural Language Processing for Historical Texts*. Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers, San Raphael, California.

Rohdenburg, G. (2013). Using the OED quotations database as a diachronic corpus. In Krug M. *et al.* (Eds.), *Research Methods in Language Variation and Change*, Cambridge University Press, pp 136-157.

Sassolini, E. *et al.* (in preparation). Historical Dictionaries as Internally Hybrid Resources: The Case of the "Grande Dizionario della Lingua Italiana".

Sassolini, E., Biffi, M., De Blasi, F., Guadagnini, E. e Montemagni, S. (2021). La digitalizzazione del GDLI: un approccio linguistico per la corretta acquisizione del testo? In *Proceedings di AIUCD 2021: DH per la società: e-guaglianza, partecipazione, diritti e valori nell'era digitale*. Raccolta degli abstract estesi della 10a conferenza nazionale, pp. 159-166.

Sassolini, E., Biffi, M. (2020). Strategie e metodi per il recupero di dizionari storici. In *Proceedings of the AIUCD 2020 Conference*, 15-17 gennaio

2020, Università Cattolica del Sacro Cuore. Milano, pp. 235-239.

Sassolini, E., Khan, A. F., Biffi, M., Monachini, M., Montemagni, S. (2019). Converting and structuring a Digital Historical Dictionary of Italian: a case study. Electronic lexicography in the 21st century. In *Proceedings of the eLex 2019 conference*, 1-3 October 2019, Sintra, Portugal. Brno: Lexical Computing CZ, s.r.o. Eds., pp. 603-621.

JABERT: A New State-of-the-Art BERT Model for Judeo-Arabic

Elisha Rosensweig,[1,2,‡] **Yitzchak Lindenbaum,**[1,2,‡] **Hillel Gershuni,**[1,2,‡]
Vered Raziel-Kretzmer,[2,§] **Daniel Caine,**[2,◇] **Avi Shmidman**[1,2,†]
[1]DICTA / Jerusalem, Israel [2]Bar-Ilan University / Ramat Gan, Israel
[†]avi.shmidman@biu.ac.il, [§]razielv@post.bgu.ac.il
[◇]daniel.caine@mail.huji.ac.il
[‡]{benshafat, gershuni, yitzilindenbaum}@gmail.com

Abstract

We present JABERT, the first BERT model pretrained specifically for historical Judeo-Arabic texts. We demonstrate that JABERT outperforms Arabic and multilingual models on the downstream task of Judeo-Arabic homograph disambiguation. Furthermore, in order to test the latter, we have curated and annotated the first Judeo-Arabic homograph test set. We release both JABERT and the Judeo-Arabic homograph test to the public for unrestricted use.

Keywords: Judeo-Arabic, BERT, WSD

1. Introduction

Judeo-Arabic (=JA) refers to a diverse family of Arabic dialects historically spoken and written by Jewish communities throughout the Middle East and North Africa, from the 8th century and onward through the middle ages. Judeo-Arabic is written in Hebrew script, incorporating Hebrew and Aramaic lexical elements while maintaining Arabic grammatical structures and core vocabulary. As a primary vehicle for Jewish intellectual, religious and everyday communication in Arabic-speaking lands, Judeo-Arabic has left behind a vast textual legacy spanning philosophy, biblical commentary, legal responsibility, scientific treatises, personal correspondence, and literary works. Notable examples include the writings of Saadia Gaon, Judah Halevi and Maimonides, whose influence extended far beyond their immediate communities.

Despite its historical and cultural significance, Judeo-Arabic remains severely underrepresented in NLP research. The language presents unique challenges as a low-resource variety: digital corpora are limited in size, texts exhibit substantial orthographic and dialectal variation across regions and time periods, and the Hebrew-script encoding creates additional barriers for standard Arabic NLP tools. While recent years have seen rapid advances in transformer-based language models for both Modern Standard Arabic and various Arabic dialects, as well as for historical Hebrew, no dedicated BERT model for Judeo-Arabic has been developed to date. This gap has hindered computational research on Judeo-Arabic texts and limited the application of modern NLP techniques to digital humanities scholarship in this domain.

In this paper, we present JABERT, the first pre-trained BERT model specifically designed for Judeo-Arabic. We test JABERT's ability to predict missing words within actual Judeo-Arabic historical documents, and we present both quantitative and qualitative evaluations. Furthermore, we compare the performance of JABERT to that of Arabic and multilingual BERT models when put to the downstream task of Judeo-Arabic homograph disambiguation. We demonstrate that JABERT outperforms these configurations.

In what follows we describe our training methodology, including corpus collection and preprocessing strategies for handling the unique characteristics of Judeo-Arabic texts, and we present an evaluation of the model's performance. We release JABERT on huggingface for unrestricted use.[1]

2. Related Work

As noted above, there does not currently exist a BERT model specifically trained for Judeo-Arabic. However, because Judeo-Arabic is effectively a mixture of Hebrew and Arabic, we survey the existing BERT models for Hebrew and for Arabic, as well as existing systems for transliterating between Hebrew and Arabic scripts.

2.1. BERT Models

2.1.1. Arabic Models

Multiple Arabic BERT models exist (e.g., AraBERT (Antoun et al.), CAMeLBERT (Inoue et al., 2021), MARBERT (Abdul-Mageed et al., 2021)). However, these are trained exclusively on Arabic script, while Judeo-Arabic uses Hebrew-script, thus preventing them from being directly applicable to Judeo-Arabic processing. Nevertheless, they can be applied via transliteration; see below, 2.2.

[1]Huggingface link: https://huggingface.co/MiDRASH-ERC/JABERT . The exact license can be found at https://creativecommons.org/licenses/by/4.0/

Proceedings of LT4HALA 2026 @ LREC 2026, pages 58–71.
11 May 2026. *©ELRA Language Resources Association (ELRA), 2026*

2.1.2. HeArBERT: a Hebrew-Arabic Model

HeArBERT (Rom and Bar, 2024) is a bilingual Hebrew-Arabic BERT that uses transliteration to map Arabic text into Hebrew script, creating a shared character space for both Semitic languages. It was not trained on Judeo-Arabic texts, but since it operated in Hebrew script, it is suitable to use as a reference point to the performance of our Judeo-Arabic model JABERT. In Section 5.3 we evaluate the performance of JABERT vis-a-vis HeArBERT.

2.2. Transliteration Methods

With a lack of models directly applicable to Judeo-Arabic, an alternative approach is to transform the Judeo-Arabic text into Arabic script, and then apply Arabic BERT models to the transliterated text. In Section 5.3 we evaluate the performance of JABERT vis-a-vis this transliteration approach.

Transliterating Judeo-Arabic was initially addressed by Bar et al. (Bar et al., 2015), which introduced statistical machine translation for character-level Judeo-Arabic to Arabic transliteration, achieving 96.9% precision. This was extended by Terner et al. (Terner et al., 2020) using RNN with CTC loss, reducing character error rate from 9.5% to 2% through synthetic data augmentation.

In a more recent work, Gonzalez et al. (2025) used a transliteration-based approach to process Judeo-Arabic. Their transliteration was comprised of two steps. The first step involved a rule-based transliteration from Hebrew to Arabic script. However, because there are cases where a given Judeo-Arabic letter stands in more than one Arabic letter, a rule-based transliteration may induce a certain margin of error (see ibid., Section 2.1). Due to this factor, they applied a second stage in which the text was auto-corrected either by commercial LLMs (e.g. GPT4o) or using dedicated GEC (grammatical error correction) models (specifically, the Seq2Seq SWEET models proposed by Alhafni et al. (2023)). The resulting auto-corrected text was then processed in Arabic BERT models. They showed that transliteration combined with post-correction enables using existing Arabic NLP tools for downstream tasks such as machine translation, thus establishing a pathway to handle Judeo-Arabic processing in the absence of a dedicated Judeo-Arabic model.

The transliteration method has three notable drawbacks. First, as noted, transliterating Judeo-Arabic into Arabic script may produce an incorrect result due to the cases where a single Judeo-Arabic letter represents more than one Arabic letter; the aforementioned autocorrection methods do not always yield a perfect result. Second, the current SOTA method from Gonzalez et al. (2025) utilizes decoder methods which do not always produce a direct word-for-word conversion, and which may instead produce a paraphrase of the original text; this poses a major difficulty for word-based labeling tasks. Finally, the Arabic models can only correctly process the Arabic portions of the transliterated text. Yet, in practice, Judeo-Arabic texts integrate many Hebrew loan-words, Hebrew citations, Hebrew Biblical verses, and more. Due to this issue, performance in Gonzalez et al. (2025) is measured by applying a code-switching detector and isolating the analysis to the Arabic portions alone. JABERT is designed to handle an entire Judeo-Arabic text as-is, thus addressing all three drawbacks.

3. Model Training

3.1. Tokenizer

We use the Word-Piece tokenization method proposed by Song et al. (2021) with the default normalizers and preprocessors suggested by HuggingFace. Following the work of Gueta et al. (2023) demonstrating the substantial advantage of larger token vocabularies for morphologically-rich languages, the tokenizer was trained with a vocabulary size of 128K tokens.

Due to the fragmentary nature of many historical Judeo-Arabic texts, we introduce two special tokens into the vocabulary to handle textual lacunae:

- "[ONEGAP]" indicates a single word with at least one missing letter. If the transcription indicated several consecutive missing characters, up to 5 characters were assumed to constitute a single word.
- "[GAP]" indicates multiple missing words or a lacuna of indeterminate length.

3.2. Training Data

The Judeo-Arabic texts included in JABERT's training corpus were amassed from three different sources: the Princeton Genizah Project (PGP; Rustow and Koeser, 2025), the Friedberg Judeo-Arabic Project by the Friedberg Jewish Manuscript Society[2] (FJMS), and quotations in Blau's *Dictionary of Mediaeval Judaeo-Arabic Texts* (Blau, 2006). Since these sources would not have been sufficient to train a competent BERT model, we augmented them with medieval Arabic texts from the Open Islamicate Texts Initiative (OpenITI; Romanov et al., 2024) transliterated into Hebrew characters, approximating the way these texts would appear if they had been written in Judeo-Arabic. We included texts by authors whose listed *hijri* year of

[2]Resource can be accessed at https://github.com/princetongenizalab/pgp_transliteration/tree/main/resources/scrapes.

death is between 390 and 905 (corresponding to 1000–1500 C.E.)[3] and transliterated them using a simplified and slightly adjusted version of the table provided in Weisberg Mitelman et al. (2024). We employed weighted stochastic decision based on typical Judeo-Arabic usage to simulate regional and scribal variation.[4]

The texts from each source were preprocessed to handle punctuation, glosses/translations, and document segmentation as appropriate. The full sets of preprocessing steps applied to each corpus are detailed in Appendix B. To the extent possible, only the original medieval text was preserved, although where lacunae were filled in by editors, the suggestions were accepted. Remaining lacunae were handled using the `[ONEGAP]` and `[GAP]` tokens (see section 3.1).

Source	Word Count	Notes
PGP	~1.6 million	Transcribed manuscripts
FJMS	~3.1 million	Printed works
Blau	~42,000	Extracted quotes
OpenITI	~1.2 billion	Transliterated

Table 1: Training data sources

3.3. Training Details and Hyperparameters

We trained our model with the HuggingFace architecture wrapped with NVIDIA libraries,[5] which are highly optimized for training compute-heavy machine learning models on NVIDIA hardware. We pre-trained the model on four NVIDIA A40 48GB GPUs for a total of 86,010 iterations, with a total of 106.56B tokens.

The training was carried out in 3 phases:

(1) Sequences of 128 tokens with a learning rate of 6e-3 for 13,000 iterations.

(2) Sequences of 128 tokens with a learning rate of 1e-4 for 57,380 iterations.

(3) Sequences of 512 tokens with a learning rate of 5e-5 for 15,630 iterations.

The training was done with a batch size of 8,192 in phases 1 & 2 and 4,096 in phase 3. Warmup proportion of 0.2843 for phases 1 & 2 (jointly, as phase 2 is a continuation of phase 1), and 0.128 for phase 3.

4. Datasets

4.1. JABERT Cairo Genizah Dataset

In order to evaluate JABERT's ability to predict masked words within actual Judeo-Arabic historical documents, we curated a new dataset of lines of Judeo-Arabic text not included in JABERT's training data. In order to do so, we employed a Judeo-Arabic expert who transcribed lines of Judeo-Arabic from Cairo Genizah manuscripts stored in the Cambridge University Library.[6] In total, our expert transcribed 3862 lines of text. For the transcription protocol according to which the Judeo-Arabic expert transcribed these lines, see Appendix C.

4.2. JABERT Homograph Disambiguation Challenge Dataset

We also sought to test JABERT's ability to distinguish between different meanings of Judeo-Arabic words. To this end we curated a challenge dataset consisting of 16 Judeo-Arabic homographs. Two Judeo-Arabic experts examined dozens of occurrences of each of these homographs within Judeo-Arabic texts, and they annotated each of these occurrences for its context-specific meaning.[7] To the best of our knowledge, this is the first instance of a Judeo-Arabic homograph challenge set. Table 2 provides a quantitative overview of this dataset.

Category	# instances	median / word
words	16	1
meanings	37	2
samples	1119	71

Table 2: Homograph Dataset at a glance

Table 3 illustrates four representative homographs from the dataset, selected to highlight various disambiguation challenges: functional (e.g., distinguishing particles from other parts of speech), lexical and syntactic (e.g., noun vs. adjective or different verb forms), and root-level ambiguity in Judeo-Arabic orthography.[8]

[3] Documents tagged as "UNCORRECTED_OCR" were not used.

[4] Specifically, 80% of the documents (chosen randomly) were transliterated with jīm (ج) rendered as ג (without geresh) and ghayn (غ) as ג'. In the remaining 20%, jīm → ג' and ghayn → ג. Hamza (ء on the line) was randomly rendered as either א or omitted entirely. The complete transliteration table is presented in Appendix A.

[5] https://github.com/NVIDIA/DeepLearningExamples/tree/master/PyTorch/LanguageModeling/BERT

[6] The documents were selected from the following four boxes of the Taylor-Schechter collection of Cairo Genizah fragments in the Cambridge University Library: T-S Misc. 24; T-S Arab. 31; T-S Arab. 43; and T-S 8F.

[7] Where there was a disagreement between the two experts, we discarded the sample.

[8] The full list of Judeo-Arabic homographs, with grammatical explication and translation, is presented in Appendix D.

#	Translit.	Arabic	Meaning
1. אלא (*'l'*)			
1	*alā*	أَلَا	“Lo”, “is it not?”
2	*illā*	إِلَّا	“except”, “other than”
12. עאלם (*'lm*)			
28	*'ālam*	عَالَم	“world”
29	*'ālim*	عَالِم	“knows”
13. כלק (*ǩlq*)			
30	*ǩalaqa*	خَلَقَ	“created”
31	*ǩalq*	خَلق	“creation”
16. אכד (*'kd*)			
36	*akkada*	أَكَّدَ	“confirmed”
37	*aǩaḏa*	أَخَذَ	“took”

Table 3: Sample homographs illustrating common ambiguities: functional (1), semantic and POS (12), and morphological or root-based (13, 16).

4.3. Release of our datasets to the public

We are pleased to release these two datasets to the public, both in the interests of reproducibility, and also so that these datasets can continue to be used as benchmarks to test and evaluate future Judeo-Arabic NLP models.[9] The datasets are released under a permissive Creative Commons CC-BY license.

5. JABERT Evaluation

To evaluate the utility of the model and the degree to which it captures the properties of Judeo-Arabic, we utilize two tasks:

- **Masked Word Prediction** - Given a Judeo-Arabic line from the JABERT Cairo Genizah Dataset (Section 4.1), we mask a word in the line and ask the model to predict the missing word. This task sets the bar quite high, as multiple words might fit reasonably in place of the mask (e.g., swapping "Sunday" for "Monday"). The task is even more challenging in our case since each line in the dataset consists of a single *physical* line from actual historical Judeo-Arabic manuscripts, wherein the lines generally begin and end mid-sentence. This reflects actual expected scholarly usage of the BERT model when working on historical manuscripts, and substantially increases the difficulty of the task. We evaluate the ability of JABERT to predict the correct word within the top-k candidates it proposes, for various values of k. Because there are no existing BERT models capable of predicting full Judeo-Arabic words as MLM predictions, this task is not a comparative one. Below we provide quantitative results followed by a qualitative analysis.
- **Homograph Disambiguation** - We evaluate the model's ability to distinguish between different senses of a single Judeo-Arabic word form. Success at this task would indicate that the model captured key semantic elements of Judeo-Arabic. For this evaluation task, we use the JABERT Homograph Disambiguation Challenge Set (Section 4.2). Furthermore, because this test relies on the contextual embedding rather than on the MLM prediction, it can be used to compare JABERT's abilities to other multilingual BERT models. Thus, for this experiment, we compare JABERT's performance with that of the industry-standard multilingual mBERT, side-by-side with the aforementioned HeArBERT model and the transliteration-based method proposed by Gonzalez et al. (2025). Here too, we provide quantitative results followed by a qualitative analysis.

Crucially, the test sets for these two evaluations consist entirely of authentic Judeo-Arabic text. This is important because, as described earlier (3.2), in order to amass a corpus large enough for pretraining JABERT, we relied heavily on transliterated medieval Arabic texts, and the reader may legitimately wonder whether the model truly captures Judeo-Arabic, or rather just Arabic transliterated into Hebrew characters. In order to address this concern, we ensured that the test sets contain only authentic Judeo-Arabic text. Thus, the results obtained on our evaluation tasks reflect the model's ability to handle actual Judeo-Arabic.

5.1. Masked Word Prediction - Quantitative Evaluation

For each single-token word[10] in each line of the JABERT Cairo Genizah Dataset, we mask the given word and run the line through JABERT's MLM head, retrieving predictions for the masked word position. In total, this yielded $16,129$ masked words on which we ran MLM inference and evaluated the model's predictive performance.

When predicting the word, we considered how far into the ranked candidates list we needed to proceed in order to find a prediction of the correct

[9] Link to download: https://github.com/ERC-Midrash/ja-models-eval

[10] Other than the first and last words on the line; masking either of these words would be an unfair challenge because BERT would not have access to the bi-directional context on the other side of the word.

word, for $k \in [1, 250]$. Figure 1 shows the percentage of samples where the model predicted the missing word correctly within the top k words as k increases. Table 4 provides some representative performance milestones for clarity. As can be seen here, despite the high difficulty level of the task, the model rapidly increases in performance until approximately $k = 25$, where it reaches 42% accuracy. From that value on we get steady but diminishing returns as k increases.

k	Percentage	k	Percentage
1	16.36%	25	42.01%
2	21.40%	50	48.38%
3	24.47%	100	54.89%
5	28.55%	150	58.48%
10	34.01%	250	63.31%

Table 4: Results for various values of k

5.2. Masked Word Prediction - Qualitative Evaluation

The quantitative analysis poses a high bar for success: to accurately predict the correct token out of 128K tokens. This is complicated further by the fact that certain tokens cannot be predicted in this fashion - for example, a date or number might be interchangeable in a given position with other, equally valid dates and numbers. As such, we turn to qualitative analysis of some examples, to demonstrate meaningful patterns that are not captured by the preceding quantitative analysis.

Consider the following examples where the correct word was not ranked first:

- **Ground Truth:** פיה (*fīhi*, "in it")
  Context: יכון אלאקראר ממא יגב עליה פיה חכם (*yakūn al-iqrār mimmā yajib ʿalayhi **fīhi** ḥukm*, "the acknowledgement is among the things for which a ruling is required **therein**")
  Model Rank: 8
  Top Candidates: לה (*lahu*, "to it"), ענדה (*ʿindahu*, "at/with it"), מעה (*maʿahu*, "with it"), פי (*fī*, "in"), עליה (*ʿalayhi*, "upon it")
  Analysis: All predictions belong to the correct semantic category of relational prepositions, demonstrating the model's ability to identify the appropriate syntactic slot despite the specific token variation.

- **Ground Truth:** מר (*mar*, honorific title)
  Context: ואלהלכה פי דלך כקול מר שמואל (*wa-l-halākha fī dhālik ka-qawl **mar** Shemūʾel*, "And the law in this matter follows **Mar** Samuel")
  Model Rank: 8
  Top Candidates: רבינו (*rabbēnū*), רבנו (*rabbānū*, both "our rabbi"), ר' (*rabbī*, "rabbi"), אלחבר (*al-ḥaver*, "the scholar"), רב (*rav*, "rabbi/master")
  Analysis: The model correctly identifies this as a slot for an honorific or rabbinic title.

- **Ground Truth:** תצח (*taṣiḥḥu*, "is valid", fem.)
  Context: הו אן אלדי ימלך בה אלאשיא אלתי תצח מלכהא בה (*huwa anna alladhī yamliku bihi al-ashyāʾ allatī **taṣiḥḥu** milkuhā bihi*, "it is that by which one possesses the things whose ownership **is valid** through it")
  Model Rank: 43
  Top Candidates: יצח (*yaṣiḥḥu*, "is valid"), יקע (*yaqaʿu*, "takes effect"), ימכן (*yumkinu*, "is possible"), יגוז (*yajūzu*, "is permissible"), יגב (*yajibu*, "is required"), all masc. forms.
  Analysis: While the manuscript uses a feminine verb, likely influenced by the preceding non-human plural *al-ashyāʾ* and the feminine pronominal suffix in *milkuhā*, the model correctly identifies the masculine agreement required by the subject *milk*. This demonstrates JABERT's grasp of normative syntax despite scribal variation.

These patterns hold broadly: predictions tend to cluster within the same part of speech and semantic field as the target word. In several instances, the model's "errors" actually reflect normative syntactic rules, such as correct gender agreement, even when the original manuscript exhibits scribal variation. This suggests that the top-1 accuracy underestimates the model's grasp of Judeo-Arabic. It captures the distributional structure of the language even when it does not recover the exact lexical item. The jump from 16.35% (top-1) to 34.01% (top-10) reflects not just additional guesses, but the model's ability to identify the correct semantic and syntactic category.

5.3. Homograph Disambiguation - Quantitative Evaluation

Next, we challenge the model to disambiguate Judeo-Arabic homographs. To succeed at this task, JABERT would need to have captured properties of Judeo-Arabic that allow it to determine the correct meaning of a word given the context.

For each homograph, we perform the following steps[11]:

- Take each sentence with the homograph, remove punctuation and mask the homograph.
- Run the sentence through the BERT model and compute the final embedding of the masked token.

[11] See our Homograph eval code here: https://github.com/ERC-Midrash/ja-models-eval

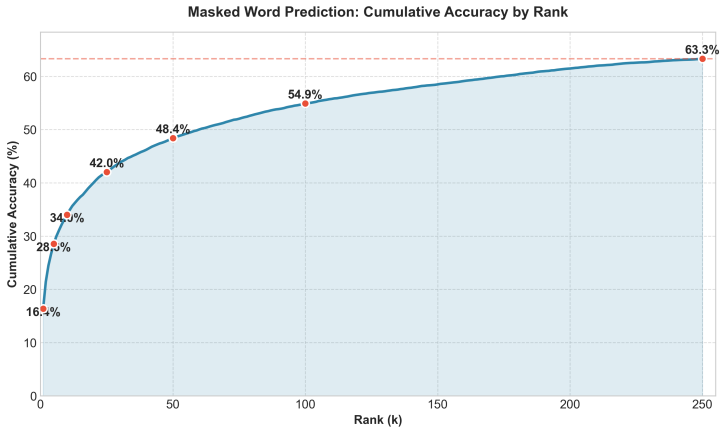


Figure 1: Cumulative accuracy for masked word prediction by rank k

- Run clustering algorithms on the computed embeddings. We attempt here both supervised and unsupervised clustering.
- Compute metrics measuring the quality of the clustering vis-a-vis the ground truth labels (= actual meaning of the homograph in each sentence).

Note that we compute the embedding on the *masked* homograph. This method solves several problems. First, it ensures the embedding is based only on the context, not the word itself, making the task both more challenging and increasing the applicability and stability of this approach across other words. Second, it prevents us from having to handle word-pieced words differently. Most importantly, it allows us to compare performance across models while remaining agnostic to their different vocabularies (i.e., tokenization). It also allowed us to compare our performance to that of the transliteration method discussed in Section 2.2.

We conducted two types of clustering on the embeddings: supervised (kNN-LOOCV, with k=5 neighbors) and unsupervised (k-means, where k is chosen to match the known number of meanings per homograph). For kNN-LOOCV the success metric is simply the number of correctly-assigned samples. For the unsupervised clustering task we considered two performance metrics to evaluate the quality of the proposed clustering *HYP* compared to the Ground Truth labels *GT*:

- **Accuracy (percentile)** - fraction of correctly-assigned words. Computing this requires mapping the *HYP* cluster labels to *GT* cluster labels, such that the accuracy measure is maximized.
- **Adjusted RAND Index** - Given *HYP*, for each pair of data-points (= embeddings) u, v, we check if they have the same relative status (same cluster / different cluster) in *HYP* as in *GT*. If they do, we call this *pairwise agreement*. The RAND index is then *(# pairwise-agreements) / (# total pairs)*. The *adjusted* RAND Index normalizes this value to reflect the degree to which the computed score deviates from the score we would get from a *random* clustering.

We compare the performance of our model to several competing models and flows:

- `mbert` - multilingual BERT.[12]
- `HeArBert` (above, 2.1.2).
- The transliteration approach by Gonzalez et al. (2025), as outlined in Section 2.2. Specifically, we transliterated the Judeo-Arabic homograph dataset into Arabic based on Fig. 1 in Appendix A in their paper, followed by running their top-performing "SWEET" GEC model on the transliterated output. Next, we applied the clustering process, using embeddings for the contextualized masked Arabic word from `CAMeLBERT-MSA` and `CAMeLBERT-CA`, models trained on Modern Standard Arabic (MSA) and Classic Arabic (CA), respectively[13]

For the transliteration process, Gonzales et. al. report that at times it causes structural and stylistic changes to the text, and we observed this phenomenon as well. Our initial homograph dataset consisted of 1118 entries (=sentences with homographs), but post-transliteration + GEC model the homograph word was recoverable perfectly only for a subset of the entries - 859 entries in total. In order to run clustering algorithms we require a minimal amount of datapoints, so we ended up here with 853 entries and 14 homographs to analyze.

[12] https://huggingface.co/google-bert/bert-base-multilingual-cased

[13] See our code here: https://github.com/ERC-Midrash/ja-models-eval.

Due to all of the above, when comparing performance, we ran both (a) full-dataset comparison without the transliteration flow (Fig. 2), and (b) limited dataset comparison, including transliteration flow (Fig. 3).

As can be seen in both Fig. 2 and 3, JABERT exceeds performance of all alternative methods in k-means, with the highest median accuracy and Adjusted RAND scores and competitive in all other quartiles, and in some cases reaching close-to-perfect alignment with Ground Truth. Also, in Fig. 4 we see similar improvements in JABERT over the alternatives when applying the supervised kNN clustering. Of all the alternative methods, the best ones were those based on transliteration followed by strong Arabic models, following Gonzales et. al., for the reduced dataset as discussed above.

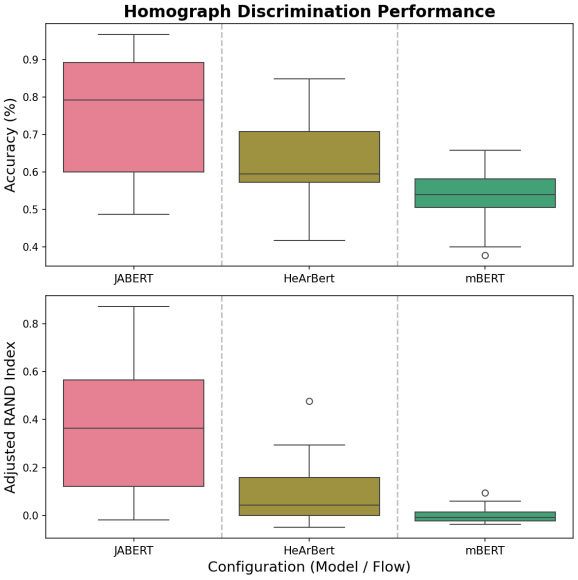


Figure 2: Performance for *k-means* clustering. Accuracy (%) and adjusted RAND scores for JABERT compared to two alternate models aggregated across the entire homograph dataset.

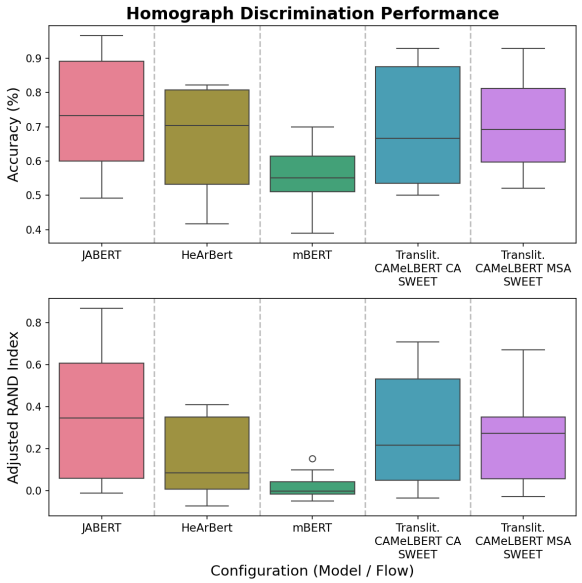


Figure 3: Performance for *k-means* clustering. Accuracy (%) and adjusted RAND scores for JABERT compared to two alternate models and the transliterate + GEC + arabic BERTs flow, aggregated across the subset of the homograph dataset where the transliteration + GEC process was stable.

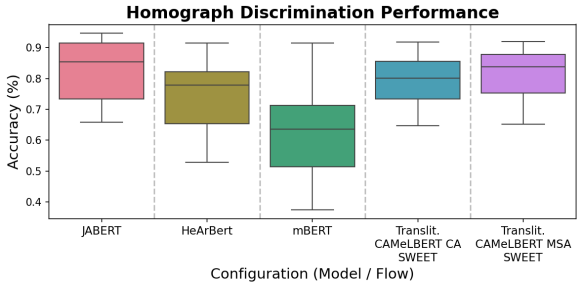


Figure 4: Performance for *kNN-LOOCV* clustering. Accuracy (%) scores for JABERT compared to two alternate models and the transliterate+arabic models flow, aggregated across the subset of the homograph dataset where the transliteration + GEC process was stable.

5.4. Homograph Disambiguation - Qualitative

We examined several cases where clustering results diverged from our annotations. Rather than straightforward errors, many of these cases point to linguistic subtleties.

Formulaic expressions. The homograph אלא (*'l'*) has two meanings: the interrogative *alā* ("Lo", "is it not?", meaning 1) and the particle *illā* ("except", meaning 2). Notably, meaning 1 appears frequently in the fixed phrase אלא תרי (*alā tarā*, "do you not see?"). As shown in Fig. 5, the model clusters these formulaic instances tightly, separating them from the few non-formulaic interrogative uses–a distinction our annotation scheme did not make.

Definiteness patterns. For the homograph עאלם (*''lm*)–representing *'ālam* ("world", meaning 28) vs. *'ālim* ("knowing", meaning 29)–one instance (701: קדר עאלם, *qadr 'ālam*, "the size of a world") clustered among the meaning 29 instances (in close proximity to instance 735, see Fig. 6). The reason is probably distributional: *'ālam* typically appears in the construct state (e.g., עאלם אלאפלאך, *'ālam al-*

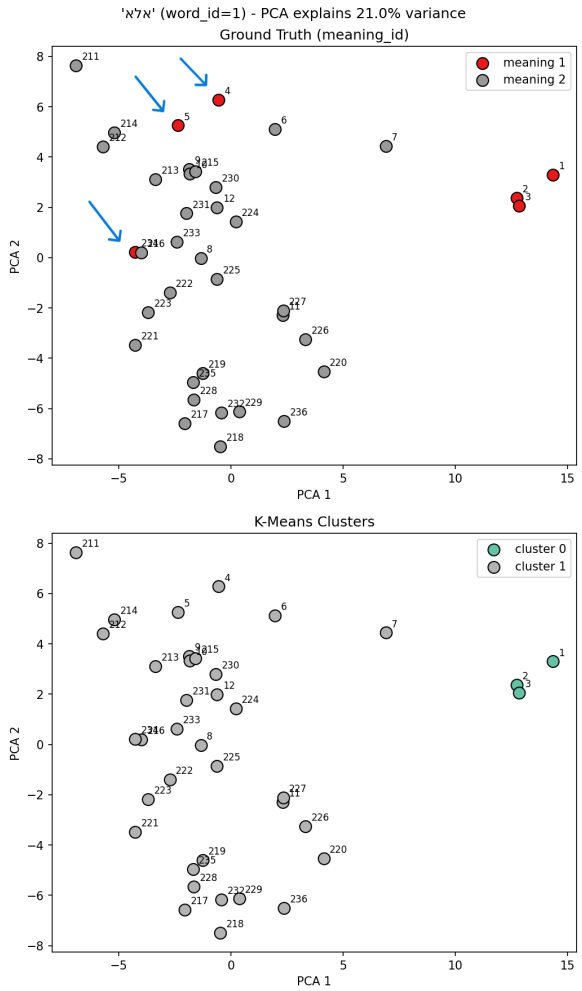


Figure 5: PCA visualization of JABERT embeddings for the homograph **אלא**. Top: Ground truth annotations for meaning 1 (*alā*, interrogative) and meaning 2 (*illā*, exceptive). Bottom: Unsupervised *k-means* clusters ($k = 2$). Note that the formulaic instances of *alā tarā* (points 1, 2, 3) are clustered together, while non-formulaic interrogative instances (e.g., 4, 5, 10, indicated by arrows) are closer to meaning 2. This suggests the model is highly sensitive to formulaic patterns that transcend our binary semantic categories.

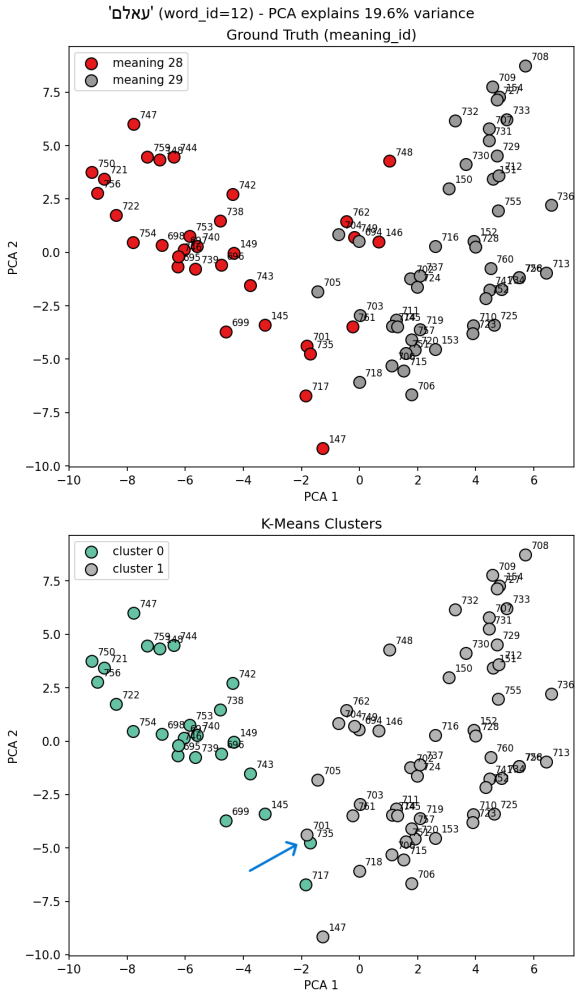


Figure 6: PCA visualization of JABERT embeddings for **עאלם**. Top: Ground Truth for meaning 28 (*ʿālam*, world) and 29 (*ʿālim*, knows). Bottom: k-means clusters ($k = 2$). The model separates senses based on syntactic distribution (definiteness and construct state); instance 701 (indicated by arrow) is an outlier due to its rare indefinite, non-annexed form.

aflāk, "the world of the [celestial] spheres") or with the definite article **אל** (*al-ʿālam*, "the world"). The indefinite, non-annexed form in 701 is rare in this meaning and much more common in meaning 29, and the model's embeddings reflect this.

Voice distinctions and annotation choices. The homograph **כלק** (*ḥlq*) was annotated as a verb (*ḥalaqa*/*ḥuliqa*, active/passive "created", meaning 30) vs. a noun (*ḥalq*, "creation/creatures", meaning 31). Instance 551 (**לדלך כלק**, *li-ḏālik ḥuliqa*, "for this purpose he was created") clusters with meaning 31, probably because its passive voice was marked as unique. Compare this to **יערף** (*yʿrf*), where we annotated active (*yaʿrifu*), passive (*yuʿrafu*), and causative (*yuʿarrif*) as three separate meanings (16–18). For *hlq*, we grouped active and passive together. The number of clusters defined ahead has thus a crucial impact on the clustering outcomes.

Orthographic identity vs. syntactic role. The homograph **פהם** (*fhm*) can be the noun *fahm* ("understanding", meaning 34) or the sequence *fa-hum* ("and they", meaning 35). Most instances cluster correctly. However, instance 872 is an outlier: the text has **פהם** where standard orthography would

have פאהם (*fāhim*, act. part., "a person who understands"). This spelling variation led to misclassification.

Collocational pull. The homograph אכד (*'kd/'ḫḍ*, when the diacritical mark is often omitted in manuscripts) represents *'akkada* ("confirmed/affirmed", meaning 36) vs. *'akhaḍa* ("took", meaning 37). instance 186 (אכד פי אלקסם, *'akkada fī al-qasam*, "affirmed by an oath") was correctly annotated as "confirmed/affirmed" but clustered with "took." The likely cause: the strong collocation between "oath" and "take" in Arabic (*'akhaḍa al-qasam*, "took an oath"). This lexical association apparently outweighed the contextual evidence for the Form II (*'akkada*) reading.

6. Future Directions

As noted, our evaluation datasets are comprised entirely of authentic Judeo-Arabic text, and thus they effectively demonstrate the ability of JABERT to handle actual cases of Judeo-Arabic, rather than the transliterated Arabic which formed much of the pretraining corpus. Nevertheless, it remains to be explored whether JABERT's capabilities derive from the quality of the genuine Judeo-Arabic training signal, the quality of the transliterated Arabic signal, or a combination of both. In future work, we plan to train a separate version of JABERT based entirely on the transliterated part of the corpus, and a further version based entirely on the authentic Judeo-Arabic part of the corpus, in order to address the question of the relative contribution and necessity of each part of JABERT's pretraining corpus.

7. Conclusion

In this paper, we introduced JABERT, the first pre-trained BERT model dedicated specifically to Judeo-Arabic. By leveraging a massive corpus of transliterated medieval Arabic alongside curated historical Judeo-Arabic manuscripts, we have addressed a significant gap in the NLP landscape for low-resource, historical Jewish languages. Our model not only sets a new SOTA for Judeo-Arabic NLP tasks, but also obviates the need for the cumbersome and brittle transliteration process which the previous SOTA methods required.

We hope that JABERT will lower barriers to computational analysis of Judeo-Arabic materials, and empower researchers to unlock the vast intellectual heritage contained within the corpus of Judeo-Arabic literature.

8. Acknowledgments

This work has been funded by the European Union (ERC, MiDRASH, Project No. 101071829; principal investigators: Avi Shmidman, Bar-Ilan University; Daniel Stökl, EPHE-PSL; Nachum Dershowitz, Tel Aviv University; and Judith Olszowy-Schlanger, EPHE-PSL), for which we are grateful. Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

9. Bibliographical References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Bashar Alhafni, Go Inoue, Christian Khairallah, and Nizar Habash. 2023. Advancements in Arabic grammatical error detection and correction: An empirical investigation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6430–6448, Singapore. Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

Kfir Bar, Nachum Dershowitz, Lior Wolf, Yackov Lubarsky, and Yaacov Choueka. 2015. Processing judeo-arabic texts. In *2015 First International Conference on Arabic Computational Linguistics (ACLing)*, pages 138–144.

Joshua Blau. 2006. *A Dictionary of Mediaeval Judaeo-Arabic Texts*. The Academy of the Hebrew Language; The Israel Academy of Sciences and Humanities.

Juan Moreno Gonzalez, Bashar Alhafni, and Nizar Habash. 2025. A tale of two scripts: Transliteration and post-correction for judeo-arabic.

Eylon Gueta, Avi Shmidman, Shaltiel Shmidman, Cheyn Shmuel Shmidman, Joshua Guedalia, Moshe Koppel, Dan Bareket, Amit Seker, and Reut Tsarfaty. 2023. Large pre-trained models with extra-large vocabularies: A contrastive analysis of Hebrew BERT models and a new one to outperform them all.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Kyiv, Ukraine (Online). Association for Computational Linguistics.

Princeton Geniza Project. 2021. Transcription history and conventions. Google Docs. Archived at https://web.archive.org/web/20260204094447/https://docs.google.com/document/d/1pyFajrGRYTzRsb2yFjfX0XmNqpz1kUO9mhlISGwfpxk/edit?tab=t.0. Accessed March 25, 2025.

Aviad Rom and Kfir Bar. 2024. Training a bilingual language model by mapping tokens onto a shared character space. *arXiv preprint arXiv:2402.16065*.

Maxim Romanov. 2023. OpenITI mARkdown. Accessed April 7, 2025.

Xinying Song, Alex Salcianu, Yang Song, Dave Dopson, and Denny Zhou. 2021. Fast wordpiece tokenization.

Ori Terner, Kfir Bar, and Nachum Dershowitz. 2020. Transliteration of Judeo-Arabic texts into Arabic script using recurrent neural networks. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 85–96, Barcelona, Spain (Online). Association for Computational Linguistics.

Daniel Weisberg Mitelman, Nachum Dershowitz, and Kfir Bar. 2024. Code-switching and back-transliteration using a bilingual model. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1501–1511, St. Julian's, Malta. Association for Computational Linguistics.

10. Language Resource References

Romanov, Maxim G. and Miller, Matthew Thomas and Savant, Sarah Bowen and Verkinderen, Peter. 2024. *OpenITI: a Machine-Readable Corpus of Islamicate Texts*. Zenodo, 2023.1.8. Accessed September 25, 2024.

Rustow, Marina and Koeser, Rebecca Sutton. 2025. *Princeton Geniza Project*. Center for Digital Humanities at Princeton, 4.28.0. Accessed March 25, 2025. Text data retrieved via https://github.com/princetongenizalab/pgp-text.

A. Appendix: Arabic to Hebrew Transliteration Table

As explained above in Section 3.2, this is a simplified and slightly adjusted version of the transliteration table provided in Weisberg Mitelman et al. (2024)

Arabic	Name	Hebrew	Notes
ا	alif	א	No change for hamza or madda
ب	bā’	ב	
ت	tā’	ת	
ث	thā’	ת’	
ج	jīm	ג’ / ג	See note 1
ح	ḥā’	ח	
خ	khā’	כ’	
د	dāl	ד	
ذ	dhāl	ד’	
ر	rā’	ר	
ز	zāy	ז	
س	sīn	ס	
ش	shīn	ש	
ص	ṣād	צ	
ض	ḍād	צ’	
ط	ṭā’	ט	
ظ	ẓā’	ט’	
ع	‘ayn	ע	
غ	ghayn	ג’ / ג	See note 1
ف	fā’	פ	
ق	qāf	ק	
ك	kāf	כ	
ل	lām	ל	
م	mīm	מ	
ن	nūn	נ	
ه	hā’	ה	
ة	tā’ marbūṭa	ה	
و	wāw	ו	
ي	yā’	י	
ى	alif maqṣūra	י	
ؤ	hamza on wāw	א	
ئ	hamza on yā’	י	
ء	hamza	א / ∅	See note 2

Table 5: Arabic to Judeo-Arabic (Hebrew Script) Transliteration Table

Notes:

1. **Jīm/Ghayn variation:** The algorithm models regional and scribal variation. 80% of the documents (chosen randomly) were transliterated with jīm (ج) rendered as ג (without geresh) and ghayn (غ) as ג’. In the remaining 20%, jīm → ג’ and ghayn → ג. This approximates the variation found in actual Judeo-Arabic manuscripts, as per our quantitative inspection of a sample of the corpus.

2. **Hamza:** Hamza (ء on the line) is randomly rendered as either א or omitted entirely, reflecting orthographic fluidity in medieval Judeo-Arabic texts.

3. **Final letter forms (*sofiot*):** Hebrew letters in word-final position are automatically converted to their final forms: מ→ם, נ→ן, צ→ץ, כ→ך, פ→ף. The algorithm correctly handles cases where the final letter carries a geresh modifier.

B. Appendix: JA Dataset Pre-process

B.1. PGP Corpus Preprocessing

The following algorithm was used to preprocess the PGP corpus for training.[14] In addition to the process in the pseudocode below, a small number of Hebrew translations inserted in between original JA transcriptions were removed. These can be found by searching for the word תרגום surrounded by word boundaries and inspecting manually.

Algorithm 1 PGP Corpus Preprocessing

Require: Raw transcription text T, word-length parameter N (default: 5)
Ensure: Cleaned text with gap markers

Phase 1: Bracket and Parenthesis Normalization
Replace empty bracket pairs `[ ]` or `( )` and adjacent non-whitespace with `[GAP]`
Replace unmatched `]` at line start or `[` at line end (along with adjacent text) with `[GAP]`
Remove content within curly braces `{...}` and rasura brackets ⟦...⟧
Remove remaining brackets, parentheses, and symbols: `[ ] ( ) / \ | ! ? _`

Phase 2: Dot Sequence Handling
for all sequences of one or more periods with surrounding (non-whitespace) text **do**
    if single period immediately following text and preceding whitespace **then**
        Keep as sentence-final punctuation
    else
        $c \leftarrow$ count of non-space characters in matched sequence
        if $c \leq N$ **then**
            Replace with `[ONEGAP]`
        else
            Replace with `[GAP]`
        end if
    end if
end for

Phase 3: Gap Consolidation
Replace consecutive gap markers with single `[GAP]`

Phase 4: Character Filtering
Remove lines containing Arabic script
Remove Latin alphanumeric characters (preserving gap markers)
Truncate documents at editorial markers ("footnotes", "הערות")

[14] The algorithm was developed based on the transcription conventions described at Princeton Geniza Project (2021) as well as some manual inspection of the corpus.

B.2. FJMS Corpus Preprocessing

The following algorithm was used to preprocess the FJMS corpus for training.[15]

Algorithm 2 FJMS Corpus Preprocessing

Require: Raw transcription text T, word-length parameter N (default: 5)
Ensure: Cleaned text with gap markers

Phase 1: Editorial Annotation Removal
Remove content within parentheses `(...)`

Phase 2: Character Filtering
Remove all characters except: Hebrew letters, period, apostrophe, quotation marks, geresh (`U+05F3`), gershayim (`U+04F4`), horizontal ellipsis (`U+2026`), and whitespace

Phase 3: Multi-Period Sequence Handling
for all runs of ≥ 2 consecutive periods (allowing internal spaces) with surrounding (non-whitespace) text **do**
    $c \leftarrow$ count of non-space characters in matched sequence
    if $c \leq N$ **then**
        Replace with `[ONEGAP]`
    else
        Replace with `[GAP]`
    end if
end for

Phase 4: Horizontal Ellipsis Handling
▷ The Unicode character `U+2026` (...), distinct from three periods
for all sequences containing horizontal ellipsis characters with surrounding text **do**
    if multiple horizontal ellipsis characters in sequence **then**
        Replace with `[GAP]`
    else
        Replace with `[ONEGAP]`
    end if
end for

Phase 5: Normalization
Replace consecutive gap markers with single `[GAP]`
Collapse multiple spaces to single space
Remove spaces immediately preceding periods

[15] The algorithm was developed based on manual inspection of the corpus.

B.3. OpenITI Corpus Preprocessing

The following algorithm was used to preprocess the OpenITI corpus for training.[16] The OpenITI corpus contains Arabic texts in Arabic script. Prior to preprocessing, these texts were transliterated into Hebrew characters using the table in Appendix A.

Algorithm 3 OpenITI Corpus Preprocessing (Post-Transliteration)

Require: Transliterated text T
Ensure: Cleaned text suitable for training

Phase 1: Metadata Removal
1: Delete all lines beginning with `#META#`

Phase 2: Line Continuation Handling
2: Replace newline followed by `~~` with a single space

Phase 3: Editorial Section Removal
3: **for all** markers $m \in \{$`|EDITOR|`, `|APPENDIX|`, `|PARATEXT|`$\}$ **do**
4: Delete all text from m until the next line beginning with `###`
5: **end for**

Phase 4: Character Filtering
6: Remove all characters except: Hebrew letters, period, apostrophe, and geresh (`U+05F3`)

Phase 5: Whitespace Normalization
7: Strip leading and trailing whitespace from each line
8: Collapse multiple consecutive spaces to single space

B.4. Blau Dictionary Preprocessing

After extracting the Judeo-Arabic quotes from Blau (2006), which have a distinct font, a few minor adjustments were made: ellipses were replaced with "`[GAP]`", dots above letters (`U+0307`) were normalized to apostrophe, and all other diacritics were removed.

B.5. Post-preprocess For All Corpora

In each corpus, after its respective preprocess, blocks of text with fewer than three words were removed to make sure that at least some minimal context was provided for each word during training. Finally, geresh (`U+05F3`) was normalized to apostrophe (`U+0027`) and gershayim (`U+04F4`) to quotes (`U+0022`).

[16] The algorithm was developed based on the transcription conventions described at Romanov (2023) as well as some manual inspection of the corpus.

C. Appendix: Annotation Protocol

For the `JABERT` Cairo Genizah Dataset (Section 4.1), which contains transcriptions of 3862 lines of naturally-occurring Judeo-Arabic text in Cairo Genizah documents, we instructed the Judeo-Arabic expert with the following transcription protocol:

- Overall, the objective is to capture the letters of the manuscripts exactly as they are, without normalizing any orthographic oddities or mistakes.
- Diacritic vocalization marks (that is, diacritics which indicate vowels) shall not be transcribed.
- Consonantal diacritic dots (that is, dots placed above letters to indicate a variant of the consonant) shall be transcribed; however, if the consonantal diacritic dot is absent from the manuscript, it shall not be filled in, even if the word appears to require it, because Judeo-Arabic scribes often choose to omit these diacritic dots.
- If the manuscript employs a single ligature in place of the two-letter sequence *aleph* and *lamed*, then it shall be transcribed as a single character, using the Unicode point designated for this ligature (U+FB4F).
- If the manuscript employs a *rafe* mark (a horizontal line above a letter), then it shall be transcribed using the corresponding Unicode point (U+05BF).
- Finally, because Cairo Genizah manuscripts can be difficult to read in places, and because scribal strokes can be ambiguous at times, and because only one Judeo-Arabic expert was employed to do the transcriptions for this first dataset, we instructed our Judeo-Arabic expert to completely skip any line which posed an ambiguity or reading difficulty. Indeed, because the aim of this dataset is to evaluate a model's ability to predict a word based on the context of the individual line, the elision of lines which happen to be difficult to decipher does affect the effectiveness of the dataset. Nevertheless, in future work, we plan to have the lines transcribed independently by a second annotator, so that we may properly evaluate inter-annotator agreement and indicate cases of disagreement.

D. Appendix: Homograph Dataset

Table 6 presents the complete list of homographs used in our evaluation dataset. Each homograph is a word spelled identically in unvocalized Judeo-Arabic text but carrying distinct meanings depending on context. The "Arabic" column shows the vocalized Arabic form that disambiguates pronunciation and meaning. Arabic verb forms are indicated: Form I (basic stem), Form II (intensive/causative), act. = active, pass. = passive. The 16 homographs yield 37 distinct meanings across 1,119 annotated samples.

#	Translit.	Arabic	Meaning
1. אלא (*'l'*)			
1	*alā*	أَلَا	Interrog.: "is it not?"
2	*illā*	إِلَّا	"except", "other than"
2. מן (*mn*)			
3	*min*	مِن	Prep.: "from"
4	*man*	مَن	Rel. pron.: "whoever"
3. מא (*mā*)			
5	*mā*	مَا	Int./Neg.: "what?"/"not"
6	*mā*	مَا	Rel. pron.: "that which"
4. צרב (*ṣrb*)			
7	*ḍarb*	ضَرب	Noun: "type", "manner"
8	*ḍaraba*	ضَرَبَ	Verb: "struck"
9	*ḍaraba*	ضَرَبَ	Verb: "gave a parable"
5. עלי (*'ly*)			
10	*'alā*	عَلَى	Prep.: "on", "upon"
11	*'alayya*	عَلَيَّ	Prep.+pron.: "on me"
6. קסם (*qsm*)			
12	*qasama*	قَسَمَ	Verb: "divided"
13	*qasam*	قَسَم	Noun: "oath"
14	*qism*	قِسم	Noun: "portion"
15	*qasama*	قَسَمَ	Verb: "swore"
7. יערף (*y'rf*)			
16	*ya'rifu*	يَعرِفُ	Verb (I act.): "knows"
17	*yu'rafu*	يُعرَفُ	Verb (I pass.): "is known"
18	*yu'arrifu*	يُعَرِّفُ	Verb (II): "informs"
8. ראי (*r'y*)			
19	*ra'ā*	رَأَى	Verb: "saw", "thought"
20	*ra'y*	رَأي	Noun: "opinion"
9. קבל (*qbl*)			
21	*qabila*	قَبِلَ	Verb: "received"
22	*qabla*	قَبلَ	Prep.: "before"
10. דכר (*ḏkr*)			
23	*ḏakara*	ذَكَرَ	Verb: "mentioned"
24	*ḏikr*	ذِكر	Noun: "memory"
25	*ḏakar*	ذَكَر	Noun: "male"
11. פרק (*frq*)			
26	*faraqa*	فَرَقَ	Verb: "separated"
27	*farq*	فَرق	Noun: "difference"
12. עאלם (*'lm*)			
28	*'ālam*	عَالَم	Noun: "world"
29	*'ālim*	عَالِم	Adj.: "knows"
13. כלק (*ẖlq*)			
30	*ẖalaqa*	خَلَقَ	Verb: "created"
31	*ẖalq*	خَلق	Noun: "creation"
14. גיר (*ġyr*)			
32	*ġayr*	غَير	Neg.: "not", "non-"
33	*ġayr*	غَير	Prep.: "other than"
15. פהם (*fhm*)			
34	*fahm*	فَهم	Noun: "understanding"
35	*fa-hum*	فَهُم	Conj.+pron.: "and they"
16. אכד (*'kd*)			
36	*akkada*	أَكَّدَ	Verb (II): "confirmed/affirmed"
37	*aẖaḏa*	أَخَذَ	Verb (I): "took"

Table 6: Judeo-Arabic homographs: word forms and meanings

BEReshiT: an Ancient Hebrew Model based on DictaBERT

Iglika Nikolova-Stoupak[1], **Maxime Amblard**[1], **Frédérique Rey**[2]
[1]LORIA, UMR 7503, Université de Lorraine, CNRS, Inria, 54000 Nancy, France
[2]Zentralinstitut für Katholische Theologie, Humboldt-Universität zu Berlin, 10099 Berlin, Germany
{iglika.nikolova-stoupak, maxime.amblard}@univ-lorraine.fr, frederique.rey@hu-berlin.de

Abstract

This project addresses the general absence of Natural Language Processing (NLP) tools when it comes to historical languages as a subset of low-resource languages that is relevant to an array of academic disciplines from linguistics to textual criticism. In particular, we train an Ancient Hebrew language model, BEReshiT, as well as BEReshiT-morph, a submodel for morphological annotation. BEReshiT is achieved through the fine-tuning of DictaBERT, a state-of-the-art model for Modern Hebrew that has also proved useful in Biblical Hebrew tasks. Layer freezing is applied in order to achieve maximal results while gaining insight about the adaptation process. In the context of an elaborate cloze test, BEReshiT demonstrates increased performance and notions of the Ancient Hebrew language compared to the source model as well as a selection of additional relevant models. The submodel BEReshiT-morph performs highly on tasks of morphological classification, reaching an F1 score of 0.97 for part-of-speech (POS) tagging. We release the main and morphological models as well as the datasets used at training and evaluation.

Keywords: BERT, layer freezing, Ancient Hebrew, cross-variety transfer, model interpretability

1. Introduction and Motivation

It is well-known that the NLP field is dominated by data and tools that pertain to high-resource languages (Joshi et al., 2020). Whilst low-resource languages have been receiving increased attention in recent years, work with them still constitutes a challenge (Hedderich et al., 2021; McGiff and Nikolov, 2025). In particular, historical languages are associated with a limited number of texts, which are moreover typically homogenous in register and complexity. Yet, these languages are of key importance within a number of research disciplines, such as linguistics, archaeology, theology and textual criticism.

We are focusing on Ancient Hebrew, the language of the Hebrew Bible and of additional texts similar in time frame and linguistic characteristics. Ancient Hebrew differs from Modern Hebrew in terms of morphology, syntax and, most notably, vocabulary. For more information about the Hebrew language and its stratification, please refer to: Khan et al. (2013), Hurwitz (2014), Hornkohl (2014) and Rezetko and Young (2014).

We are herein presenting our language model, BEReshiT, along with the submodel BEReshiT-morph, which predicts the morphological characteristics of Ancient Hebrew words in context. BEReshiT is achieved through the fine-tuning of a state-of-the-art BERT-based model specialising in Modern Hebrew, DictaBERT (Shmidman et al., 2023). In order to make optimal use of the source model as well as to gain insight about the relationship between BERT architecture and the Hebrew language (in its different varieties), we trained twelve versions of the model, each corresponding to a different scenario of layer freezing.

The trained models, as well as a representative selection of established Hebrew or multilingual models, were evaluated on specially crafted datasets, which underline the specificities of the Ancient Hebrew language as well as notable differences with its Modern counterpart. These datasets, as well as the strongest BEReshiT model and the BEReshiT-morph submodel, are publicly released in the following repository: https://zenodo.org/records/18750374. Our main contribution consists in the Ancient Hebrew model and associated submodel artefacts rather than in novelty in terms of aspects such as training and evaluation. The models may be used in future work with Ancient Hebrew texts that have been newly discovered or digitalised or that have not yet received high levels of attention (notably, morphological annotation).

2. Background

There are several Hebrew language models to date. In particular, the Dicta team[1] has developed high performing models that focus on different aspects and varieties of the language. DictaLM are recent Hebrew-based large language models (LLMs) (Shmidman et al., 2024b). In this study, we chose to focus on a BERT-based model (Devlin et al., 2019) for several reasons. BERT models systematically capture rich morphological and syntactic information across layers, which is particularly important for a morphologically complex language such as Hebrew. In addition, the architecture has been shown to allow for successful experiments that imply explainability and interlingual comparison. Also, task- and language-specific models such as DictaBERT

[1]https://dicta.org.il/

Proceedings of LT4HALA 2026 @ LREC 2026, pages 72–88.
11 May 2026. *©ELRA Language Resources Association (ELRA), 2026*

have achieved strong performance on NLP tasks, often dominating the state of the art. Finally, the relatively small size of BERT models makes them more suitable than other high-performing models, such as LLMs, in view of computational and environmental constraints.

MsBERT is trained for the task of filling in lacunae in Hebrew manuscripts (Shmidman et al., 2024a). In turn, BEREL (Shmidman et al., 2022) specialises in Rabbinic Hebrew, a stratum that has significant links with Ancient Hebrew in terms of linguistics and register. However, we opted for the Modern Hebrew model DictaBERT (Shmidman et al., 2023) as a model to finetune for the following reasons: it is associated with a strong morphological submodel, whose performance has brought improvement even on tasks related to Ancient Hebrew (Nikolova-Stoupak et al., 2025); also, BEREL has already been trained on the Hebrew Bible as well as additional sources that reference it and is therefore not a good candidate for domain adaptation through fine-tuning on available Ancient Hebrew text. To our knowledge, no uniquely Ancient/Biblical Hebrew language model exists to date.

BERT- and RoBERTa-based models have been trained for a large variety of languages, including historical ones such as Latin (Bamman and Burns, 2020) and early modern French (Gabay et al., 2022). HmBERT is a multilingual model (a.k.a. one-model) of historical languages[2] that specialises in Named Entity Recognition (NER) (Schweter et al., 2022). The model improves on single-model benchmarks for the German language whilst falling short for the other languages. Fine-tuning of previously trained models has also been used in adaptation to low-resource historical languages. For instance, Lendvai et al. (2023) fine-tune several multilingual BERT models on historical varieties of Slavic languages for the purpose of identification of manuscripts and related tasks, achieving highest results when using models that already specialise in Cyrillic[3].

Various experiments have been carried out as a quest to gain insight about BERT's structure with regards to language knowledge. Jawahar et al. (2019) conclude that the model captures linguistic information about English in tree-like structures. Working with English, French and Swedish on the task of proficiency level determination, Muñoz Sánchez et al. (2024) explore layer freezing at BERT fine-tuning, discovering that optimal architectures differ for each language.

[2]historical German, English, French, Swedish and Finnish

[3]KoichiYasuoka/bert-base-slavic-cyrillic-upos and anonsubmissionmk/bert-base-macedonian-bulgarian-cased

3. Data

3.1. Model Training

We trained the BEReshiT models on the whole Hebrew[4] text of the Hebrew Bible (373 093 words/28 379 verses) and clean[5] Hebrew verses from the Dead Sea Scrolls (DSS), a series of both Biblical and extra-Biblical manuscripts discovered from 1947 onwards (67 806 words/5955 lines)[6]. The two sets of texts were used as available, respectively, through the Open Scriptures Hebrew Bible (OSHB)[7] and ETCBC's The Dead Sea Scrolls Text-Fabric dataset[8]. In accordance with DictaBERT, we used the established Hebrew character set consisting of consonantal letters only, excluding diacritical marks such as vowel signs (*nikkudot*) and cantillation marks. In the case of the Hebrew Bible, we duplicated one quarter of the data and included it a second time along with *nikkudot* in order to ensure that the model has knowledge of this later convention whilst not overburdening it with what it would treat as new vocabulary.

For the purpose of training the morphological submodel, we created a token-level dataset based on the same data along with morphological information as present in the two data sources and mapped in cases of difference in notation. To limit the number of morphological features per token, prefixes were separated from the word they modify.

3.2. Model Evaluation

We composed a 200-verse dataset that seeks to capture the general qualities of the Ancient Hebrew language. For half of the dataset, we selected Biblical verses as divided into discrete categories in relation to literary genre or time-period that have been standardly distinguished in Biblical Hebrew scholarship as exhibiting different linguistic features: "narrative"[9], "legal"[10], "prophetic"[11], "poetry"[12], "wisdom'[13] and "late"[14]. Note that the list is not unequivocally accepted and that the same text

[4]Aramaic parts were discarded.

[5]Lines with reconstructed or missing text were discarded.

[6]For more information about the Dead Sea Scrolls, please refer to djd (1955–2009).

[7]https://hb.openscriptures.org/

[8]https://github.com/ETCBC/dss

[9]30 verses; from "Gen", "Exod", "Num", "Josh", "Judg", "Ruth", "1 Sam", "2 Sam", "1 Kgs", "2 Kgs", "Esth", "Jonah"

[10]15 verses; from "Deut", "Lev"

[11]15 verses; from "Isa", "Jer", "Ezek", "Hos", "Joel", "Amos", "Obad", "Mic", "Nah", "Hab", "Zeph", "Hag", "Zech", "Mal"

[12]15 verses; from "Lam", "Ps", "Song"

[13]10 verses; from "Eccl', "Job', "Prov'

[14]15 verses; from "1 Chr", "2 Chr", "Dan", "Ezra", "Neh"

could be classified as belonging to more than one category ("Eccl" may be "late" as well as "wisdom"), in which case we have made selections in view of quantitative balance.

In order to specifically test our models' proficiency in Ancient as opposed to Modern Hebrew, we based the second half of the dataset around five phenomena as absent/rare in the latter but commonly encounterable in the former (20 verses per phenomenon): "wayyiqtol narrative chains", "combination of infinitive absolute and finite verb", "verb subject object", "suffixed possessive pronouns" and "particles". The last category includes verses that contain the particles ה[15], הנה־נה[16], נא[17] and the relative particle אשר (in place of the particle –ש, which came to gradually be more common). The included phenomena are not meant to be exhaustive of the differences between the concerned language varieties.

For the purpose of evaluating the performance of the BEREL and GPT models on data that thay have not been trained on, we also created an alternative dataset of 200 DSS lines (100 Biblical and 100 extra-Biblical). In the case of the Biblical lines, BEREL has already encountered the same text with alternative orthography, and the extra-Biblical lines are entirely unfamiliar to it. We do not have this level of knowledge concerning the GPT model's training, but a similar tendency is expected in view of the data's general online presence.

The morphological model was evaluated on 10% of the original dataset.

The data used for model evaluation was removed from the training datasets.

4. Methods

12 different versions of the BEReshiT model were trained, each corresponding to a different number of "frozen" layers within the original DictaBERT model. As with DictaBERT, standard BERT-based training was followed. Tokenisation was based on full words rather than subword tokens, as proven optimal in relation to DictaBERT (Shmidman et al., 2023). A simple grid search was used in the configuration of each model. The strongest performing model in terms of accuracy, which had 9 frozen layers (3 trainable layers), was trained for 3 epochs with a learning rate of 0.0001 and word-mask probability of 0.2. All models were trained on one GPU Nvidia L40S 45GB. Training time ranged between 16.72 min (for 12 trainable layers) and 79.42 min (for 9 trainable layers). The models were evaluated on a cloze test based on the evaluation dataset as presented in 3.2. A random word was masked within each verse/line.

For the purpose of inter-model comparison, we opted for the following selection of models: DictaBERT, BEREL[18], mBERT (Devlin et al., 2019) and GPT-4o (OpenAI, 2024). A comparison of our models with DictaBERT allows for clear evaluation of the improvement resulting from the fine-tuning process. It is also meaningful to juxtapose our models to BEREL, which specialises in Rabbinic Hebrew, a stratum that resembles Ancient Hebrew more than Modern Hebrew does. Whilst remaining within the framework of BERT models, mBERT[19], which was trained on 2k tokens of Modern Hebrew data, allows for a comparison between language-specific and multilingual transformer models in relation to the Hebrew language[20]. Finally, OpenAI's GPT-4o is included as a representative of LLMs. The model is well documented as exhibiting strong performance on non-English text (Zhu et al., 2024; Harigai et al., 2025; Shvartz et al., 2025). Within our prompt, the model is instructed to return solely the missing word in each example and to make a guess even if unsure. In the case of GPT and BEREL, evaluation is also performed on the separate DSS-based dataset as likely unseen by them during train time.

As a way to gain further insight about the compared models' characteristics, we also performed qualitative evaluation based on a sample of their predictions, focusing on the following: orthographic variations, (near-)synonyms, complex verb forms and infrequent words.

To develop the morphological model, task fine-tuning was applied using the BEReshiT encoder[21] and multi-task morphology heads for all morphological features. Loss was computed only for non-NA labels. Tokens were predicted with reference to their local context rather than in isolation. The values for features irrelevant for the predicted POS (e.g. "conjugation" for non-verbs) were forced to be "NA". Once again, a grid search was conducted. The strongest model (mean F1 0.76) was trained for 5 epochs at a learning rate of 0.0001 and with maximal input length of 256 tokens. Macro-F1 was utilised as primary metric as a means to provide a comprehensive view on the model's capacities, independent of the relevant complexity of the predictable information. A comparison was performed between our morphological model (named BEReshiT-morph) and DictaBERT's counterpart,

[15] question particle

[16] presentative particle/discourse marker; related forms are also included

[17] denotes entreaty/politeness

[18] as per its most recent version, 3.0

[19] as per BERT-BASE-MULTILINGUAL-CASED

[20] As the model uses WordPiece subwords, in the case of the gold label becoming multiple tokens, a pseudo-likelihood is computed through masking one token at a time.

[21] Our strongest model was used.

DictaBERT-morph. For the purpose, the relevant morphological categories were mapped to the best of our abilities (see Table 3).

5. Results

5.1. Quantitative

5.1.1. Main Models

Please refer to the results presented in Table 1. For more detailed results, including per category in addition to dataset, please refer to Appendix A. Overall, the reported accuracies are low, which is explainable through the challenging nature of the evaluation dataset (e.g. a large percentage of low-frequency words in Biblical text). Consistently lowest performance by a large margin is associated with mBERT. The best global performance is demonstrated by the GPT and BEREL models. BEReshiT scores better than BEREL solely for the "prophetic" category. However, it is crucial to note that BEREL is trained on the Sefaria repository[22], which includes the full text of the Hebrew Bible (Shmidman et al., 2022), thereby causing train-test contamination. When only the DSS examples from the same dataset are considered, the model's accuracy drops significantly to 0.18 (2 correct predictions out of 11); to go further, both of the model's correct answers come from Biblical scrolls[23], which it has already encountered during training, albeit with different orthography. In turn, GPT is trained on large amounts of undisclosed online data, which is likely to also include Biblical texts. Once again, when only DSS texts are considered, the accuracy drops to the low value of 0.09, the only correct answer being from the Biblical scroll 1QIsaa.

When GPT and BEREL results are disregarded due to train-test contamination, BEReshiT demonstrates the highest accuracy for our main evaluation dataset, surpassing DictaBERT by a margin ranging from 0.02 to 0.09 in its different configurations. Increase is especially high for the following phenomena: "particles", "vso", and "wayyiqtol"[24]. The globally strongest BEReshiT model is the one with 9 frozen layers (for subset "by phenomenon", the models with 3 and 4 frozen layers perform slightly better).

We went on to evaluate the two contaminated models as well as DictaBERT on the additional DSS-only dataset[25]. In this context, performance decreased significantly, results being by far worse in the case of extra-Biblical texts as likely completely unseen by the models. The fact that DictaBERT's performance was identical to BEREL's on unseen text reassures us that the choice of DictaBERT as the model to fine-tune has not constituted a compromise in terms of potential performance.

An accuracy of 0 is occasionally demonstrated for portions of the data: notably, by mBERT as well as by most models for the "wisdom" category. In these cases, the alternative metric "probability of gold" (i.e. the probability that the model attributes to the correct label, whether it is its first prediction or not) helps provide an interpretable non-zero (albeit still very low) value. "Probability of gold" is close in value to "accuracy" for all experiments, implying that there are no significant cases of correct labels being rated highly whilst not representing the top choice. Interestingly, the metric is typically lower than "accuracy" for all models except mBERT, for which the tendency is reversed. The model can therefore be judged to spread its probabilities more evenly among candidates. Note that "probability of gold" is not retrievable for the proprietary model GPT.

5.1.2. Morphological Submodel

Table 2 shows the performance of the model BEReshiT-morph in predicting each of 9 morphological features (see Appendix B.1 for more detailed results). The mean F1 score across the full evaluation dataset comes at 0.76. When the different coverage of each feature is accounted for (column "cov.-weighted"), the value increases to 0.82[26]. Notably, the result for POS tagging is 0.97. Albeit not uniformly, performance is slightly lower for the DSS portion of the dataset and slightly higher when only extra-Biblical DSS tokens are considered. Note that here, as train-test contamination is not an issue, these trends reflect solely linguistic differences in the subsets.

[22] https://www.sefaria.org/texts

[23] 1QIsaa and 4Q5, which denote, respectively, the books Isa and Gen.

[24] respectively, from 0.15 to 0.35, 0.15 to 0.35, and 0.3 to 0.5

[25] Naturally, BEReshiT models, which were trained on this data, were not evaluated.

[26] Not every category is relevant for every evaluated token. For instance, POS tags are available for 0.84 of the evaluation data, whilst "suffix type" labels - for 0.02.

	main dataset			DSS-only		
	overall	by category	by phenomenon	overall	Biblical	extra-Biblical
GPT-4o	**0.54 (-)**	**0.58 (-)**	0.50 (-)	**0.24 (-)**	**0.39 (-)**	**0.09 (-)**
mBERT	0.05 (0.06)	0.04 (0.05)	0.06 (0.07)			
DictaBERT	0.27 (0.21)	0.33 (0.25)	0.22 (0.16)	0.12 (0.07)	0.16 (0.09)	0.08 (0.05)
BEREL	**0.54 (0.49)**	0.53 (0.49)	**0.54 (0.49)**	0.20 (0.16)	0.32 (0.26)	0.08 (0.05)
BEReshiT (f.l.:1-11)	0.32 (0.22)	0.36 (0.26)	0.27 (0.17)			
BEReshiT (f.l.:1-10)	0.34 (0.25)	0.37 (0.29)	0.30 (0.22)			
BEReshiT (f.l.:1-9)	0.36 (0.25)	0.40 (0.29)	0.32 (0.21)			
BEReshiT (f.l.:1-8)	0.35 (0.25)	0.39 (0.28)	0.31 (0.21)			
BEReshiT (f.l.:1-7)	0.33 (0.24)	0.38 (0.28)	0.28 (0.20)			
BEReshiT (f.l.:1-6)	0.30 (0.22)	0.35 (0.26)	0.25 (0.19)			
BEReshiT (f.l.:1-5)	0.29 (0.21)	0.33 (0.24)	0.25 (0.18)			
BEReshiT (f.l.:1-4)	0.34 (0.25)	0.34 (0.26)	0.33 (0.23)			
BEReshiT (f.l.:1-3)	0.34 (0.25)	0.34 (0.27)	0.33 (0.23)			
BEReshiT (f.l.:1-2)	0.32 (0.21)	0.33 (0.24)	0.30 (0.19)			
BEReshiT (f.l.:1)	0.33 (0.23)	0.35 (0.26)	0.30 (0.20)			
BEReshiT (no f.l.)	0.32 (0.23)	0.36 (0.26)	0.27 (0.19)			

Table 1: Results of the evaluation of the discussed models on the main dataset and the DSS-only dataset (as well as their subsets). Primary metric: accuracy. Secondary metric (noted between parentheses): probability of the gold label, where retrievable. The best global results are indicated in **bold** and the best results when models with likely train-test contamination are disregarded are underlined. Numbers are rounded to the second digit after the decimal point. f.l.: frozen layer(s).

	by morphological feature (F1)									**overall (F1)**	
	pos	subpos	stem	conj.	person	gender	number	state	suffix type	**mean**	**cov.-weighted**
full	0.97	0.88	0.72	0.79	0.74	0.73	0.72	0.64	0.63	0.76	0.82
DSS	0.94	0.76	0.72	0.54	0.72	0.71	0.67	0.62	0.62	0.70	0.77
extra-Biblical DSS	0.94	0.78	0.73	0.60	0.96	0.92	0.96	0.62	0.49	0.78	0.85

Table 2: Morphological evaluation results for BEReshiT-morph: full dataset, DSS subset (14.98%) and extra-Biblical part of the DSS subset (5.66%). The reported metric is macro-F1. Numbers are rounded to the second digit after the decimal point.

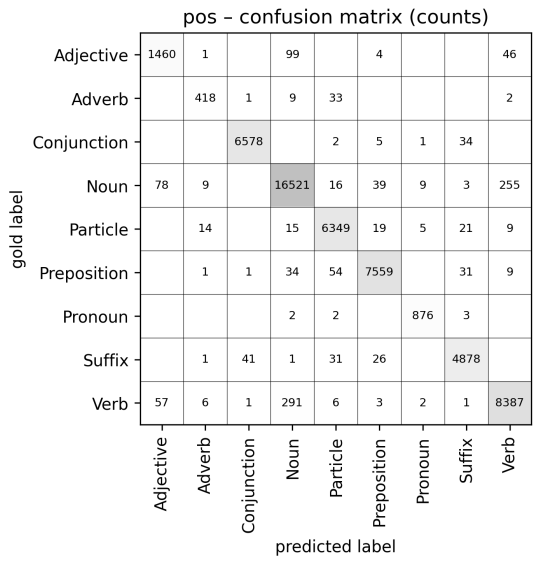


Figure 1: Confusion matrix of model BEReshiT-morph for the feature “pos”.

What follows is an overview of the model’s most common errors. Feature “pos” is mistakenly labelled “noun” when the gold label is “verb” 3.32% of times, and the reverse takes place 1.51% of times (see Fig. 1). “2nd person” is occasionally marked as “3rd” (1.31%) and vice-versa (0.65%). “Plural” number is marked as “singular” in 1.87% of cases, and “feminine” gender is marked as “masculine” in 4.46%, whilst the reverse occurrences are much less significant. Within feature “subpos”, “proper” is wrongly labelled as “common” in 2.33% of instances. Also, “common” is sometimes predicted as “UNK” (1%). Wrong “UNK” labels are also prominent when it comes to gold values “participle” (feature “conjugation”), “piel” (feature “stem”) and

"pronoun" (feature "suffix type")[27]. For the full confusion matrices per morphological feature, please refer to Appendix B.2.

As a next step, we took on to compare the performance of BEReshiT-morph to that of DictaBERT-morph, a morphological submodel based on DictaBERT. The first challenge came in the face of the two models' different tokenisation (subword-based for our model vs word-based for DictaBERT-morph). To guarantee fairness, we carried out the evaluation on DictaBERT-morph using a non-segmented version of the dataset, following which we segmented all prefixes and associated them with the labels that the model indicated as relevant to them. We proceeded to reduce the evaluation dataset to only the verses where tokenisation was now identical between the two models, which left us with a total of 920 verses. Finally, we mapped the morphological labels as existent in the two models (for the original label sets, please refer to Appendix C). This implied the occasional removal, combination and division of labels as well as the replacement of others with "NA". Please see Table 3 for the derived labels. As expected due to the morphological differences between Ancient and Modern Hebrew, DictaBERT-morph's performance in the task is significantly inferior (see Table 4). When only DSS tokens as well as only extra-Biblical DSS tokens are considered, the gap between the two models' performance increases further.

5.2. Qualitative

We conducted an additional microscopic, qualitative-based analysis, with the goal of gaining deeper knowledge about the performance of the investigated models, including different versions of BEReshiT (namely, the strongest performing version, where 9 layers are frozen during training, and the version that underwent full training).

See Table 5 for a representative example of the models' output. Common problems that are associated with all models, but most prominent with mBERT, are predictions consisting merely of punctuation, word fragments and simple particles, prepositions or pronouns. mBERT proposed punctuation in 73 out of the 200 examples of the main evaluation dataset, followed by DictaBERT (23), whilst for the BEReshiT models with 3 and 12 trained layers, there are respectively only 2 and 1 examples of the phenomenon.

Orthographic differences. The BEREL model is associated with the highest number of "wrong answers" that are in fact simply spelling variations of the gold label (15 over all 400 examples) i.e. one is the plene (including matres lectionis, typically more archaic) and the other is the defective spelling of the same word. Interestingly, there is no specific tendency as to the type of spelling preferred by the model as well as by DictaBERT and GPT, which are also associated with several examples of a predicted orthographic variant of the gold label[29]. There are only two examples of orthographic differences with the gold label for the BEReshiT

Feature	DictaBERT-morph label	Mapped label	BEReshiT-morph label
POS	VERB AUX	verb	verb
	NOUN	common noun	noun (common)
	PROPN	proper noun	noun (proper)
	ADJ NUM	adjective	adjective
	ADV	adverb	adverb particle (accusative-marker)
	ADP	preposition	preposition particle (object-marker)
	CCONJ	conjunction	conjunction
	DET	definite article	particle (article/definite)
	PRON	pronoun	pronoun
	SCONJ	relative particle	particle (interrogative)
		interrogative particle	particle (interrogative)
	INTJ	*NA*	
	SYM	*NA*	
	X	*NA*	
	suffix[28]	suffix	suffix
tense	Past	perfect	perfect
	Pres	participle	participle
	Fut	imperfect,wayyiqtol, jussive	imperfect wayyiqtol jussive
	Imp	imperative	imperative
		NA	cohortative
		NA	infinitive-absolute
		NA	infinitive-construct
person	1	1	1
	2	2	2
	3	3	3
	1, 2, 3	*NA*	
gender	Masc	masc	masc
	Fem	fem	fem
	Fem,Masc	comm	comm
number	Sing	sing	sing
	Plur	plur	plur
	Dual	dual	dual
suffix type	PRON ADP_PRON	pronoun	pronoun
		NA	directional *he*

Table 3: Mapping of the morphological features of models DictaBERT-morph and BEReshiT-morph. The values between parentheses denote BEReshiT-morph "subpos" labels.

[27] Respectively, in 4.1%, 6.86%, and 7.42% of cases.

[29] e.g. BEREL predicted תשכון (tiškon "dwell-FUT.2SG.M") in place of תשכן but also דוד (david "David") in place of דויד.

	by morphological feature (F1)						overall (F1)	
	pos	tense	person	gender	number	suffix type	**mean**	**cov.-weighted**
full dataset								
BEReshiT-morph	0.96	0.72	0.74	0.73	0.73	0.50	0.76	0.82
DictaBERT-morph	0.54	0.42	0.68	0.49	0.46	0.46	0.51	0.52
DSS								
BEReshiT-morph	0.94	0.37	1.00	0.95	0.93	0.49	0.78	0.91
DictaBERT-morph	0.55	0.34	0.61	0.46	0.43	0.49	0.47	0.50
extra-Biblical DSS								
BEReshiT-morph	0.95	0.48	1.0	0.94	0.95	1.0	0.89	0.93
DictaBERT-morph	0.30	0.26	0.40	0.30	0.24	0.21	0.28	0.29

Table 4: Comparison of the morphological submodels BEReshiT-morph and DictaBERT-morph using mapped features and labels for the full evaluation dataset vs the DSS subset (8.48%) vs the extra-Biblical DSS subset (4.09%). The reported metric is macro-F1. Numbers are rounded to the second digit after the decimal point.

Reference	**GPT-4o**	**mBERT**	**DictaBERT**	**BEREL**	**BEReshiT**
משק	משק	*punctuation*	*punctuation*	משק	סורר
mešeq	mešeq	–	–	mešeq	sorer
“steward, heir”	“steward, heir”	–	–	“steward, heir”	“rebellious”
ויאמר אברם אדני יהוה מה תתן לי ואנכי הולך ערירי ובן *משק* ביתי הוא דמשק אליעזר (Gen 15:2)					
But Abram said, “O Lord GOD, what will you give me, for I continue childless, and the *heir* of my house is Eliezer of Damascus?” (Gen 15:2)					

Table 5: Representative example of predictions provided by the investigated models. Typically for train-test contamination, BEREL and GPT guess the gold label. mBERT and DictaBERT propose punctuation. BEReshiT proposes a wrong vocabulary item.

model that underwent full training: לוא/לא (lo “no”) and אלהם/אליהם (alehem “to them”). In contrast, the model with three trained layers predicts the gold label for the later (the defective spelling), remaining with only a single example of the phenomenon.

Synonyms. Another type of prediction that deviates from the gold label but that may be considered correct is that of a synonym or close synonym. In the context of the verse ואל־בני ישראל תאמר איש איש מבני ישראל ומן־הגר הגר בישראל אשר יתן מזרעו למלך מות יומת עם הארץ ירגמהו באבן (Lev 20:2)[30][31], the gold label איש (iš “man, one”), which is predicted correctly by the contaminated BEREL model[32], is synonymous with the prediction put forward by both BEReshiT models as well as by their ancestor, DictaBERT - כל[33]. To go a little further, there are cases where an incorrect prediction is, while not strictly speaking synonymous, still relevant in terms of both morphology and semantics. For instance, the two BEReshiT models propose the word מים[34] in place of ארץ[35] in the following context: פנית לפניה ותשרש שרשיה ותמלא ארץ (Ps 80:10)[36]. Here, only the GPT model makes the correct prediction, whilst both BEREL and Dicta propose merely punctuation.

Complex verb forms. Morphologically complex words, such as conjugated verbs in less frequently used tenses, are challenging for all models. For instance, instead of the verb יחפרו[37], the

[30]“Any of the Israelites or of the aliens who reside in Israel who give any of their offspring to Molech shall be put to death; the people of the land shall stone them to death.”

[31]Biblical references follow the New Revised Standard Version (NRSV) versification. English translations are taken from the NRSV.

[32]but, interestingly, not by GPT, which predicts מזרעו (mizaro “from his seed”)

[33]kol “all, everyone”

[34]mayim “water”

[35]ereṣ “land”

[36]“You cleared the ground for it; it took deep root and filled the land.”

[37]yaḥperu “dig-FUT.3PL.M”; יחפרו בעמק וישיש בכח יצא לקראת נשק “It paws violently, exults mightily; it goes out to meet the weapons.” (Job 39:21)

models GPT and DictaBERT propose other POS (respectively, the noun גיבור[38] and the noun/adjective הקרב[39]), whilst BEREL, mBERT and BEREshiT select verbs with more common meanings and conjugations (respectively, בא[40], יצא[41] and וישב[42]). Occasionally, in cases of complex verb forms as gold labels, the two BEReshiT models make different predictions. For example, the word כפופים (kefufim "bend.down-PTCP.M.PL"[43]) is predicted as a fragmentary plural suffix by the BEReshiT model with three trainable layers and as אפים (apaim, "faces; noses") by the fully trainable one. In both cases, the morphological aspects of number and gender (plural masculine) are correct.

Infrequent words. Finally, we were interested in finding out how the different models behave in relation to rarely occurring words. Models that exhibit train-test contamination are likely to perform significantly better with such examples, as with the outputs of models BEREL and GPT in Table 5. In particular, we looked at hapax legomena (i.e. words that appear only a single time in a given text), a common phenomenon in the Hebrew Bible. Within the main evaluation dataset, we defined as hapax legomena masked words that appear only once in the full Hebrew Bible dataset[44]. For the Biblical DSS examples, we included words that do not reappear in the dataset and appear up to once in the full Hebrew Bible dataset[45]. In the case of extra-Biblical DSS examples, we counted only words that do not appear in the Hebrew Bible or elsewhere in the dataset. However, it is possible that words with the same root but different morphological features are present within the concerned data.

In the case of the main evaluation dataset, out of the 10 discovered hapax legomena, BEREL and GPT each made two correct predictions, whilst the two BEReshiT models and DictaBERT all guessed correctly solely the word מסיג (masig "remove/shift-PTCP.M.SG.SUBST")[46]. mBERT made no correct predictions. BEREL and GPT did significantly worse in relation to the DSS dataset, once again demonstrating the effects of train-test contamination. 13 Biblical and 8 extra-Biblical hapax legomena were found in the DSS dataset. No model made a correct prediction for an extra-Biblical hapax, and BEREL made a single correct prediction for a Biblical hapax: ויושביה (ve-yošveha "and-inhabit-PTCP.M.PL.CSTR+3FS")[47]. It is worth noting, however, that the word appears 7 times in the Hebrew Bible dataset without the prefixed conjunction and is therefore, strictly speaking, not a hapax.

6. Discussion

The BEReshiT model demonstrates high results on a specially crafted dataset that emphasises the characteristics of Ancient Hebrew and its differences with the language's Modern counterpart. GPT-4o and BEREL surpass the model in performance, but their behaviour on alternative data proves that the root cause for their high results is train-test contamination. The multilingual model mBERT performs very poorly, pointing at the higher suitability of a language-specific BERT-based model given the Hebrew language's resourcedness and characteristics.

Starting from the model with 11 frozen layers, performance increased until reaching its peak for the model with 9 frozen layers (best overall result as well as best result for the "by category" portion of our main evaluation dataset). Then, the results started deteriorating gradually, whilst seeing another increase for 4 and 3 frozen layers. Traditionally, BERT's upper layers have been associated with syntax-related information (Tenney et al., 2019; Hewitt and Manning, 2019). Our models' increased performance when these layers are fine-tuned speaks of efficient re-learning of syntax, notably captured by the subset "by category". Interestingly, the subset "by phenomenon" exhibited a differing trend, reaching a peak in results when earlier layers were also trained, possibly proving a higher relevance for the model's knowledge of vocabulary and semantics, language aspects that have been associated with these layers.

7. Conclusion and Future Work

We presented the Ancient Hebrew model BEReshiT, which was achieved through the fine-tuning of the state-of-the-art Modern Hebrew model, DictaBERT. When evaluated on elaborately

[38] gibbor "hero, warrior"

[39] ha-qerav/ha-qarov "the battle; the nearby"

[40] ba "come-PST.3SG.M"

[41] yaṣa "go.out-PST.3SG.M"

[42] vayešev "sit-PST.3SG.M"; both BEReshiT models make this prediction

[43] יהוה פקח עורים יהוה זקף כפופים יהוה אהב צדיקים "the Lord opens the eyes of the blind. The Lord lifts up those who are bowed down; the Lord loves the righteous." (Ps. 146:8)

[44] Words that appear twice due to the doubling of part of the dataset were not counted as hapax legomena, although they may be such within the Hebrew Bible, as the BEReshiT models have already encountered them during training.

[45] as the data is the same with possibly different orthography

[46] ארור מסיג גבול רעהו ואמר כל העם אמן " 'Cursed be anyone who moves a neighbor's boundary marker.' All the people shall say, 'Amen!' " (Deut 27:17

[47] ויושביה כמו כן ימותון וישועתי לעולם תהיה וצדקתי לוא תחת "and those who live on it will die like gnats,but my salvation will be forever, and my deliverance will never be ended." (1QISAA: 42 line 21, part of Isa 51:6)

constructed tasks, BEReshiT exhibits better performance than its ancestor model as well as a selection of other relevant Hebrew and multilingual models (when train-test contaminated models are disregarded). A closer qualitative look shows that even when wrong, the model's predictions tend to be sound in terms of morphology and semantics. The highest performing BEReshiT model is the one with 9 frozen layers. We are releasing this model and BEReshiT-morph, a strong submodel trained for morphological tagging of Ancient Hebrew text.

The proposed BEReshiT model would benefit from further evaluation on different tasks and possibly different data. The latter would be tricky to realise due to the scarcity of Ancient Hebrew data. Possible directions may include the use of letters and inscriptions from the same historical period, a selection of which is present through the Qumran-Digital repository [48] or of synthetic data (possibly derived from Modern Hebrew based on hand-crafted rules or style-transfer methods). Our future plans also include the application of BEReshiT within a pipeline of stemmatological analysis of Ancient Hebrew text that uses morphological tagging in the definition of textual variants.

8. Ethics Statement

The datasets used for model training and evaluation are based on publicly available data.

9. Limitations

The exhibited results are highly dependent on the utilised evaluation datasets. Different data as well as different task settings, such as masking at the subword level (e.g. of isolated suffixed pronouns), could have decreased the challenge. The definition of hapax legomena would have also differed given alternative tokenisation.

The comparison of our morphological submodel and DictaBERT-morph also comes with limitations: the evaluation dataset was reduced to identically tokenised verses only; also, the two models retrieve different morphological information. Although we have mapped the represented morphological features in an exhaustive way, the logic of label attribution is partly subjective and may differ between the models.

10. Bibliographical References

1955–2009. *Discoveries in the Judean Desert*. Clarendon Press, Oxford.

David Bamman and Patrick J. Burns. 2020. Latin bert: A contextual language model for classical philology.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186.

Simon Gabay, Pedro Ortiz Suarez, Alexandre Bartz, Alix Chagué, Rachel Bawden, Philippe Gambette, and Benoît Sagot. 2022. From freem to d'alembert: a large corpus and a language model for early modern french.

Ayaka Harigai, Oshitaka Toyama, Mitsutoshi Nagano, Mirei Abe, Masahiro Kawabata, Li Li, Jin Yamamura, and Kei Takase. 2025. Response accuracy of gpt-4 across languages: insights from diagnostic radiology questions. *Japanese Journal of Radiology*.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Aaron D. Hornkohl. 2014. *Ancient Hebrew Periodization and the Language of the Book of Jeremiah: The Case for a Sixth-Century Date of Composition*, volume 74 of *Studies in Semitic Languages and Linguistics*. Brill, Leiden.

Avi Hurwitz. 2014. *A Concise Lexicon of Late Biblical Hebrew: Linguistic Innovations in the Writings of the Second Temple Period*. Vetus Testamentum Supplements. Brill, Leiden.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the*

[48] https://www.qumran-digital.org/

57th Annual Meeting of the Association for Computational Linguistics, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Pratik M. Joshi, Sebastin Santy, Ameya Godbole, Kalika Bali, and Chandra Sekhar Mukherjee. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293. Association for Computational Linguistics.

Geoffrey Khan, Shmuel Bolozky, Steven E. Fassberg, Gary A. Rendsburg, Aaron D. Rubin, Ora R. Schwarzwald, and Tamar Zewi, editors. 2013. *Encyclopedia of Hebrew Language and Linguistics*, 1 edition, volume 1-4. Brill, Leiden.

Piroska Lendvai, Uwe Reichel, Anna Jouravel, Achim Rabus, and Elena Renje. 2023. Domain-adapting BERT for attributing manuscript, century and region in pre-Modern Slavic texts. In *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, pages 15–21, Singapore. Association for Computational Linguistics.

Josh McGiff and Nikola S. Nikolov. 2025. Overcoming data scarcity in generative language modelling for low-resource languages: A systematic review. *arXiv preprint*.

Ricardo Muñoz Sánchez, David Alfter, Simon Dobnik, Maria Irena Szawerna, and Elena Volodina. 2024. Jingle BERT, jingle BERT, frozen all the way: Freezing layers to identify CEFR levels of second language learners using BERT. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 137–152, Rennes, France. LiU Electronic Press.

Iglika Nikolova-Stoupak, Maxime Amblard, Sophie Robert-Hayek, and Frédérique Rey. 2025. A classifier of word-level variants in witnesses of biblical Hebrew manuscripts. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21313–21329, Vienna, Austria. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4o system card. https://openai.com/index/gpt-4o-system-card/.

Robert Rezetko and Ian Young. 2014. *Historical Linguistics and Biblical Hebrew: Steps toward an Integrated Approach*, volume 9 of *Ancient Near East Monographs*. SBL Press, Atlanta.

Stefan Schweter, Luisa März, Katharina Schmid, and Erion Çano. 2022. hmbert: Historical multilingual language models for named entity recognition.

Avi Shmidman, Joshua Guedalia, Shaltiel Shmidman, Cheyn Shmuel Shmidman, Eli Handel, and Moshe Koppel. 2022. Introducing berel: Bert embeddings for rabbinic-encoded language.

Avi Shmidman, Ometz Shmidman, Hillel Gershuni, and Moshe Koppel. 2024a. MsBERT: A new model for the reconstruction of lacunae in Hebrew manuscripts. In *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)*, pages 13–18, Hybrid in Bangkok, Thailand and online. Association for Computational Linguistics.

Shaltiel Shmidman, Avi Shmidman, Amir DN Cohen, and Moshe Koppel. 2024b. Adapting llms to hebrew: Unveiling dictalm 2.0 with enhanced vocabulary and instruction capabilities.

Shaltiel Shmidman, Avi Shmidman, and Moshe Koppel. 2023. Dictabert: A state-of-the-art BERT suite for modern Hebrew.

Elad Shvartz, Leah Attal, Omri Zur, Zaki Nujeidat, Gilad Plopsky, and Daniel Bahir. 2025. Bilingual comparison of the performance of gpt-4o and gpt-4 on multilingual medical questions. *European Journal of Radiology*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Wenhao Zhu, Hongyu Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*.

A. Detailed Evaluation Results of the BEReshiT Model

A.1. Evaluation of dataset “Syntactic” by category

	overall					late					legal					narrative					poetry					prophetic					wisdom				
	acc	f1	pr	r	prob.gold	acc	f1	pr	r	prob.gold	acc	f1	pr	r	prob.gold	acc	f1	pr	r	prob.gold	acc	f1	pr	r	prob.gold	acc	f1	pr	r	prob.gold	acc	f1	pr	r	prob.gold
GPT-4o	**0.58**	**0.41**	**0.41**	**0.41**	–	0.47	0.30	0.30	0.30	–	0.47	0.33	0.33	0.33	–	**0.70**	**0.54**	**0.56**	**0.53**	–	**0.73**	**0.58**	**0.58**	**0.58**	–	**0.60**	**0.43**	**0.43**	**0.43**	–	**0.30**	**0.18**	**0.18**	**0.18**	–
mBERT	0.04	0.05	0.03	0.02	0.05	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.02	0.07	0.04	0.05	0.04	0.07	0.00	0.00	0.00	0.00	0.07	0.13	0.07	0.07	0.07	0.07	0.00	0.00	0.00	0.00	0.00
DictaBERT	0.33	0.19	0.20	0.20	0.25	0.26	0.15	0.15	0.17	0.14	0.40	0.26	0.26	0.26	0.30	0.53	0.33	0.34	0.34	0.40	0.27	0.17	0.17	0.17	0.25	0.20	0.12	0.12	0.12	0.19	0.00	0.00	0.00	0.00	0.00
BEREL	0.53	0.36	0.35	0.37	**0.49**	**0.53**	**0.37**	**0.36**	**0.38**	**0.42**	**0.60**	**0.47**	**0.47**	**0.47**	**0.62**	0.63	0.43	0.42	0.43	**0.59**	0.67	0.56	0.56	0.56	**0.66**	0.33	0.20	0.20	0.20	**0.28**	0.20	0.12	0.12	0.12	**0.20**
BEReshiT (f.l.:1-11)	0.36	0.21	0.21	0.22	0.26	0.20	0.11	0.10	0.12	0.15	0.47	0.32	0.32	0.32	0.33	0.53	0.33	0.34	0.33	0.41	0.33	0.20	0.20	0.20	0.23	0.33	0.20	0.20	0.20	0.18	0.00	0.00	0.00	0.00	0.00
BEReshiT (f.l.:1-10)	0.37	0.21	0.20	0.22	0.29	0.33	0.19	0.17	0.22	0.23	0.47	0.32	0.32	0.32	0.43	0.57	0.35	0.35	0.35	0.44	0.20	0.12	0.12	0.12	0.20	0.33	0.19	0.19	0.21	0.18	0.00	0.00	0.00	0.00	0.00
BEReshiT (f.l.:1-9)	0.40	0.22	0.22	0.24	0.29	0.33	0.19	0.17	0.22	0.23	0.53	0.38	0.38	0.38	0.44	0.6	0.38	0.38	0.38	0.43	0.27	0.15	0.15	0.15	0.20	0.33	0.20	0.20	0.23	0.17	0.00	0.00	0.00	0.00	0.00
BEReshiT (f.l.:1-8)	0.39	0.22	0.22	0.23	0.28	0.33	0.19	0.17	0.22	0.22	0.53	0.38	0.38	0.38	0.43	0.57	0.36	0.36	0.36	0.43	0.27	0.16	0.16	0.16	0.19	0.33	0.20	0.20	0.23	0.18	0.00	0.00	0.00	0.00	0.00
BEReshiT (f.l.:1-7)	0.38	0.22	0.22	0.24	0.28	0.40	0.25	0.24	0.26	0.27	0.40	0.26	0.26	0.26	0.37	0.50	0.30	0.31	0.30	0.40	0.27	0.15	0.15	0.15	0.19	0.40	0.26	0.25	0.27	0.23	0.10	0.06	0.06	0.06	0.04
BEReshiT (f.l.:1-6)	0.35	0.20	0.21	0.22	0.26	0.20	0.12	0.12	0.12	0.17	0.40	0.26	0.26	0.26	0.36	0.53	0.34	0.35	0.34	0.41	0.27	0.15	0.15	0.15	0.18	0.40	0.26	0.25	0.29	0.17	0.00	0.00	0.00	0.00	0.00
BEReshiT (f.l.:1-5)	0.33	0.18	0.18	0.19	0.24	0.20	0.11	0.10	0.12	0.19	0.47	0.32	0.32	0.32	0.32	0.53	0.33	0.34	0.33	0.39	0.20	0.12	0.12	0.12	0.17	0.27	0.15	0.14	0.17	0.17	0.00	0.00	0.00	0.00	0.00
BEReshiT (f.l.:1-4)	0.34	0.19	0.19	0.20	0.26	0.20	0.11	0.11	0.11	0.17	0.47	0.33	0.33	0.33	0.41	0.57	0.38	0.39	0.38	0.41	0.13	0.07	0.07	0.07	0.13	0.33	0.19	0.19	0.21	0.21	0.00	0.00	0.00	0.00	0.00
BEReshiT (f.l.:1-3)	0.34	0.19	0.20	0.20	0.27	0.27	0.15	0.14	0.16	0.21	0.40	0.26	0.26	0.26	0.35	0.53	0.35	0.36	0.36	0.43	0.20	0.12	0.12	0.12	0.15	0.33	0.20	0.20	0.22	0.25	0.00	0.00	0.00	0.00	0.00
BEReshiT (f.l.:1-2)	0.33	0.18	0.18	0.19	0.24	0.27	0.16	0.16	0.16	0.18	0.47	0.32	0.32	0.32	0.38	0.50	0.31	0.32	0.30	0.38	0.13	0.08	0.08	0.08	0.11	0.33	0.20	0.19	0.22	0.16	0.00	0.00	0.00	0.00	0.00
BEReshiT (f.l.:1)	0.35	0.18	0.18	0.20	0.26	0.27	0.15	0.14	0.16	0.22	0.53	0.38	0.38	0.38	0.43	0.57	0.38	0.38	0.39	0.39	0.20	0.12	0.12	0.12	0.16	0.20	0.10	0.10	0.12	0.15	0.00	0.00	0.00	0.00	0.00
BEReshiT (no f.l.)	0.36	0.20	0.20	0.21	0.26	0.33	0.20	0.20	0.20	0.20	0.40	0.26	0.26	0.26	0.34	0.53	0.33	0.34	0.33	0.42	0.20	0.12	0.12	0.12	0.16	0.40	0.26	0.25	0.27	0.19	0.00	0.00	0.00	0.00	0.00

Table 6: Evaluation of dataset “Syntactic” by genre. Metrics: accuracy, F1 score (macro), precision (macro), recall (macro), average probability of gold label. The best global results per model and per subset are indicated in **bold** and the best results when models with likely train-test contamination are disregarded are underlined Numbers are rounded to the second digit after the decimal point. f.l.: frozen layer(s)

A.2. Evaluation of dataset “Syntactic” by phenomenon

	overall					inf.abs.fin.					part.					poss.					vso					wayyiqtol				
	acc	f1	pr	r	prob.gold	acc	f1	pr	r	prob.gold	acc	f1	pr	r	prob.gold	acc	f1	pr	r	prob.gold	acc	f1	pr	r	prob.gold	acc	f1	pr	r	prob.gold
GPT-4o	0.50	0.33	0.33	0.33	–	**0.55**	**0.39**	**0.39**	**0.39**	–	0.35	0.21	0.21	0.21	–	**0.55**	**0.38**	**0.38**	**0.38**	–	0.55	0.38	0.38	0.38	–	0.50	0.34	0.34	0.34	–
mBERT	0.06	0.03	0.03	0.04	0.07	0.10	0.05	0.06	0.05	0.06	0.10	0.05	0.06	0.05	0.01	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.01	0.10	0.05	0.05	0.06	0.13
DictaBERT	0.22	0.13	0.13	0.14	0.16	0.40	0.25	0.25	0.25	0.26	0.15	0.08	0.08	0.08	0.10	0.10	0.06	0.06	0.06	0.03	0.15	0.10	0.10	0.10	0.12	0.30	0.18	0.18	0.18	0.29
BEREL	**0.54**	**0.39**	**0.39**	**0.39**	**0.49**	**0.55**	**0.39**	**0.39**	**0.39**	**0.52**	**0.55**	**0.40**	**0.41**	**0.39**	**0.46**	0.40	0.25	0.25	0.25	**0.45**	**0.65**	**0.50**	**0.50**	**0.50**	**0.62**	**0.55**	**0.38**	**0.38**	**0.38**	**0.51**
BEReshiT (f.l.:1-11)	0.27	0.15	0.14	0.16	0.17	0.40	0.26	0.25	0.26	0.25	0.25	0.14	0.13	0.15	0.12	0.15	0.09	0.09	0.09	0.04	0.2	0.10	0.10	0.11	0.12	0.35	0.23	0.23	0.23	0.33
BEReshiT (f.l.:1-10)	0.30	0.16	0.16	0.17	0.22	0.35	0.21	0.20	0.22	0.30	0.30	0.17	0.16	0.18	0.16	0.15	0.08	0.08	0.08	0.07	0.30	0.18	0.18	0.18	0.19	0.40	0.25	0.25	0.25	0.36
BEReshiT (f.l.:1-9)	0.32	0.18	0.17	0.19	0.21	0.35	0.21	0.20	0.22	0.30	0.35	0.21	0.20	0.22	0.15	0.15	0.08	0.07	0.09	0.08	0.35	0.22	0.22	0.22	0.19	0.40	0.26	0.26	0.26	0.35
BEReshiT (f.l.:1-8)	0.31	0.17	0.17	0.19	0.21	0.40	0.25	0.24	0.26	0.30	0.25	0.13	0.13	0.15	0.14	0.15	0.08	0.07	0.09	0.09	0.35	0.22	0.22	0.22	0.19	0.40	0.25	0.25	0.25	0.34
BEReshiT (f.l.:1-7)	0.28	0.15	0.14	0.17	0.20	0.35	0.21	0.20	0.22	0.26	0.25	0.16	0.16	0.16	0.16	0.15	0.08	0.08	0.08	0.05	0.25	0.15	0.15	0.15	0.19	0.40	0.26	0.26	0.28	0.35
BEReshiT (f.l.:1-6)	0.25	0.13	0.12	0.14	0.19	0.35	0.21	0.21	0.22	0.26	0.20	0.10	0.10	0.11	0.14	0.05	0.03	0.03	0.03	0.04	0.25	0.14	0.14	0.15	0.15	0.40	0.26	0.26	0.26	0.35
BEReshiT (f.l.:1-5)	0.25	0.13	0.12	0.14	0.18	0.35	0.21	0.21	0.22	0.26	0.25	0.14	0.14	0.14	0.15	0.10	0.05	0.05	0.05	0.05	0.15	0.08	0.07	0.09	0.13	0.40	0.26	0.26	0.26	0.34
BEReshiT (f.l.:1-4)	0.33	0.19	0.18	0.20	0.23	0.45	0.31	0.30	0.33	0.32	0.30	0.17	0.16	0.18	0.17	0.25	0.15	0.15	0.15	0.09	0.30	0.18	0.18	0.18	0.20	0.35	0.22	0.21	0.23	0.35
BEReshiT (f.l.:1-3)	0.33	0.19	0.18	0.21	0.23	0.40	0.26	0.25	0.28	0.32	0.30	0.19	0.18	0.20	0.15	0.20	0.11	0.10	0.12	0.09	0.35	0.23	0.23	0.23	0.22	0.40	0.25	0.24	0.26	0.35
BEReshiT (f.l.:1-2)	0.30	0.17	0.16	0.18	0.19	0.35	0.21	0.20	0.22	0.26	0.30	0.17	0.17	0.17	0.14	0.15	0.08	0.08	0.08	0.06	0.20	0.11	0.11	0.12	0.14	0.50	0.34	0.34	0.34	0.36
BEReshiT (f.l.:1)	0.30	0.16	0.16	0.19	0.20	0.30	0.17	0.17	0.18	0.23	0.25	0.14	0.14	0.14	0.16	0.15	0.08	0.07	0.09	0.09	0.30	0.18	0.18	0.18	0.16	0.50	0.35	0.34	0.36	0.36
BEReshiT (no f.l.)	0.27	0.14	0.13	0.16	0.19	0.35	0.21	0.21	0.22	0.26	0.20	0.11	0.11	0.11	0.15	0.10	0.05	0.04	0.06	0.04	0.30	0.19	0.19	0.19	0.15	0.40	0.26	0.25	0.27	0.35

Table 7: Evaluation of dataset “Syntactic” by phenomenon. Metrics: accuracy, F1 score (macro), precision (macro), recall (macro), average probability of gold label. The best global results per model and per subset are indicated in **bold** and the best results when models with likely train-test contamination are disregarded are underlined Numbers are rounded to the second digit after the decimal point. f.l.: frozen layer(s)

B. Morphological Model Evaluation

B.1. BEReshiT-morph: Detailed Evaluation Results

	Morphological features																																							
	POS					sub-POS					stem					conj.					person					gender					number					state					suffix type				
	acc	pr.	r.	f1	cov.	acc	pr.	r.	f1	cov.	acc	pr.	r.	f1	cov.	acc	pr.	r.	f1	cov.	acc	pr.	r.	f1	cov.	acc	pr.	r.	f1	cov.	acc	pr.	r.	f1	cov.	acc	pr.	r.	f1	cov.	acc	pr.	r.	f1	cov.
full	.98	.97	.97	.97	.84	.98	.89	.88	.88	.38	.93	.80	.68	.72	.12	.92	.82	.77	.79	.11	.98	.74	.74	.74	.16	.98	.73	.73	.73	.36	.99	.73	.72	.72	.36	.95	.65	.63	.64	.19	.98	.64	.61	.63	.07
OSHB	.98	.97	.97	.97	.86	.98	.89	.89	.89	.38	.93	.69	.59	.63	.12	.92	.82	.77	.80	.11	.99	.74	.74	.74	.17	.98	.73	.73	.73	.36	.99	.73	.73	.73	.36	.96	.65	.64	.65	.19	.98	.65	.61	.63	.08
DSS	.96	.94	.94	.94	.77	.96	.78	.76	.76	.35	.89	.74	.72	.72	.13	.89	.56	.52	.54	.13	.97	.72	.72	.72	.11	.97	.72	.71	.71	.35	.98	.69	.65	.67	.35	.92	.63	.61	.62	.18	.97	.62	.62	.62	.02

Table 8: Detailed evaluation results for model BEReshiT-morph (full dataset, OSHB subset, DSS subset). The following metrics are considered: accuracy, precision (macro), recall (macro), F1-score (macro), coverage (the fraction of evaluated token positions where the gold label is applicable). Numbers are rounded to the second digit after the decimal point.

B.2. BEReshiT-morph: Confusion Matrices

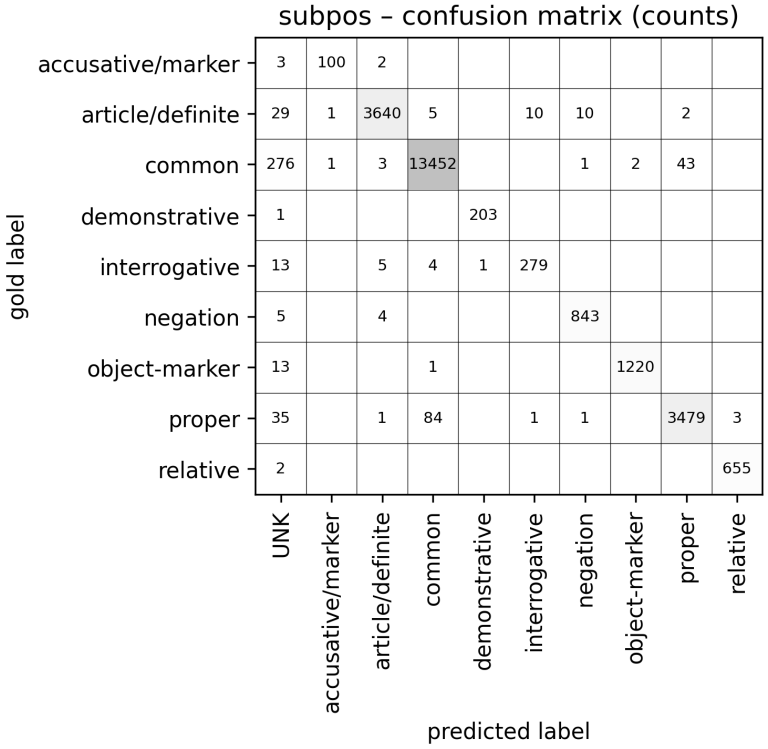


Figure 2: Confusion matrix for the feature “sub-POS”

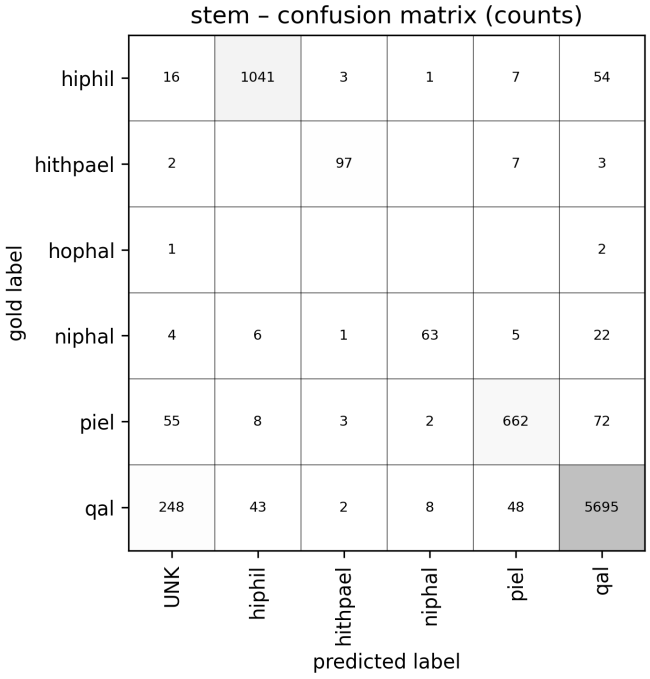


Figure 3: Confusion matrix for the feature “stem”

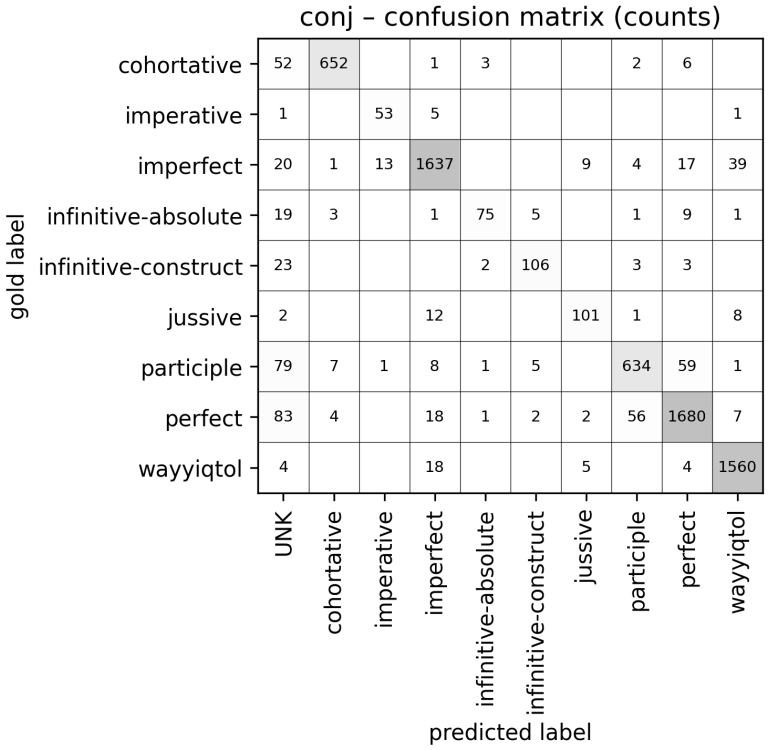


Figure 4: Confusion matrix for the feature “conjugation”

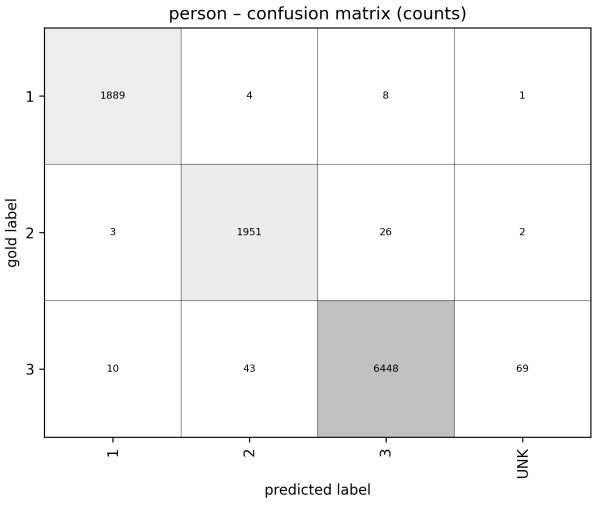


Figure 5: Confusion matrix for the feature “person”

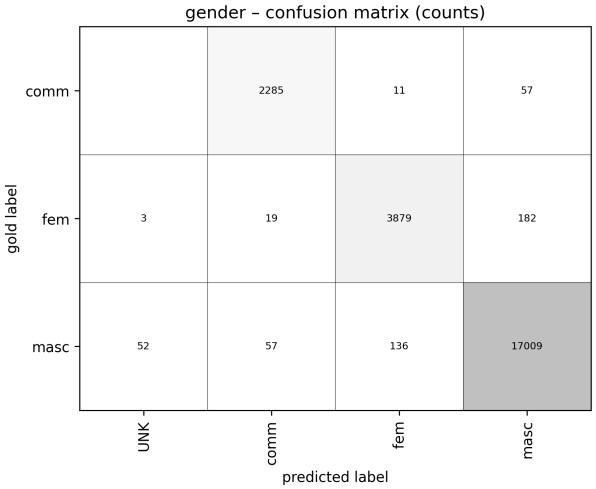


Figure 6: Confusion matrix for the feature “gender”

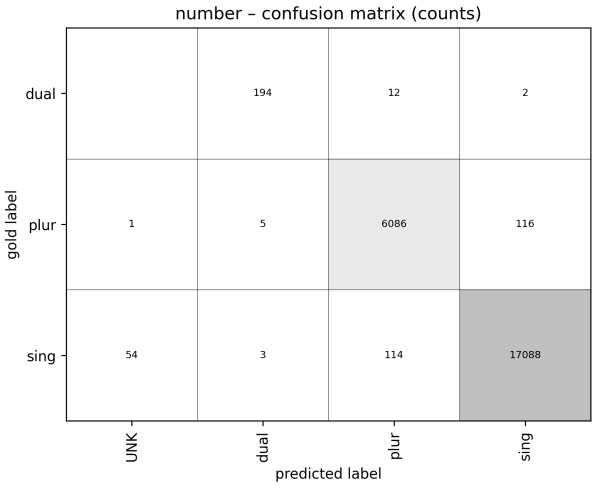


Figure 7: Confusion matrix for the feature “number”

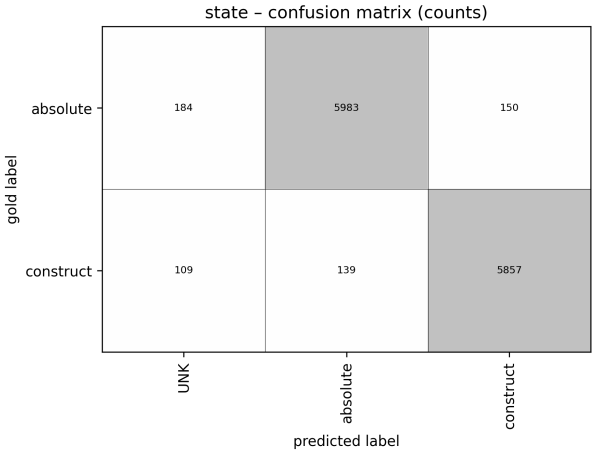


Figure 8: Confusion matrix for the feature “state”

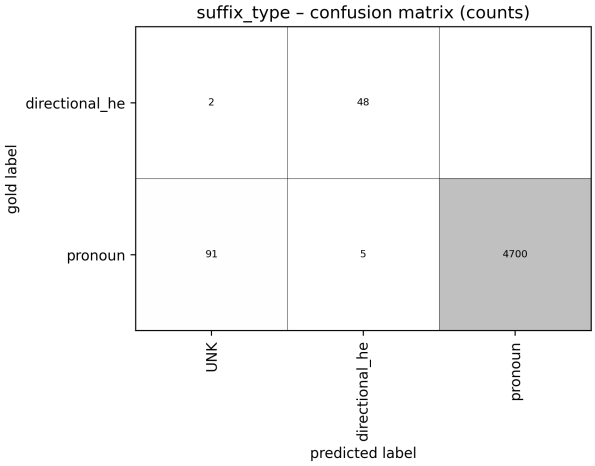


Figure 9: Confusion matrix for the feature “suffix type”

B.3. DictaBERT-morph: Confusion Matrix for POS

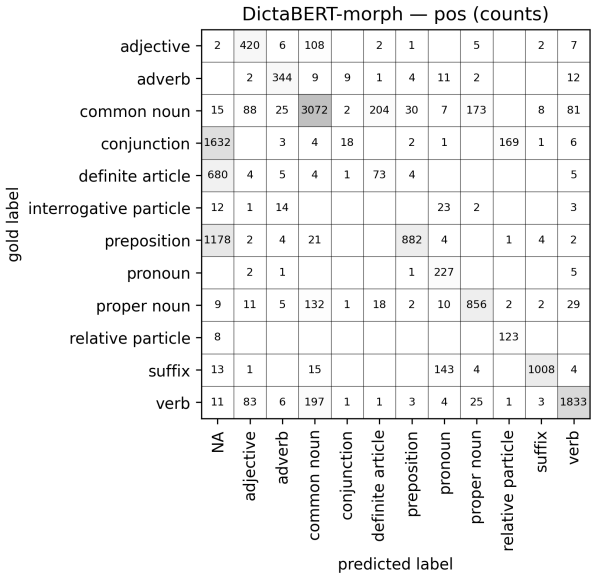


Figure 10: Confusion matrix for the feature “POS” for the model DictaBERT-morph.

C. Features and Labels as Present in the BEReshiT-morph and DictaBERT-morph Models

Feature	Label	Example	Ref.
POS	verb	יאמר	2 Sam 1:3
	noun	אזנים	Ps 40:7
	adjective	רקים	Prov 28:19
	adverb	שם	Gen 26:23
	pronoun	הוא	Num 9:13
	preposition	על	1 Kgs 6:35
	conjunction	ו	2 Chr 24:10
	particle	לא	Ps 119:157
	suffix	י	Ps 35:3
sub-POS	common	ימים	Judg 14:14
	proper	זבולן	Num 1:30
	article / definite	ה	Ps 24:10
	demonstrative	הנה	Esth 6:5
	interrogative	ה	2 Sam 3:33
	relative	אשר	Deut 23:11
	negation	לא	Lam 1:10
	accusative / marker	רק	2 Chr 28:10
	object-marker	את	Judg 19:29
conjugation	perfect	דבר	Jer 27:13
	imperfect	יהיו	2 Chr 12:8
	wayyiqtol	יערכו	2 Chr 14:9
	participle	באים	Jer 31:27
	imperative	נאבלה	Dan 1:12
	infinitive-construct	כתוב	Ps 149:9
	infinitive-absolute	רצח	Hos 4:2
	jussive	יחשב	2 Sam 19:20
	cohortive	המרות	Job 17:2
stem	qal	אמר	Exod 1:22
	niphal	יבא	Gen 2:22
	piel	ידברו	Jer 9:4
	hiphil	יזעק	Judg 4:10
	hophal	יכלכל	2 Chr 2:5
	hithpael	התרגז	Isa 37:29
person	1	אקח	Gen 14:23
	2	שמעת	Lam 3:61
	3	הוא	Jer 25:12
gender	masc	יבאו	2 Sam 20:14
	fem	ה-	Ezek 22:22
	common	אני	Job 33:6
number	sing	שקל	Exod 38:29
	pl	יחשבו	Ps 41:8
	dual	שני	Jer 46:12
suffix type	directional *he*	ה	Num 10:29
	pronoun	י	Eccl 2:3

Table 9: Morphological features determinable by the model BEReshiT-morph. Examples are retrieved from the golden labels (some tokens are at subword level).

Feature	Label	Example	Ref.
POS	VERB	ימול	Gen 17:13
	NOUN	בני	1 Sam 9:2
	ADJ	מלאה	Num 7:62
	ADV	מאד	Jer 48:29
	PRON	זה	Jer 52:28
	PROPN	חור	1 Chr 4:1
	ADP	ב	2 Kgs 18:17
	ADP_PRON	חללו	Ezek 7:22
	DET	ה	Josh 8:5
	NUM		
	AUX	היתה	1 Chr 4:10
	CCONJ	ו	1 Sam 14:37
	SCONJ	כי	Isa 52:5
	INTJ	הו	Isa 3:5
	PUNCT		
	SYM		
	X		
tense	Past	שגיתי	Job 6:24
	Pres	משמרים	Jonah 2:9
	Fut	יאכל	Eccl 5:16
	Imp	רדה	Gen 46:3
person	1	עלי	Job 27:23
	2	הביאת	Isa 43:23
	3	היתה	1 Chr 4:10
	1, 2, 3	חצובים	Neh 9:25
gender	Masc	אחי	Deut 2:8
	Fem	היתה	Gen 17:16
	Fem,Masc	בני	Exod 30:19
number	Sing	ראיתי	1 Sam 25:25
	Plur	ידם	Lev 10:3
	Dual	שנתים	1 Kgs 22:52

Table 10: Morphological features determinable by the model DictaBERT-morph. Examples are taken from the model's predictions, where available and correct (some tokens are at subword level).

Automatic Detection of Metaphorical Expressions in Classical Japanese Using WLSP-Enhanced BERT

Hang Zhu[1], **Mitoki Ohara**[2], **Rei Kikuchi**[3], **Kanako Komiya**[1],
Masayuki Asahara[4], **Sachi Kato**[5]
[1]Tokyo University of Agriculture and Technology, Tokyo, Japan
[2]Chuo University, Tokyo, Japan
[3]National Institute of Technology, Nagano College, Nagano, Japan
[4]National Institute for Japanese Language and Linguistics, Tokyo, Japan
[5]Hokkaido University, Hokkaido, Japan
s259390y@st.go.tuat.ac.jp, 82303991araho@gmail.com, r_kikuchi@nagano-nct.ac.jp,
kkomiya@go.tuat.ac.jp, masayu-a@ninjal.ac.jp, katosachi@let.hokudai.ac.jp

Abstract

Metaphor detection is a fundamental task in natural language processing, yet research on historical languages remains limited. While progress has been made in modern Japanese metaphor detection, classical Japanese texts present unique challenges due to their distinct vocabulary, grammar, and metaphorical patterns. This paper addresses this gap by applying a BERT-based metaphor detection method enhanced with semantic classification information from the Word List by Semantic Principles (WLSP) to classical Japanese texts. We evaluate our approach on CHJ-Metaphor, a newly available corpus featuring metaphor annotations for three medieval Japanese works from the Corpus of Historical Japanese (CHJ). Our method achieves an F1-score of 82.18 through 5-fold cross-validation. Notably, qualitative analysis by domain experts reveals that our model successfully identifies genuine metaphors overlooked during manual annotation, demonstrating its potential as a tool for improving annotation quality in large-scale corpus construction. These results confirm the effectiveness of WLSP-enhanced approaches for metaphor detection in classical Japanese and suggest promising directions for applying similar techniques to other historical languages.

Keywords: Metaphor Detection, Classical Japanese, Historical Language Processing, BERT

1. Introduction

Metaphorical expressions play a crucial role in language understanding, and their automatic detection remains an important challenge in natural language processing (Group Pragglejaz, 2007). In this framework, a word is considered metaphorical when its meaning in context differs from its most basic, concrete, or historically earliest sense (Group Pragglejaz, 2007). Recent years have witnessed significant advances in metaphor detection research using Transformer-based models such as BERT (Choi et al., 2021; Li et al., 2023a), though most studies focus primarily on modern languages.

For Japanese, Kato et al. (2022) constructed BCCWJ-Metaphor based on the Balanced Corpus of Contemporary Written Japanese (Maekawa et al., 2014), providing systematic metaphor annotations following the Metaphor Identification Procedure (MIP) (Group Pragglejaz, 2007). Building upon this foundation, Zhu et al. (2025) leveraged semantic classification information from the Word List by Semantic Principles (National Institute for Japanese Language and Linguistics, 1964) to achieve an F1-score of 75.07 for modern Japanese metaphor detection.

However, automatic metaphor detection in classical Japanese texts remains largely unexplored. Classical Japanese differs substantially from modern Japanese in vocabulary and grammar, and metaphorical usage patterns may have evolved over time. The Corpus of Historical Japanese (National Institute for Japanese Language and Linguistics, 2023) provides a rich resource spanning texts from the Nara period (710-794) to the Meiji era (1868-1912), but metaphor detection for these historical texts presents unique challenges.

To address this research gap, we apply the WLSP-based BERT approach, previously validated on modern Japanese, to classical Japanese metaphor detection. We evaluate our method on CHJ-Metaphor, which provides metaphor annotations for classical Japanese texts from CHJ enhanced with WLSP semantic classifications (Asahara et al., 2022). Specifically, we conduct experiments on three currently annotated works from the medieval period (1185-1573, comprising the Kamakura and Muromachi eras): *Konjaku Monogatarishu* (Tales of Times Now Past), *Hojoki* (An Account of My Hut), and *Toraakira-bon Kyogen-shu* (Toraakira's Kyogen Collection).

Our contributions are threefold: (1) We demonstrate the first application of automatic metaphor detection to classical Japanese texts, achieving an F1-score of 82.18 through 5-fold cross-validation. (2) We provide a comprehensive analysis across works with varying characteristics, revealing that performance remains stable across metaphor rates

Proceedings of LT4HALA 2026 @ LREC 2026, pages 89–95.
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

ranging from 4.41% to 25.82%. (3) Through expert validation, we demonstrate that our model identifies genuine metaphors overlooked during manual annotation, suggesting its potential utility for improving annotation quality in large-scale corpus construction.

2. Related Work

2.1. Metaphor Detection Approaches

Metaphor detection approaches can be broadly categorized into three types: dictionary and rule-based methods (Dodge et al., 2015), statistical approaches (Klebanov et al., 2014), and deep learning-based methods (Mao et al., 2019). Recent advances have been dominated by Transformer-based models (Tong et al., 2021), with various architectural innovations proposed to capture metaphorical meaning.

MelBERT (Choi et al., 2021) builds upon RoBERTa (Liu et al., 2019) and compares target word embeddings with contextual embeddings to detect metaphors. FrameBERT (Li et al., 2023b) incorporates semantic frame information from FrameNet (Baker et al., 2024). BasicBERT (Li et al., 2023a) learns basic sense representations from non-metaphorical examples in the training data. Other approaches include MIss RoBERTa WiLDe (Babieno et al., 2022), which leverages Wiktionary definitions, and MisNet (Zhang and Liu, 2022), which enhances linguistic knowledge for metaphor detection.

Our approach differs from these methods by leveraging the systematic semantic classification structure of WLSP, providing theoretically grounded basic sense determination.

Recent work has also explored metaphor detection beyond English, including cross-lingual transfer approaches for low-resource languages (Hülsing and Schulte Im Walde, 2024), highlighting the broader challenge of extending metaphor detection methods across linguistic boundaries.

2.2. Japanese Historical Corpora and Metaphor Research

The Corpus of Historical Japanese (National Institute for Japanese Language and Linguistics, 2023) encompasses diverse textual genres spanning from the Nara period (710-794) to the Meiji era (1868-1912). Asahara et al. (2022) developed CHJ-WLSP by annotating CHJ texts with WLSP (National Institute for Japanese Language and Linguistics, 1964) semantic classification codes. Building upon CHJ-WLSP, Kikuchi et al. are constructing CHJ-Metaphor through manual metaphor annotation. This study uses the currently completed portions of the corpus under construction.[1]

3. Dataset

3.1. CHJ-Metaphor

CHJ-Metaphor is a metaphor expression dataset for classical Japanese, created by adding metaphor annotations to CHJ-WLSP (Asahara et al., 2022). The corpus is designed to include medieval period works across three genres: narratives (*Konjaku Monogatari-shu*, *Uji Shui Monogatari*, *Jikkinsho*), essays (*Hojoki*, *Tsurezuregusa*), and kyogen plays (*Toraakira-bon Kyogen-shu*).

Metaphor identification follows the same procedure as BCCWJ-Metaphor for modern Japanese, employing MIP and its extended version MIPVU (Steen et al., 2010), along with Nakamura's (Nakamura, 1977) metaphor type classification. MIP identifies metaphorical words by comparing their contextual meaning with their most basic (prototypical) sense: a word is metaphorical if these two meanings differ in a systematic way (Group Pragglejaz, 2007). MIPVU extends MIP to handle implicit metaphors, personification, and metaphorically used proper nouns. Nakamura's classification further distinguishes metaphorical expressions by how they are comprehended: Type A (simile, where a linguistic marker signals the comparison), Type B (combinatorial metaphors, arising from selectional restriction violations), and Type C (contextual metaphors, where metaphoricity emerges from broader discourse context).

The annotation covers six types of metaphorical expressions: 指標比喩 "simile" (explicit comparison using indicators such as ように "like" or 如し "as"), 換喩 "metonymy" (referring to something by mentioning an associated concept), 提喩 "synecdoche" (using a part to represent the whole or vice versa), 文脈比喩 "Discourse-level metaphors" (metaphorical meaning emerging from context), 結合比喩 "Word-level metaphors" (metaphorical meanings created by unconventional word combinations that violate selectional restrictions), and その他 "others" (including 枕詞 "makurakotoba," a conventional poetic technique in classical Japanese).

This study uses three currently completed works. As shown in Table 1, *Konjaku Monogatari-shu* contains the largest amount of data with 79,892 tokens, while *Hojoki* (2,668 tokens) and *Toraakira-bon Kyogen-shu* (2,542 tokens) are substantially smaller. The corpus contains 85,102 tokens in total.

Table 2 presents the metaphor rate (proportion of metaphorical tokens) for each work, which

[1] The CHJ-Metaphor corpus will be made publicly available after the annotation is completed.

Work	Tokens
Konjaku Monogatari-shu	79,892
Hojoki	2,668
Toraakira-bon Kyogen-shu	2,542
Total	85,102

Table 1: Dataset Size

Work	Rate (%)
Konjaku Monogatari-shu	4.41
Hojoki	25.82
Toraakira-bon Kyogen-shu	12.79
Overall	5.33

Table 2: Metaphor Rate

varies substantially across the three texts. *Konjaku Monogatari-shu*, despite having the most data, exhibits the lowest metaphor rate at 4.41%. In contrast, *Hojoki* shows the highest metaphor rate at 25.82%, nearly six times higher. *Toraakira-bon Kyogen-shu* falls in between with a metaphor rate of 12.79%. The overall metaphor rate across all works is 5.33%.

The metaphor rate varies across literary works. The influence of genre and style on metaphor rate may be analyzed further as annotation of additional works progresses.

4. Methodology

We apply the WLSP-based BERT approach, previously validated on modern Japanese (Zhu et al., 2025), to classical Japanese metaphor detection. This section describes the core components of our method.

4.1. Prototypical Sense Determination

Determining the prototypical sense of polysemous words is a fundamental challenge in applying MIP. Our approach leverages the prototypical sense information annotated to WLSP by Yamazaki et al. (2015); Yamazaki and Kashino (2017).

Notably, the prototypical sense used in this study is based on modern Japanese rather than classical Japanese, as WLSP was originally developed for contemporary language. This methodological choice reflects the practical consideration that annotators, as contemporary speakers, inevitably interpret classical texts through a modern lens when identifying metaphorical expressions.

Yamazaki et al. assigned confidence scores on a five-level scale to each sense of polysemous words in WLSP, indicating the degree to which each sense represents the prototypical sense. For a target word w (lemma l) appearing in context with classification code c_{context}, we determine the prototypical sense as follows. We first identify the set C_{proto} containing all classification codes with the highest confidence score for lemma l in WLSP. If $c_{\text{context}} \in C_{\text{proto}}$, we select it as the prototypical sense classification code c_{proto}. Otherwise, we randomly select one code from C_{proto}. If no candidates exist in WLSP, we set $c_{\text{proto}} = c_{\text{context}}$.

4.2. Prototypical Example Retrieval

Using the determined prototypical sense classification code c_{proto}, we retrieve prototypical usage examples from BCCWJ-WLSP-auto (Asada et al., 2024). We extract all instances where both the lemma l and classification code c_{proto} match. If multiple examples exist, we randomly select one. If no matching example is found, we use the lemma l itself as the prototypical example.

4.3. BERT-based Metaphor Detection Model

For a sentence $S = \{w_1, \ldots, w_n\}$ containing the target word w_t, we first augment it with WLSP semantic classification information:

$$S' = (w_1, \ldots, w_n, [\text{SEP}], f_1, \ldots, f_k) \quad (1)$$

where $\{f_i\}$ represents WLSP semantic features appended after `[SEP]`. WLSP organizes Japanese vocabulary into a four-level hierarchical taxonomy: *division* (類), *section* (部門), *medium category* (中項目), and *classification item* (分類項目), covering approximately 96,000 headwords. We use [MASK] tokens when classification information is unavailable.

We employ a multi-layer token type ID system to distinguish different information types within the augmented sequence. Each token is assigned a role: target word, local context (words within punctuation boundaries around the target), semantic features (WLSP classification information), or background (other tokens).

We encode the target sentence S' and the prototypical example $P = u_{\text{proto}}$ using Japanese BERT:

$$\mathbf{v}_{S',1}, \ldots, \mathbf{v}_{S',n} = \text{BERT}(S') \quad (2)$$

$$\mathbf{v}_{P,1}, \ldots, \mathbf{v}_{P,m} = \text{BERT}(P) \quad (3)$$

We then compute a vector $\mathbf{h}_{\text{MIP}}$ that captures the semantic relationship between the contextual meaning and the prototypical sense:

$$\mathbf{h}_{\text{MIP}} = f([\mathbf{v}_{S',t}; \mathbf{v}_{P,t'}]) \quad (4)$$

where $f(\cdot)$ is a linear transformation and $[\cdot;\cdot]$ denotes concatenation. The model is trained using cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1-y_i)\log(1-p_i)] \quad (5)$$

To address class imbalance (approximately 5% metaphorical tokens in the dataset), we apply weighted cross-entropy loss by assigning a higher weight to metaphorical instances, ensuring that the model learns to detect rare metaphorical patterns effectively.

4.4. Experimental Settings

We use bert-base-japanese-v3[2] as our base model, with a maximum sequence length of 512, a batch size of 16, a learning rate of 3e-5, and train it for 5 epochs. We evaluate performance using 5-fold cross-validation.

Evaluation metrics include F1-score, precision, and recall, all computed at the token level based on CHJ-WLSP morphological units (short units).

5. Experiments and Results

5.1. Overall Performance

Table 3 presents the overall results of 5-fold cross-validation on CHJ-Metaphor.

Model	F1	Prec.	Rec.
WLSP-enhanced	82.18	82.83	81.58

Table 3: Overall Performance

Our WLSP-enhanced BERT approach achieves an F1-score of 82.18%, precision of 82.83%, and recall of 81.58% on classical Japanese texts, demonstrating the effectiveness of applying methods validated on modern Japanese to classical texts.

5.2. Performance by Work

Tables 4 and 5 show a performance breakdown across the three works.

Konjaku Monogatari-shu achieves the highest accuracy (98.41%) with a precision of 82.55% and a recall of 81.21%. *Hojoki* shows a precision of 85.26% and a recall of 84.76% with an accuracy of 92.28%. *Toraakira-bon Kyogen-shu* demonstrates a precision of 80.00% and a recall of 78.76% with an accuracy of 94.77%.

[2] https://huggingface.co/tohoku-nlp/bert-base-japanese-v3

Work	Acc.
Konjaku Monogatari-shu	98.41
Hojoki	92.28
Toraakira-bon Kyogen-shu	94.77

Table 4: Accuracy by Work

Work	Prec.	Rec.
Konjaku Monogatari-shu	82.55	81.21
Hojoki	85.26	84.76
Toraakira-bon Kyogen-shu	80.00	78.76

Table 5: Precision and Recall by Work

All three works achieve over 78% precision and recall with well-balanced performance. *Konjaku Monogatari-shu* shows the highest accuracy, *Hojoki* achieves the most balanced precision-recall performance, and *Toraakira-bon Kyogen-shu* shows intermediate performance. Detailed analysis is provided in Section 6.

6. Discussion

6.1. Performance Differences Across Works

The performance differences across the three works can be attributed to their distinct characteristics in data size and metaphor rate.

Konjaku Monogatari-shu achieves strong performance with the largest data size (79,892 tokens), which contributes to learning stability. Despite the low metaphor rate (4.41%), the model maintains balanced precision-recall, demonstrating that the weighted loss function and abundant training data enable the model to learn metaphorical patterns effectively while handling class imbalance.

Hojoki achieves the most balanced precision-recall performance (85.26% precision, 84.76% recall). The high metaphor rate (25.82%) allows the model to learn metaphorical patterns more effectively, as metaphorical instances are well represented during training. The relatively lower accuracy (92.28%) reflects the increased proportion of metaphorical tokens, which reduces the relative contribution of non-metaphorical tokens to overall accuracy calculations.

Toraakira-bon Kyogen-shu shows a precision of 80.00% and a recall of 78.76%, reflecting its intermediate characteristics in both data size and metaphor rate.

Overall, the method demonstrates stable performance across metaphor rates ranging from 4.41% to 25.82%, confirming the effectiveness of the proposed approach for classical Japanese texts with

varying characteristics.

6.2. Comparison with Existing Approaches

Direct comparison with existing metaphor detection models such as MelBERT and FrameBERT is not feasible in this setting, as these models rely on language resources specific to modern languages (e.g., FrameNet, English training corpora) that have no equivalent for classical Japanese.

Regarding the contribution of WLSP semantic features, ablation experiments on modern Japanese metaphor detection (Zhu et al., 2025) demonstrate that the WLSP-enhanced model significantly outperforms a vanilla BERT baseline, confirming the effectiveness of the semantic augmentation. The present study applies this validated framework to classical Japanese, where constructing a comparable ablation baseline would require additional annotation effort beyond the scope of this work.

6.3. Differences from Modern Japanese

The strong performance on classical Japanese texts may be attributed to several linguistic features.

First, classical Japanese vocabulary is more limited compared to modern Japanese, potentially resulting in lower polysemy. Second, classical literary texts are characterized by formulaic expressions, and metaphorical patterns tend to be more regular. These factors may collectively contribute to the model's strong performance on classical texts.

6.4. Qualitative Evaluation of Detection Results

Following the experiments, the creators of CHJ-Metaphor conducted a detailed qualitative analysis of the detection results, examining instances classified as false positives by the model to determine whether they constituted genuine metaphors overlooked during the original annotation phase. The analysis reveals both strengths and limitations of the proposed method.

6.4.1. Successful Detection of Overlooked Metaphors

Some instances classified as false positives were revealed to be genuine metaphors that had been overlooked during manual annotation. The following examples illustrate cases where the model successfully detected such subtle metaphors.

In *Konjaku Monogatari-shu*, the expression 音 "sound" was identified as a metaphor by the model despite being missed during annotation:

> (1) 聖人は，誓ひを発して，音を挙て法花経を読奉る．(Classical Japanese)
> (*Hijiri wa, chikai wo okoshite, oto wo agete, hokkekyō wo yomi-tatematsuru*)
> 'A saint made a vow and, raising his voice, read the Lotus Sutra in reverence.'

Here, 音 (*oto*, "sound") is used to represent 声 (*koe*, "voice"), where the superordinate category stands for a specific member, constituting 提喩 "synecdoche." Furthermore, the verb 挙て (*agete*, "to raise") is also metaphorical: its prototypical sense involves physically lifting a concrete object, whereas here it is applied to sound/voice (a perceptual rather than manipulable entity), constituting a cross-domain mapping annotated as 換喩 "metonymy" under the CHJ-Metaphor scheme. Such synecdoche and metonymy are particularly difficult to notice during manual annotation due to their subtle nature, yet they occur frequently in classical texts.

Similarly, 入滅 "entering nirvana" was detected as a metaphor in the following passage:

> (2) 専ら入滅の日を思て此の会に参り値はば，罪業を滅して浄土に生れむ事を疑ひ有じ．(Classical Japanese)
> (*Moppara nyūmetsu no hi wo omoite kono-e ni mairi atawaba, zaigō wo messhite jōdo ni umaren koto wo utagahi araji*)
> 'If one attends this gathering with the sole intention of the day of entering nirvana, one would have no doubt about being reborn in the Pure Land after extinguishing one's karmic sins.'

The expression 入滅 (*nyūmetsu*, "entering nirvana") encodes a virtual event script where *dying* is expressed as *entering nirvana*, constituting 換喩 "metonymy" based on event-level contiguity.

These examples demonstrate that the model is capable of detecting subtle metaphors that are easily overlooked during manual annotation, particularly 提喩 "synecdoche," 換喩 "metonymy," and conventionalized metaphors. From a language resource construction perspective, this suggests that the system can serve as an effective supplement to human annotation efforts.

6.4.2. Limitations in Metaphor Detection

However, the analysis also reveals limitations, particularly in handling 結合比喩 "word-level metaphors." In *Hojoki*, the following sentence contains the word-level metaphor 浮雲の思ひ "thoughts like floating clouds":

> (3) 古京はすでに荒れて，新都はいまだ成らず．ありとしある人は皆浮雲の思ひをなせり．(Classical Japanese)

(*Kokyō wa sudeni arete, shinto wa imada narazu. Aritoshi aru hito wa mina ukigumo no omoi wo naseri*.)
'The ancient capital has already fallen to ruin, and the new capital is not yet complete; all people have minds like floating clouds.'

Both 浮雲 (*ukigumo*, "floating clouds") and 思ひ (*omoi*, "thoughts") (underlined in romanization) should be identified as metaphorical, forming a 結合比喩 "word-level metaphor." However, the model failed to detect these core metaphorical tokens while incorrectly marking the surrounding tokens 皆 (*mina*, "all") and なせり (*naseri*, "made/have") as metaphorical. This suggests that the model struggles with identifying the precise token boundaries of extended metaphorical expressions within complex 結合比喩 "word-level metaphors."

Hojoki is particularly rich in 文脈比喩 "discourse-level metaphors" and extended 指標比喩 "similes," leading to mismatches between annotators'and the model's decisions regarding which tokens within a metaphorical passage should be marked. Nonetheless, the model does not appear to miss entire metaphorical sentences or clauses.

6.5. Future Work

Several directions remain for future research.

First, dataset expansion is necessary. This study uses three currently completed works from CHJ-Metaphor (*Konjaku Monogatari-shu*, *Hojoki*, *Toraakira-bon Kyogen-shu*), but the corpus includes additional works under annotation (*Uji Shui Monogatari*, *Jikkinsho*, *Tsurezuregusa*). Once annotation is completed, a more comprehensive evaluation will enable detailed verification of the method's generalizability.

Second, cross-period generalization requires investigation. This study focuses on works from the medieval period, but the applicability to other historical periods remains unexplored.

7. Acknowledgements

This work was supported by JSPS KAKENHI Grants JP22K12145 and JP25K00459, the NINJAL Collaborative Research Projects "Empirical Computational Psycholinguistics Using Annotated Data" and "Extending the Diachronic Corpus through an Open Co-construction Environment", the Kayamori Foundation of Informational Science Advancement Research Grant "Extraction of Conceptual Metaphors Using Natural Language Processing," and the Mitsubishi Foundation Research Grant. We are also grateful to Professor Makoto Yamazaki and Professor Wakako Kashino for providing us with the basic sense data of the WLSP.

8. Bibliographical References

Soma Asada, Kanako Komiya, and Masayuki Asahara. 2024. Comprehensive assignment of Word List by Semantic Principles sense numbers to the Balanced Corpus of Contemporary Written Japanese. In *Proceedings of the 30th Annual Conference of the Association for Natural Language Processing*, pages 2767–2772. In Japanese.

Masayuki Asahara, Nao Ikegami, Tai Suzuki, Taro Ichimura, Asuko Kondo, Sachi Kato, and Makoto Yamazaki. 2022. CHJ-WLSP: Annotation of 'word list by semantic principles' labels for the corpus of historical Japanese. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 31–37, Marseille, France. European Language Resources Association.

Mateusz Babieno, Masashi Takeshita, Dusan Radisavljevic, Rafal Rzepka, and Kenji Araki. 2022. MIss RoBERTa WiLDe: Metaphor identification using masked language model with Wiktionary lexical definitions. *Applied Sciences*, 12(4):2081.

Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1763–1773, Online. Association for Computational Linguistics.

Ellen Dodge, Jisup Hong, and Elise Stickles. 2015. MetaNet: Deep semantic automatic metaphor analysis. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 40–49, Denver, Colorado. Association for Computational Linguistics.

Group Pragglejaz. 2007. MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22(1):1–39.

Anna Hülsing and Sabine Schulte Im Walde. 2024. Cross-lingual metaphor detection for low-resource languages. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 22–34, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

Sachi Kato, Rei Kikuchi, and Masayuki Asahara. 2022. Assignment of metaphorical expression information based on MIP to the Balanced Corpus

of Contemporary Written Japanese. In *Proceedings of the 28th Annual Conference of the Association for Natural Language Processing*, pages 1427–1432. In Japanese.

Beata Beigman Klebanov, Ben Leong, Michael Heilman, and Michael Flor. 2014. Different texts, same metaphors: Unigrams and beyond. In *Proceedings of the Second Workshop on Metaphor in NLP*, pages 11–17, Baltimore, MD. Association for Computational Linguistics.

Yucheng Li, Shun Wang, Chenghua Lin, and Frank Guerin. 2023a. Metaphor detection via explicit basic meanings modelling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 91–100, Toronto, Canada. Association for Computational Linguistics.

Yucheng Li, Shun Wang, Chenghua Lin, Frank Guerin, and Loic Barrault. 2023b. FrameBERT: Conceptual metaphor detection with frame embedding learning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1558–1563, Dubrovnik, Croatia. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Rui Mao, Chenghua Lin, and Frank Guerin. 2019. End-to-end sequential metaphor identification inspired by linguistic theories. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3888–3898, Florence, Italy. Association for Computational Linguistics.

Akira Nakamura. 1977. *Theory and Classification of Figurative Expressions*. Number 57 in Report of the National Institute for Japanese Language and Linguistics. Shuei Shuppan, Tokyo. In Japanese.

Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna A. Kaal, Tina Krennmayr, and Tom Pasma. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. John Benjamins Publishing.

Xiaoyu Tong, Ekaterina Shutova, and Martha Lewis. 2021. Recent advances in neural metaphor processing: A linguistic, cognitive and social perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4673–4686, Online. Association for Computational Linguistics.

Makoto Yamazaki and Wakako Kashino. 2017. Annotation of prototypical sense information for polysemous words in the Word List by Semantic Principles. *Journal of Natural Language Processing*, 24(3):111–126. In Japanese.

Makoto Yamazaki, Wakako Kashino, Kiyoko Uchiyama, Kazuko Sunaoka, Ikuto Tajima, Keiko Yamamoto, Yu-Seok Han, and Yun-Soo Seol. 2015. Annotation of basic sense determination information to the Word List by Semantic Principles, revised and expanded edition. In *Proceedings of the 58th Conference of the Mathematical Linguistics Society*. In Japanese.

Shenglong Zhang and Ying Liu. 2022. Metaphor detection via linguistics enhanced siamese network. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4149–4159, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Hang Zhu, Rowan Hall Maudslay, Kanako Komiya, Sachi Kato, and Masayuki Asahara. 2025. Large-scale japanese metaphor corpus construction: Expanding bccwj-metaphor with automated annotation. In *Proceedings of the 39th Pacific Asia Conference on Language, Information and Computation*. To appear.

9. Language Resource References

Baker, Collin F. and Fillmore, Charles J. and Lowe, John B. 2024. *FrameNet*. International Computer Science Institute.

Maekawa, Kikuo and Yamazaki, Makoto and Ogiso, Takehiko and Maruyama, Masaya and Ogura, Hideki and Kashino, Wakako and Koiso, Hiroshi and Yamaguchi, Makoto and Tanaka, Masahiro and Den, Yasuharu. 2014. *Balanced Corpus of Contemporary Written Japanese*. National Institute for Japanese Language and Linguistics.

National Institute for Japanese Language and Linguistics. 1964. *Word List by Semantic Principles*. Shuei Shuppan.

National Institute for Japanese Language and Linguistics. 2023. *Corpus of Historical Japanese*. National Institute for Japanese Language and Linguistics.

Domain-Aware Error Correction for Citation NER in Medieval Hebrew Responsa

Shmuel Liebeskind[1], **Maayan Zhitomirsky-Geffet**[1], **Binyamin Katzoff**[1], **Nati Ben-Gigi**[1], **Jonathan Schler**[2]
[1]Bar-Ilan University, Ramat-Gan, Israel
[2]Holon Institute of Technology (HIT), Holon, Israel
israellieb@gmail.com, {maayan.zhitomirsky-geffet, Binyamin.Katzoff }@biu.ac.il,
{nati.bengigi, schler }@gmail.com

Abstract

Citation identification in historical and ancient texts poses challenges that extend beyond surface-level pattern recognition, including implicit references, morphological fusion, and discourse-driven ambiguity. In this work, we address citation Named Entity Recognition (NER) in medieval Hebrew Responsa literature using a modular, LLM-based correction pipeline. Rather than treating large language models as end-to-end predictors, we leverage them as structured components: an initial prompt-based expert tagger, complementary LLM judges for systematic error detection, and domain-aware correction grounded in philological regularities. Our approach requires no end-to-end fine-tuning and only minimal labeled supervision (a small validation set for training a lightweight error-detection classifier), narrowing the performance gap to strong supervised models trained on domain-specific data. The results suggest that explicit error handling and interpretability-driven design offer a promising direction for historical NLP in low-resource settings.

Keywords: Historical NLP, Citation Extraction, Medieval Hebrew, Large Language Models, Low-Resource Languages

1. Introduction

Historical and ancient language corpora present enduring challenges for natural language processing, especially when the task necessitates detailed structural interpretation rather than superficial pattern recognition. The Medieval Hebrew Responsa literature, a collection encompassing over 1,300 years of rabbinic legal discourse, illustrates this challenge. Citations in these texts are pivotal to academic reasoning; however, they exhibit significant heterogeneity: references are implicit, abbreviated, morphologically fused with prefixes, and frequently integrated into intricate discourse structures that lack standardized formatting.

Accurate citation Named Entity Recognition (NER) in this domain is therefore not merely a technical preprocessing step, but a prerequisite for large-scale digital humanities research, including citation network construction, intellectual lineage tracing, and diachronic analysis of legal discourse. Prior work has demonstrated that strong supervised models, such as BERT-CRF architectures trained on domain-specific corpora, can achieve high performance on this task. However, these approaches require substantial task-specific annotation and retraining, which is costly and difficult to replicate across historical domains and languages.

In this work, we explore an alternative paradigm: leveraging large language models (LLMs) not as end-to-end taggers, but as structured, rule-aware components within a correction pipeline tailored to historical texts. Our approach builds on a carefully engineered initial prompting stage that functions as a hybrid between a rule-based expert system and a lexicon-guided annotator. This system prompt encodes domain expertise about medieval Hebrew Responsa literature, explicitly defines a hierarchical NER taxonomy, injects lexical anchors for common rabbinic abbreviations, and enforces deterministic tagging constraints that reflect Hebrew morphology and citation practice. The system prompt was generated and iteratively refined by the LLM using solely the validation set, implementing domain knowledge through a validation-driven adaptation process without any access to the test set.

While this prompt-driven baseline already yields competitive performance, its remaining errors are systematic rather than random. We address these errors using an LLM-as-Judge correction pipeline, in which complementary judge personas evaluate token-level predictions and identify likely mistakes. Correction decisions are then guided by domain-aware heuristics that encode philological regularities, such as the distinction between quoted lemma text and actual citations, or between discourse-level mentions of sages and authoritative references. Notably, our approach does not rely on explicit modeling of citation structure; instead, improvements arise from targeted error detection and correction grounded in domain knowledge.

Our contributions are fourfold: (1) we intro-

Proceedings of LT4HALA 2026 @ LREC 2026, pages 96–105.
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

duce a prompt-as-expert baseline for historical citation NER, developed and refined through a validation-driven iterative process; (2) we propose a lightweight LLM-as-Judge correction framework that identifies systematic token-level errors without directly generating replacement labels; (3) we present a domain-aware correction strategy that achieves substantial performance gains without relying on structural heuristics or task-specific fine-tuning; and (4) we demonstrate an annotation-efficient pipeline that achieves competitive performance without large-scale supervised training, reducing reliance on costly expert annotation and enabling rapid adaptation to new historical corpora.

Together, these findings show that in historical and ancient language settings, LLM-based systems are most effective when embedded within interpretable, error-aware pipelines rather than used as monolithic predictors.

2. Related Work

Our work builds on three lines of research: citation and reference extraction, NLP for Hebrew and historical texts, and the use of LLMs as evaluators or corrective components.

Citation and Reference Extraction. Early approaches to citation parsing and reference extraction relied on rule-based systems and Conditional Random Fields (CRFs), typically targeting modern scholarly literature with standardized formats (Councill et al., 2008). Neural sequence models such as BiLSTM-CRF and later BERT-CRF architectures significantly improved performance in well-resourced settings (Lample et al., 2016; Devlin et al., 2019). However, these approaches generally assume stable orthography and explicit citation markers, assumptions that break down in historical corpora.

Within the domain of Hebrew and Jewish texts, prior work has addressed reference detection under more constrained conditions. HaCohen-Kerner et al. (2010) focused on modern Responsa literature, where citation conventions are relatively standardized. Zhitomirsky-Geffet and Prebor (2019) applied pattern-based extraction methods to identify rabbinic sages in the Mishnah. More recently, Ben-Gigi et al. (2025) demonstrated that domain-specific pretraining is crucial for medieval Responsa citation NER, achieving strong results with a BEREL-CRF model trained on Rabbinic Hebrew.

Beyond modern bibliometric parsing, a parallel line of research has addressed citation extraction in historical and classical corpora. Romanello (2013) developed methods for canonical citation extraction from classical texts, later extended to large-scale citation network construction (Kokash et al., 2024). Berti (2019) introduced a framework for extracting fragmentary citations of classical authors, demonstrating that domain-specific structural knowledge is indispensable for texts with non-standard reference conventions. Colavizza et al. (2023) proposed the Humanities Citation Index (HuCI), offering a systematic methodology for citation indexing across historical disciplines. Our work extends this tradition to the medieval Hebrew Responsa domain, where morphological fusion, abbreviation density, and implicit references pose additional challenges beyond those encountered in Greek and Latin corpora.

NLP for Historical and Ancient Languages. A growing body of work highlights the challenges of applying NLP techniques to historical and ancient texts, including diachronic variation, non-standard spelling, and limited annotated data (Piotrowski, 2012; Ehrmann et al., 2016). In response, researchers have emphasized interpretability, robustness, and reuse of linguistic knowledge over purely end-to-end optimization (Bollmann, 2019). Our approach aligns with this paradigm by explicitly encoding philological constraints and systematic error patterns within a modular pipeline.

LLMs as Judges and Corrective Modules. Recent work has explored the use of LLMs as evaluators, critics, or judges of model outputs, particularly in generation and dialogue tasks (Liu et al., 2023; Zheng et al., 2023). These approaches typically frame LLMs as meta-evaluators rather than primary predictors. We extend this paradigm to token-level structured prediction in a historical setting, using LLMs to localize systematic errors in NER outputs. Unlike prior LLM-as-Judge work focused on preference ranking or benchmarking, our framework integrates judging with lightweight supervision and domain-aware correction.

In contrast to existing approaches, we do not aim to replace supervised models or to infer citation structure explicitly. Instead, we demonstrate how LLMs can function as modular, interpretable components that amplify domain knowledge and improve robustness in low-resource historical NLP scenarios.

3. Task and Data

3.1. Task Definition

We address the *internal component identification* task for citation Named Entity Recognition (NER) in medieval Hebrew Responsa literature, following the formulation introduced by Ben-Gigi et al. (2025). Rather than treating each citation as a single span, the task requires identifying and labeling

the atomic components that jointly constitute a reference. This fine-grained decomposition is a prerequisite for downstream tasks such as citation normalization and citation network construction. Unlike flat or nested-span representations, component-level decomposition is essential because downstream tasks such as citation normalization require separate access to each element (e.g., distinguishing the book title from the chapter identifier within the same span) to resolve a reference to a canonical entry. Nested spans would preserve surface boundaries but obscure the functional role of each token.

Each token is assigned one of seven entity types: BN (Book Name, e.g., tractates); AN (Author Name); AA (Author Adjective derived from names); BA (Book Adjective, rare); R (Reference identifiers like chapter/folio); RW (Reference Words like "section" or "page"); and O (Outside, for non-citation tokens).

To illustrate the annotation schema, consider the following sentence fragment from a typical responsum: u-khtav [O] ha-Rambam [AN] be-P"A [R] me-Hilkhot [RW] Isurei [BN] Bi'ah [BN] "And the Rambam wrote, in Chapter 1 of the Laws of Forbidden Relations". Here, ha-Rambam is tagged AN (Author Name), be-P"A (an abbreviation for "in Chapter 1") is tagged R (Reference), me-Hilkhot is tagged RW (Reference Word meaning "laws of"), and Isurei Bi'ah is tagged BN (Book Name). The conjunction u-khtav ("and he wrote") is tagged O (Outside) as a discourse verb, not part of the citation itself.

This schema captures both lexical and structural aspects of rabbinic citations, which are often implicit, abbreviated, and embedded within running discourse.

3.2. Corpus and Experimental Design

All experiments are based on the citation-annotated Responsa corpus released by Ben-Gigi et al. (2025). The corpus consists of medieval Hebrew Responsa texts spanning approximately 1000–1500 CE and covering diverse geographic and cultural traditions, including Ashkenaz, France, Spain, and North Africa. The texts represent a wide range of authors, stylistic conventions, and orthographic practices.

Domain experts in Rabbinic Hebrew and rabbinic literature followed a detailed annotation guideline to produce annotations. The corpus is small compared to modern NER benchmarks due to high annotation costs and expertise, representing low-resource historical NLP settings.

In the original work, the corpus is divided into training and test sets, and results are reported using cross-validation over the training data, reflecting a supervised fine-tuning paradigm. Our setting differs fundamentally. Since our approach does not involve end-to-end model training or fine-tuning, there is no requirement for a large labeled training corpus. Instead, we retain the original test set unchanged to ensure direct comparability with prior results, while constructing a smaller development set sampled from the original training data.

Specifically, we sample a subset of sentences from the original training split to serve as a combined validation and development set. This subset is randomly sampled while preserving the original tag distribution, and is matched to the test set in size and entity composition. This validation set is used exclusively for prompt refinement, iterative error analysis, and training the lightweight error detection classifier. No information from the test set is used during these stages, ensuring strict separation between development and evaluation.

We note that the corpus exhibits considerable internal variability across geographic traditions and time periods. While random sampling preserves global tag distributions, it may under- or over-represent certain stylistic or orthographic conventions. Robustness checks using multiple independent samples, or stratified sampling by tradition and period, would further strengthen reproducibility and are planned for future work.

This experimental design reflects the core methodological goal of our work: enabling effective citation extraction without requiring large-scale supervised training, while maintaining full comparability with prior supervised baselines.

3.3. Data Splits and Statistics

Table 1 summarizes the resulting data splits used in our experiments. We report statistics only for the validation/development set and the test set, as no full training split is required for our method.

Split	Sentences	Tokens	Entity Ratio
Validation	294	3,144	55.2%
Test	263	2,733	52.9%

Table 1: Dataset statistics. Entity ratio denotes the proportion of tokens annotated as citation components (non-*O*); for example, an entity ratio of 55.2% indicates that 55.2% of tokens in the split belong to citation entities, while the remaining 44.8% correspond to regular text (*O*).

Table 2 shows the token-level distribution of entity types in both the validation and test sets. As anticipated, O tokens make up about half of the data, while R and BN are the largest entity classes. Adjective categories (AA, BA) are uncommon, but play a significant linguistic role.

In addition to the primary citation entity categories, the corpus includes auxiliary structural labels grouped under the *N/A* category. These tokens correspond to citation-adjacent or structural

elements such as reference indicators, boundary markers, editorial additions, and discourse signals. While these elements are excluded from named entity evaluation as non-target categories, they provide important contextual cues that assist in identifying citation boundaries and improving overall reference resolution.

Tag	Val (N)	Val (%)	Test (N)	Test (%)
O	1,407	44.8	1,287	47.1
R	609	19.4	481	17.6
BN	346	11.0	277	10.1
RW	260	8.3	186	6.8
AN	135	4.3	176	6.4
AA	85	2.7	101	3.7
BA	30	1.0	6	0.2
N/A	272	8.7	219	8.0

Table 2: Token-level entity distribution in the validation and test sets.

The skewed distribution of entity types in the corpus directly informs several design choices in our method. In particular, the extreme sparsity of adjectival categories—most notably *BA*, which accounts for less than 1% of annotated tokens—makes them especially vulnerable to over-prediction in prompt-based settings. This observation motivates our decision to explicitly mark such categories as rare in the system prompt and to prioritize reference tokens (*R*) in ambiguous contexts.

Similarly, the dominance of *R* and *BN* tokens, combined with the high proportion of non-entity (*O*) tokens, explains both the conservative fallback rules at the prompt stage and the need for a downstream correction mechanism to suppress systematic false positives. The presence of structurally informative but non-evaluated tokens (*N/A*) further reinforces the importance of structural reasoning when identifying citation spans.

These data characteristics highlight why domain-aware constraints and post-hoc error correction are essential for robust citation NER in this corpus.

These distributional properties also shape the error patterns observed at evaluation time, a point we return to in the analysis section, where we examine how sparsity, ambiguity, and boundary effects interact with the correction pipeline.

It is important to note that the relatively high entity ratios reflect the corpus construction methodology, which focuses on citation-rich passages rather than uniformly sampled running text. As a result, citation components are substantially more frequent than in general-domain NER datasets.

The test set is strictly held out and used only for final evaluation. At no stage does the correction pipeline access gold labels from the test data, ensuring a clean separation between development and evaluation and preventing information leakage.

4. Method

Figure 1 provides an overview of the full correction pipeline, illustrating the separation between prompt-based initial tagging, systematic error detection using LLM judges, and domain-aware correction. The following sections describe each component in the order shown in the figure. Our approach consists of a four-stage pipeline designed to separate initial tagging, error detection, and error correction, while explicitly encoding domain knowledge relevant to medieval Hebrew Responsa literature. Crucially, all components are developed under strict train/validation/test separation.

Task Setting and Baseline Assumptions We address the task of token-level Named Entity Recognition (NER) for internal citation components in medieval Hebrew Responsa texts. Given the limited availability of annotated historical data and the high cost of task-specific fine-tuning, we adopt a few-shot prompting setup as a realistic baseline. Rather than treating the LLM as an end-to-end solution, we position it as a modular component within a structured pipeline.

4.1. System Prompt as a Structured Expert Module

The core of our baseline is a carefully engineered system prompt that functions as a hybrid between a rule-based expert system and a lexicon-guided annotator. The prompt was refined over 2–3 validation-only iterations, with the goal of operationalizing philological knowledge rather than optimizing surface-level performance.

Prompt Construction and Refinement. The initial system prompt was not created spontaneously, but generated by the LLM based on specific task definitions and annotation guidelines derived from previous research on medieval Hebrew citation NER, particularly the schema established by Ben-Gigi et al. (2023). This document encompasses the citation taxonomy, boundary conventions, and instances of prevalent rabbinic abbreviations and reference patterns.

Using these specifications as input, the LLM generated an initial system prompt that operationalizes the task as a rule-aware expert annotator. This prompt was then refined through an iterative process on the validation set only, during which the model evaluated its own predictions, identified systematic errors, and adjusted the prompt accordingly. No information from the test set was used at any stage of prompt construction or refinement.

This process characterizes prompt engineering as a lightweight, validation-driven adaptation, as

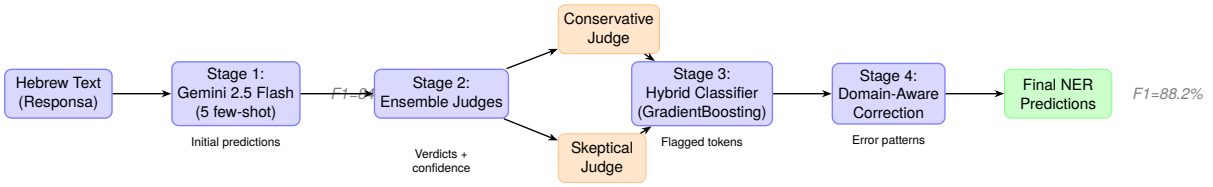


Figure 1: LLM-as-Judge correction pipeline. Stage 1 generates initial predictions; Stage 2 uses complementary judge personas to evaluate each token; Stage 3 employs a GradientBoosting classifier (trained on validation set only) to flag likely errors; Stage 4 applies domain-specific corrections.

opposed to manual rule creation or task-specific fine-tuning.

The final prompt consists of the following components:

Persona and Domain Narrowing. The model is instructed to assume the role of an expert in medieval Hebrew Responsa literature, explicitly prioritizing Rabbinic Hebrew and Aramaic conventions over Modern Hebrew usage. This narrows the model's attention to the appropriate linguistic register and discourse norms.

Hierarchical NER Taxonomy with Prioritization. We define a closed set of entity types (BN, AN, AA, BA, R, RW, O) with explicit definitions. The prompt prioritizes reference tokens (R) as the most important class and marks adjectival modifiers (BA) as rare. This serves as an implicit prioritization heuristic in ambiguous situations.

Lexical Anchors and Abbreviation Awareness. The prompt includes canonical examples of common rabbinic abbreviations and reference markers (e.g., *P"Q* for *Perek Kama* [first chapter], *A"A* for *Amud Aleph* [page side A], and *HaRambam* for Maimonides). These anchors function as a lightweight lookup table, reducing false negatives for highly frequent but opaque forms.

Some reference markers originate from the original authors, while others, such as page indicators and structural divisions, were introduced by later editors. From the perspective of citation NER, both types function as valid anchors for locating references and are therefore treated uniformly within our tagging and correction framework.

Anticipatory Disambiguation Constraints. To address known confusion patterns, the prompt explicitly contrasts similar-looking forms that belong to different categories (e.g., *D'AZ* [of Tractate Avodah Zarah] as a book name versus *P"Q* [First Chapter] as a reference marker). These contrastive explanations guide the model to attend to morphemic roots rather than surface prefixes.

Deterministic Tagging Rules. Finally, the prompt enforces strict structural constraints: all tokens must be tagged in order; Hebrew prefixes remain attached to their base tokens; and When uncertainty persists, the model defaults to tagging the token as *R*. This fallback strategy promotes tagging consistency and reduces fragmentation errors at the prompt stage. While this bias intentionally favors recall, potential over-generation is subsequently addressed by the downstream error detection and correction pipeline

All prompt-based annotation and judge evaluations were performed using the Gemini-2.5 Flash LLM, without any task-specific fine-tuning. This model was selected for practical considerations: it was the largest freely accessible LLM at the time of experimentation, provided reliable structured output generation, and supported the long-context windows required for our prompting setup. We do not claim that Gemini-2.5 Flash is necessarily the best-performing model for this task; no systematic cross-model comparison was conducted, and we expect that other large LLMs would yield comparable results given appropriate prompt re-optimization.

4.2. Initial LLM Tagging

Using this system prompt, we apply a few-shot prompting setup with 5 curated examples that emphasize rare entity types and complex citation patterns. This stage produces the initial token-level NER predictions that serve as input to subsequent correction stages.

4.3. LLM-as-Judge for Error Detection

To identify systematic errors in the initial predictions, we employ an LLM-as-Judge framework with two complementary judge personas:

- **Conservative Judge:** biased toward high precision and skeptical of entity assignments.
- **Skeptical Judge:** biased toward high recall and attentive to missed entities in entity-dense contexts.

The 0.8 confidence threshold for the Skeptical Judge was selected based on validation set

tuning. Thresholds of 0.7, 0.8, and 0.9 were compared; 0.8 provided the best balance between flagging genuine errors and avoiding excessive false alarms, yielding the highest downstream correction F1 on the validation set.

Each judge independently evaluates the predicted tag for each token and outputs a binary verdict (correct / incorrect) with a confidence score. Judges are not allowed to propose alternative tags; their role is limited to error detection.

4.4. Hybrid Error Detection Classifier

Judge outputs alone are insufficient due to frequent disagreement in ambiguous cases. We therefore train a lightweight Gradient Boosting classifier to predict the likelihood that a token is incorrectly tagged, using scikit-learn default hyperparameters. The classifier was trained using 25 features across six categories: judge confidence scores and agreement levels (4), aggregated verdict indicators and vote counts (4), agreement pattern types (4), one-hot encoding of proposed tag types (7), token-level linguistic features including word length and Hebrew prefix detection (3), and positional features capturing token location within the sentence and citation span (3).

This feature design enables the classifier to combine structural, semantic, and consensus-based signals to resolve ambiguous tagging decisions

The classifier is trained only on the validation set using gold labels for error detection. Once trained, it is frozen and applied to the test set without access to ground truth labels, ensuring no leakage.

Among the six feature categories, judge agreement and confidence scores were the strongest predictors of classification errors, followed by tag-type indicators such as predicted entity label, and positional features including token location and entity density.

We emphasize that the Gradient Boosting classifier constitutes a lightweight supervised component trained on gold error-detection labels from the validation set. The pipeline therefore involves minimal labeled supervision at the error-detection stage, while the core NER predictions at Stages 1, 2, and 4 remain unsupervised. We adopt the term "minimal supervision" rather than "no fine-tuning" to accurately characterize this design.

4.5. Domain-Aware Error Correction

Tokens flagged as likely errors are passed to a correction stage that applies domain-aware prompting. Correction prompts encode philological regularities observed during validation analysis, including:

- **Lemma formulas:** Recognition of markers such as *D"H* (abbreviation for *Dibbur Ha-Matkhil*) or *Be-Dibbur Ha-Matkhil* ("In the comment beginning with"), which signal that subsequent tokens are quoted text rather than location references (R).
- **Discourse Context:** Distinguishing discourse mentions of sages from authoritative citations. For instance, the system corrects instances where names are over-tagged in narrative contexts, such as *Ben Shamoa Omer* ("Ben Shamoa says"), which should be Outside (O), versus actual citations like *Ke-de-pirash Ben Shamoa* ("As Ben Shamoa explained"), where the name is an Author Name (AN).
- **Entity Ambiguity:** Resolving common confusions between authors and their works, specifically identifying whether a marker like *Ha-Rosh* refers to the person (AN) in the context of *Katav Ha-Rosh* ("The Rosh wrote") or the book (BN) in the context of *Be-Sefer Ha-Rosh* ("In the book of the Rosh").

Corrections are applied conservatively and only to flagged tokens, preserving the original predictions elsewhere.

Table 3 summarizes the distinct prompt configurations used across the pipeline.

5. Results

Table 4 shows the main test set results. The prompt-based baseline achieves a high F1 score, indicating that a well-designed system prompt with domain knowledge and constraints can capture citation behavior in medieval Hebrew texts.

We also evaluated both one-shot and few-shot prompting configurations. The performance differences between them were marginal and not statistically significant, indicating that explicit task instructions and domain-aware constraints contributed more to performance than the inclusion of demonstration examples.

Applying the LLM-as-Judge correction pipeline yields a substantial improvement of approximately four F_1 points over the prompt-only baseline. This gain is driven primarily by a marked increase in precision, alongside a smaller but consistent improvement in recall. These results indicate that the correction pipeline is particularly effective at suppressing systematic false positives while still recovering a meaningful number of missed citations.

Notably, the corrected prompt-based system narrows the performance gap to a strong supervised BEREL-CRF model trained on domain-specific data, despite requiring no task-specific fine-tuning or additional annotated data. This highlights the

Table 3: Overview of the distinct prompt configurations used across the pipeline. Each component employs a role-specific prompt with different objectives and inductive biases, enabling modular tagging, error detection, and targeted correction.

Aspect	Tagger	Conservative Judge	Skeptical Judge	Correction Module
Primary role	Assign citation-related tags to all tokens	Validate predicted tags and protect correct assignments	Aggressively identify potential tagging errors	Apply targeted fixes to flagged tokens only
Input	Raw text	Text + predicted tags	Text + predicted tags	Text + predicted tags + flagged token positions
Output	Token-level tags	Correct/Incorrect verdict per token	Correct/Incorrect verdict per token	Corrected tags for flagged tokens
Inductive bias	Favor recall; when uncertain, default to *R*	Favor precision; default to CORRECT when evidence is ambiguous	Favor recall; default to INCORRECT under uncertainty	Preserve non-flagged predictions exactly
Typical error patterns addressed	Boundary ambiguity; abbreviation confusion	Over-tagging of non-citation material	Missed or weakly marked citation boundaries	Systematic philological error patterns

Method	P	R	F1
Prior Work			
BEREL-CRF	90.9	88.4	89.6
Prompting Methods			
Zero-shot	81.4	87.4	84.3
Few-shot	78.7	90.9	84.4
Our Approach			
+ LLM-as-Judge	**84.5**	**92.3**	**88.2**

Table 4: Test set results. Our pipeline improves few-shot by +3.8 F1 points, approaching BEREL-CRF performance without task-specific fine-tuning.

effectiveness of explicit error detection and correction as an alternative to retraining in low-resource historical settings.

We started with citation subtype analyses as a diagnostic tool. We found that explicit citation structure modeling did not affect performance gains. To maintain a structure-agnostic and portable pipeline, we exclude such analyses from the final system and report only token-level error detection and correction results.

5.1. Ablation Study

To assess the contribution of individual components within the correction pipeline, we conduct an ablation study summarized in Table 5. Starting from the prompt-based baseline, we incrementally add (i) LLM judges with majority voting, (ii) a hybrid error detection classifier, and (iii) the full domain-aware correction stage.

Note: The ablation results in Table 5 are evaluated on the validation set, whereas Table 4 reports test set performance. The difference in few-shot F1 (85.3 on the validation set versus 84.4 on the test set) reflects natural variation between the two splits and does not indicate an inconsistency.

Configuration	F1	Δ
Few-shot only	85.3	—
+ Ensemble Judges + Classifier	87.5	+2.2
+ Correction	**89.8**	+2.3
Error Detection (Flagging)		
Ensemble Voting only	55.9	—
+ Hybrid Classifier	**66.4**	+10.5

Table 5: Ablation study (validation set). The hybrid classifier improves error detection F1 by +10.5% over voting alone by using token-level features. Note: The few-shot baseline F1 differs from Table 4 because the two tables report on different evaluation splits (validation vs. test).

Introducing LLM judges alone yields a moderate improvement, indicating that judge agreement provides a useful but incomplete signal for identifying erroneous predictions. Adding the hybrid classifier further improves performance, demonstrating that combining judge signals with token-level features substantially enhances error detection. The full correction pipeline achieves the highest F_1 score, confirming that performance gains arise from the interaction between systematic error detection and targeted, philologically grounded correction.

Notably, we exclude ablations based on explicit

citation subtype modeling, as preliminary experiments showed that correction gains did not depend on such structural assumptions. This design choice reflects our goal of maintaining a structure-agnostic and portable pipeline.

6. Analysis

We analyze the behavior of the correction pipeline to better understand the sources of improvement and remaining limitations.

6.1. Judge Agreement Analysis

We first analyze agreement patterns between the two LLM judges to assess the reliability of error detection signals. When both judges agree that a prediction is correct, accuracy reaches 94.2%, indicating a high-confidence region where correction is unnecessary. Conversely, when both judges flag a prediction as erroneous, 71% of such cases correspond to true errors, demonstrating that joint disagreement provides a strong signal for corrective intervention. Across the test set, the LLM judges flag approximately 14% of tokens as potential errors. The majority of these cases correspond to genuine misclassifications. Disagreements between judges occur primarily in boundary cases involving short functional tokens, abbreviations, or morphologically fused forms, which are known sources of ambiguity in Rabbinic Hebrew.

Complementary Judge Biases. The two LLM judges are deliberately designed to embody complementary error-detection biases. The *Conservative Judge* is provided with extensive guidance on valid citation patterns and is instructed to avoid false positives, defaulting to CORRECT when evidence is ambiguous. In contrast, the *Skeptical Judge* operates as an aggressive auditor, explicitly instructed to assume that tags may be incorrect and to flag any token whose correctness confidence falls below 0.8, thereby prioritizing recall over precision.

This asymmetric design—one judge protecting correct predictions and the other actively hunting for errors—explains the agreement patterns observed in practice. Joint CORRECT decisions correspond to a high-precision regime, while joint INCORRECT decisions yield a strong true positive signal for error detection. Disagreement cases naturally occupy an intermediate confidence region, motivating the use of a hybrid classifier to resolve ambiguity.

Why a Classifier? Simple voting rules are a natural baseline for aggregating judge decisions, but ineffective. Using majority voting alone for error detection yields only 55.9% F_1. In contrast, the hybrid error detection classifier enhances performance by +10.5 percentage points to 66.4% F_1.

This improvement is primarily driven by the classifier's ability to resolve "split-decision" cases, where the two judges disagree. Across the test set, such disagreements occur for 191 tokens. By incorporating token-level features alongside judge signals, the classifier successfully disambiguates many of these cases, leading to more accurate error localization than voting alone.

These results justify the inclusion of a lightweight supervised classifier and demonstrate that effective error detection requires integrating complementary judgment signals with local contextual information.

Importantly, the two judges exhibit complementary biases: one judge adopts a more conservative stance, favoring high precision in error detection, while the other is more skeptical, prioritizing recall. This asymmetry proves beneficial, as disagreements between judges frequently arise in such boundary cases. The hybrid error detection classifier effectively integrates these complementary signals, improving robustness beyond simple majority voting.

These results validate that structured error localization, not redundant or correlated judgments, is the source of performance gains and support the design decision to use multiple judge personas.

6.2. Error Reduction Analysis

To quantify the practical impact of the correction stage, we analyze the distribution and effect of corrected tokens. Across the test set, the pipeline flags 377 tokens (13.8%) across 189 sentences as candidates for correction. This selective intervention indicates that the system operates conservatively, modifying only a small subset of predictions rather than reprocessing all tokens.

Precision increases by 5.8% and recall by 1.4 points during the correction phase. This pattern indicates that the pipeline effectively reduces false positives and recovers missed citation entities. The increase in precision and recall indicates that gains are not solely due to conservative pruning, but also to meaningful error correction.

These findings are consistent with the ablation results, which show that performance improvements emerge only when error detection is followed by domain-aware correction, rather than from judge signals alone.

Dominant Correction Patterns. Qualitative inspection reveals that most successful corrections follow a small number of recurring philological patterns. A dominant category involves lemma formulas, where quoted text following formulaic markers is frequently misclassified as a citation by the

baseline model. The correction stage consistently reassigns such tokens to the non-citation class. Another common pattern concerns discourse-level mentions of sages or authorities that are not intended as authoritative references, which the correction heuristics successfully suppress.

Reference Type as a Diagnostic Signal. Although reference type information is not used by the correction pipeline, we analyze model behavior across reference categories as a diagnostic tool. Reference types were automatically inferred by the LLM in a post-hoc analysis, without access to gold labels, based solely on the predicted citation spans and local context.

Table 6 shows that error rates vary substantially across reference types. Explicit structural references (e.g., chapter or folio identifiers) exhibit the lowest error rates, while implicit or discourse-level references are considerably more error-prone. Notably, lemma-based references—where quoted text follows formulaic markers—account for a disproportionate share of false positives, consistent with the qualitative correction patterns discussed above.

Ref. Type	N	Init	Final	△F1	Net
Explicit	67	76.5	87.0	+10.4	+48
Consecutive	50	86.2	90.6	+4.4	+25
Recursive	17	80.4	83.2	+2.8	+5
Hierarchical	75	88.7	89.7	+1.0	+7
Partial	38	90.7	88.2	-2.5	-4
Unknown	10	90.0	81.5	-8.5	-9

Table 6: Error analysis by reference type (test set). Net = errors fixed − errors introduced. Clear structures (explicit, consecutive) benefit most; ambiguous types show regressions.

These findings explain why explicit modeling of reference type did not improve performance: reference categories correlate with error likelihood but cannot reliably determine token-level prediction accuracy. However, the analysis highlights systematic challenges that may inform future work on citation modeling and normalization.

Stability and Practical Considerations. While LLM-based predictions exhibit some stochasticity across runs, the correction pipeline acts conservatively, intervening only on tokens flagged as likely errors. In practice, this stabilizes predictions rather than amplifying variance. Computationally, the pipeline incurs additional latency due to multiple LLM calls, making it more suitable for offline or batch processing scenarios typical of digital humanities research.

In terms of computational cost, the full pipeline processes approximately 3–5 sentences per minute. when run sequentially against the LLM‘s API, with most of the latency attributable to the three LLM calls per sentence (initial tagger and two judges) and optional rule-based correction. For the 263-sentence test set, end-to-end processing takes approximately 40 minutes. This latency is acceptable for offline digital humanities workflows but would require batching or parallelization for large-scale corpora comprising thousands of Responsa.

Overall,The analysis supports our central claim that performance gains result from systematic error detection and domain-aware correction based on philological knowledge, rather than implicit citation structure modeling or increased model complexity.

7. Conclusion

We developed a modular LLM-based correction pipeline for medieval Hebrew, addressing linguistic variation and limited data. To avoid end-to-end prediction, we use LLMs as structured, rule-aware components in an error-driven pipeline.

We discovered that systematic error detection and targeted correction can improve performance without task-specific fine-tuning. Validated system prompts provide a lightweight expert base, while complementary LLM judges identify errors for conservative, domain-aware correction based on philological regularities.

Beyond performance improvements, the approach substantially reduces reliance on large annotated corpora. While prior work depends on fully annotated training sets, our pipeline operates effectively using only a small validation set and iterative LLM-guided refinement, enabling efficient adaptation and scalable annotation workflows in low-resource historical domains.

Although the overall pipeline architecture (prompt-based expert tagging, LLM judges, and a hybrid classifier, domain-aware correction) is language- by design, practical portability requires re-engineering two corpus-specific components: the lexical anchors and abbreviation tables embedded in the system prompt, and the philological correction heuristics of Stage 4. While applying this pipeline to other historical traditions (e.g., Talmudic commentary, medieval Latin legal texts, or classical Greek citations) requires domain-expert input to replace these specific modules components, the underlying structural framework and judge-classifier architecture remain directly reusable. This highlights the potential of interpretable, error-aware LLM pipelines as a practical, scalable alternative to traditional supervised approaches for ancient and specialized texts.

8. Reproducibility

To support reproducibility and further research, all system components are available at https://github.com/semlie/Correction-Pipeline-for-Citation. This release includes: (1) the complete system and judge prompts used across all pipeline stages; (2) the Gradient Boosting classifier implementation and trained model weights; (3) the feature extraction code and domain-aware correction rule set; and (4) the validation/test split definitions. The annotated corpus is available through the original release by Ben-Gigi et al. (2025).

9. Bibliographical References

Nati Ben-Gigi, Maayan Zhitomirsky-Geffet, Jonathan Schler, and Binyamin Katzoff. 2025. Automatic construction of the citation network from the medieval jewish responsa literature. *ACM Journal on Computing and Cultural Heritage*, 18(2):1–18.

Monica Berti. 2019. *Digital classical philology: Ancient Greek and Latin in the digital revolution*, volume 10. Walter de Gruyter GmbH & Co KG.

Marcel Bollmann. 2019. A large-scale comparison of historical text normalization systems. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)*, pages 3885–3898.

Giovanni Colavizza, Silvio Peroni, and Matteo Romanello. 2023. The case for the humanities citation index (huci): a citation index by the humanities, for the humanities. *International Journal on Digital Libraries*, 24(4):191–204.

Isaac Councill, C Lee Giles, and Min-Yen Kan. 2008. Parscit: an open-source crf reference string parsing package. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Maud Ehrmann, Chiara Palladino, and Giuseppe GA Celano. 2016. Diachronic named entity recognition and linking. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 1–9. The COLING 2016 Organizing Committee.

Yaakov HaCohen-Kerner, Nadav Schweitzer, and Yaakov Shoham. 2010. Automatic identification of biblical quotations in hebrew-aramaic documents. In *International Conference on Knowledge Discovery and Information Retrieval*, volume 2, pages 320–325. SCITEPRESS.

Natallia Kokash, Matteo Romanello, Ernest Suyver, and Giovanni Colavizza. 2024. The brill knowledge graph: A database of bibliographic references and index terms extracted from books in humanities and social sciences. *Research Data Journal for the Humanities and Social Sciences*, 9(1):1–21.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 260–270.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 2511–2522.

Michael Piotrowski. 2012. *Natural language processing for historical texts*. Morgan & Claypool Publishers.

Matteo Romanello. 2013. Creating an annotated corpus for extracting canonical citations from classics-related texts by using active annotation. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 60–76. Springer.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Maayan Zhitomirsky-Geffet and Gila Prebor. 2019. Sagebook: Toward a cross-generational social network for the jewish sages' prosopography. *Digital Scholarship in the Humanities*, 34(3):676–695.

From Lemmas to Links: A Lemma Bank for Ancient Greek

Colin Swaelens†, Francesco Mambrini*, Marco Passarotti*
†Language & Translation Technology Team & Dpt. of Linguisitcs; Ghent University
*CIRCSE, Università Cattolica del Sacro Cuore (Milano)
colin.swaelens@ugent.be, francesco.mambrini@unicatt.it, marco.passarotti@unicatt.it

Abstract

This paper presents the initial construction of a lemma bank for Ancient Greek, developed according to the Linked Data principles. The need for interoperable linguistic infrastructures capable of supporting interoperability among historical variation, divergent annotation practices, and resource-specific lemmatisation conventions was highlighted by the increasing availability of digital linguistic resources. This lemma bank is developed as the core component of the Linking Greek knowledge base and inspired by the architecture of the LiLa project for Latin. The proposed lemma bank adopts a descriptive, lemma-centric approach that preserves alternative canonical forms and dialectal variation while enabling consistent linking across lexical and semantic resources. Its population combines data extracted from the Ancient Greek WordNet and the Liddell–Scott–Jones lexicon, integrating semantic structure, part-of-speech information, and lexicographically encoded gender assignment. Additional normalisation steps, including the rule-based correction of closed-class part-of-speech categories and harmonisation to the Universal Dependencies tagset, were applied to improve consistency and computational usability. The resulting dataset provides a foundation for interlinking corpora, lexica, and NLP tools for Ancient Greek within a Linked Open Data framework.

Keywords: Linked Open Data, Ancient Greek, Linguistic Resources

1. Introduction

The rapid growth of the field of ancient language processing (ALP) can be attributed, in part, to the increasing availability of both lexical and textual resources (Sommerschield et al., 2023). For Ancient Greek, this development has resulted in a rich landscape of digital resources. These include textual resources like the Perseus Digital Library[1] and the First1KGreek Project[2], syntactically annotated corpora such as the Ancient Greek Dependency Treebanks (Bamman and Crane, 2011) and the Pedalion Trees (Keersmaekers et al., 2019), semantic resources such as the Ancient Greek WordNet (Bizzoni et al., 2014), as well as lexical reference works, among which Liddell-Scott Jones (LSJ)[3].

These resources, however, adopt different annotation strategies, display a variety of data formats, and employ incompatible label sets: an issue well attested for many other languages as well. For Ancient Greek, this fragmentation is further shaped by historical and linguistic factors: prior to the emergence and subsequent diffusion of the Koine (4th c. BC), Greek did not constitute a single, standardised language, but rather a collection of regional dialects, some of which differ substantially in their phonology (Example 1a), morphology (Example 1b), and lexicon (Example 1c).

(1) a. Attic θάλαττα *thalatta* vs.
Ionic θάλασσα *thalassa* (sea)

b. Attic εἶναι *einai* vs.
Doric ἔμμεν *emmen* (to be)

c. Attic αὐτόν *auton* vs.
Ionic μιν *min* (acc. sg. 3rd pers. pron.)

This linguistic variation amplifies the heterogeneity already inherent in the existing resources, which negatively affects their use and hinders interoperability. The LSJ, for example, records both variants of Example 1a as entries, while the Greek–Dutch reference dictionary (Sluiter et al., 2024) consistently selects the Attic variant as headword. This may lead to artificially low scores in evaluation settings—for instance, when lemmatisation systems are penalised for producing a dialectal variant that is linguistically valid, yet absent from the chosen gold standard, as described in Swaelens et al. (2024).

To address issues of interoperability across resources, the principles of Linked Open Data (LOD) have been increasingly applied to linguistic data. Within this context, a lively research community has emerged around Linguistic Linked Open Data (LLOD), which promotes the publication of linguistic resources in accordance with Linked Data standards and Semantic Web technologies. See, for instance, the NexusLinguarum COST Action (CA18209), which fostered collaboration in the field

[1] https://www.perseus.tufts.edu/
[2] https://opengreekandlatin.github.io/First1KGreek/
[3] https://github.com/helmadik/LSJLogeion

Proceedings of LT4HALA 2026 @ LREC 2026, pages 106–111.
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

of LLOD. By representing linguistic information as interlinked entities identified by persistent URIs, LOD enables explicit connections between otherwise distinct datasets while preserving their internal annotation choices.

One of the most fully developed instantiations of the LOD principles is the LiLa (Linking Latin) project (Passarotti et al., 2020), which builds a linked data-based knowledge base of linguistic resources and natural language processing tools for Latin. The LiLa knowledge base's principal component is the lemma bank. Lemmas provide a stable point of reference across resource types: they occur as entries in lexical resources, their inflected forms appear as tokens in textual resources, and they constitute the unit over which many NLP tools operate. This makes them a natural anchoring unit for interoperable linguistic modelling.

Building on the experience of LiLa and reusing its underlying architecture, the Linking Greek project has initiated the construction of a knowledge base of interoperable linguistic resources for Ancient Greek as linked data. This paper describes the development of its principal component: a collection of Greek lemmas, hereafter referred to as the *lemma bank*. The lemma bank is designed to serve as an anchoring layer between word occurrences in corpora and their corresponding entries in lexical resources, thereby enabling consistent linking across heterogeneous datasets within the knowledge base.

2. Background

Building on the foundations of the Semantic Web (Berners-Lee et al., 2001), the Linked Data paradigm is organised around a set of four core principles that enable the data to be identified, queried and semantically interpreted (Berners-Lee, 2006):

1. Use Uniform Resource Identifiers (URIs) to name "things"
2. Use HTTP URIs so that people can look up those names
3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
4. Include links to URIs so that they can discover more things

These principles have been successfully applied in a number of linguistic knowledge bases, illustrating how Linked Data can be used to model linguistic resources in practice. The LiLa knowledge base for Latin is a prominent example, as it represents linguistic objects (occurrences of words in texts, lexical entries in dictionaries, parts-of-speech, etc.) as URIs that can be accessed via HTTP. By explicitly linking lemmas to textual occurrences and annotations, LiLa demonstrates how heterogeneous datasets can be integrated while preserving their internal structure and annotation schemes. In addition, LiLa interlinks several lexical resources, including an etymological dictionary (de Vaan, 2008), thereby enabling connections between Latin lemmas and their Proto-Indo-European roots[4]. This integration illustrates how Linked Data principles facilitate cross-linguistic linking.

Similar approaches have been adopted for other (historical) languages. For Old Irish, a Linked Data knowledge base has been developed that converts the GoldLex lexical dataset into a lemma bank (Fransen et al., 2024), focussing primarily on nominal morphology and demonstrating the applicability of LLOD principles to historically less-resourced languages. A comparable application of the Linked Data paradigm can be found in LiIta, a knowledge base for Italian that already initiated its core component, the lemma bank (Litta et al., 2024)[5]. Together, these resources illustrate how Linked Data principles can be applied to linguistic data at different levels of scope and maturity, with lemma-centric modelling emerging as a recurring design pattern across languages.

3. The Greek Lemma Bank

Initiating the Linking Greek knowledge base with a lemma bank is a natural design choice, already motivated by the architecture of its Latin counterpart (Passarotti et al., 2020), in which lemmas function as anchoring units across different types of lexical, linguistic and textual resources.

In this approach, lemmatisation is treated as a shared organisational layer across different kinds of resources. Lemmas serve as indexing units in dictionaries and thesauri, enable the aggregation of semantically related entries, and facilitate lexical access to corpora, an especially crucial function for morphologically rich languages. In Ancient Greek, for example, a single verbal lemma can represent up to 242 forms of a verb, the participle forms not included.

On this basis, the lemma emerges as the most productive interface between lexical resources, annotated corpora, and NLP tools.

3.1. Resources

To construct the initial version of the Ancient Greek lemma bank, three complementary resources were considered: the lemma list of the morphological parsing tool Morpheus (Crane, 1991), the LSJ,

[4] https://lila-erc.eu/data-page/.
[5] https://www.liita.it

Resource	Number	Primary focus
Morpheus	45,868	Morphology
LSJ	116,924	Lexicography
AGWN	112,513	Semantics

Table 1: Overview of the primary resources considered for the construction of the Greek lemma bank.

and the Ancient Greek Word Net (Bizzoni et al., 2014). These resources represent distinct perspectives on the Greek lexicon, reflecting respectively a parser-oriented, a lexicographic, and a semantic approach to lemma representation. Each resource contributes a different lemma inventory, yet none of them, in isolation, provides a sufficient basis for a lemma bank intended to support interoperability across heterogeneous linguistic resources. To construct the initial version of the Ancient Greek lemma bank, three complementary resources were considered: the lemma list of the morphological parsing tool Morpheus (Crane, 1991), the LSJ, and the Ancient Greek Word Net (Bizzoni et al., 2014). These resources represent distinct perspectives on the Greek lexicon, reflecting respectively a parser-oriented, a lexicographic, and a semantic approach to lemma representation. Each resource contributes a different lemma inventory, yet none of them, in isolation, provides a sufficient basis for a lemma bank intended to support interoperability across heterogeneous linguistic resources.

While alternative approaches based on lemmatised corpora or treebanks were considered, we opted to prioritise lexical resources. In annotated corpora like Diorisis (Vatri and McGillivray, 2018) and Opera Graeca Adnotata (Celano, 2024), lemma assignment is often context-dependent and may vary according to syntactic interpretation, which makes it less suitable as a stable, canonical reference layer. For the purposes of constructing a lemma bank as an interoperability layer, a consistent and context-independent representation of lemma identity is required. Lexical resources, by contrast, provide a more controlled and abstraction-oriented representation of lemmas, which better aligns with the objectives of a lemma-centric Linked Data infrastructure.

Morpheus' lemma list provided a useful starting point, since it records not only lemmas but also associated linguistic information like part-of-speech and, for nouns, gender, in addition to the dialectal labelling, as noted above. Nevertheless, its coverage is comparatively limited and therefore insufficient as the sole backbone of the knowledge base (Table 1).

Coverage was not a limitation for the LSJ, whose lexicographic content is both extensive and reliable. The dictionary also records a substantial amount of dialectal variation (cf. Example 1); however, extracting this information automatically proved challenging, mainly because the relevant relations are encoded in heterogeneous and, at times, inconsistent ways. For phonological variation (cf. Example 1a), the entry for θάλαττα *thalatta* ('sea') explicitly states that the headword and the related verbal and adjectival forms, constitutes the Attic counterpart of θάλασσα (Example 2a). Yet comparable alternations are elsewhere encoded via cross-references using *v.*, as in τάττω *tattō* 'to put' *v.* τάσσω (Example 2b). Importantly, the same *v.* relation is also used for derivational or morphosyntactic links that are not dialectal variants at all, such as χάριν (an adverbial use of the accusative of χάρις) *v.* χάρις (Example 2c). As a consequence, it was not possible to automatically extract and link these forms to one another.

(2) a. θάλαττα, θαλαττεύω, θαλάττιος, etc., Att. for θάλασσα, etc.

b. τάττω v. τάσσω

c. χάριν v. χάρις

The Ancient Greek Wordnet provides a large-scale, systematically structured lemma inventory that approaches the coverage of the *LSJ*. In line with the WordNet paradigm, AGWN is organised around synsets representing lexicalised concepts, with lemmas functioning as the linguistic realisations of semantic senses rather than as dialectally normalised headwords. As a result, dialectal and phonological variation is not represented consistently: while AGWN includes both θάλασσα *thalassa* and θάλαττα *thalatta*, it only lists τάσσω *tassō* and not its Attic counterpart τάττω *tattō*. AGWN further provides part-of-speech and morphological information to support computational processing, although its annotation scheme does not fully align with Universal Dependencies conventions. Despite these limitations, AGWN offers a conceptually clean and explicitly structured view on lemma identity, which complements the lexicographic orientation of *LSJ* and the parser driven design of Morpheus.

3.2. The ontology of Ancient Greek canonical forms

Alternative canonical solutions adopted by different resources are modelled using the property `lila:lemmaVariant`, a symmetric relation defined in the LiLa ontology that connects different inflected forms of the same word that can be chosen for lemmatisation. The ontology also retains the class `lila:Hypolemma`, introduced in LiLa to model forms that may function as lemmas and are members of the regular inflectional paradigm of another lemma, with a different part of speech (e.g. participles: ADJ, members of verbal paradigms).

While this modelling option is preserved for conceptual compatibility, current evidence from Greek corpora suggests that participles are not typically used as canonical lemmas; consequently, no instances of `lila:Hypolemma` are instantiated in the present version of the Greek lemma bank. Orthographic differences that do not affect inflectional behaviour are represented separately using the `ontolex:writtenRep` property provided by OntoLex-Lemon. The ontology distinguishes the following set of core entities that capture the abstraction levels required to model Ancient Greek canonical forms and their variation:

- **Lemma**: the index of all inflected forms that is conventionally identified as canonical. For example, the lemma γράφω *graphō* (to write) indexes inflected forms such as γράφει *graphei* (A. ind. pres. 3rd. sg.) and ἔγραψεν *egrapsen* (A. ind. aor. 3rd. sg.).
- **Lemma Variant**: the symmetric property to link multiple forms that could be used as lemma for a specific token. An example could be the co-existence of the active and mediopassive voice of verbs in a lexicon, such as γράφω *graphō* and γράφομαι *graphomai*.
- **Written Representation**: the different spellings or peculiar inflections canonical forms can have. An example we already discussed above is the alternation of θάλασσα *thalassa* and θάλαττα thalatta. This property is used only when the variation in the realisation of the lemma affects only the orthography of a form, provided that its morphology is not altered.

These entities are connected through a small set of explicit properties that formalise how canonical forms, their variants, and their representations interact within the lemma bank:

- **lila:hasLemma**: links a lemma to its (inflected) word forms occurring in textual resources.[6]
- **lila:lemmaVariant**: a symmetric property linking lemma variants that represent alternative canonical solutions for the same lexical item.[7]
- **ontolex:writtenRep**: associates a lemma with its orthographic realisations.[8]

The population of the Greek lemma bank was initiated from the Ancient Greek WordNet, which served as the primary source for lemma extraction. Using the AGWN API index call, we retrieved the full set of lemmas together with all associated part-of-speech analyses. This proved particularly valuable, as a single written form may receive multiple part-of-speech assignments, which could be directly modelled as lemma variants within the ontology. At the same time, inspection of the AGWN data revealed inconsistencies in part-of-speech annotation, plausibly related to its largely automatic annotation pipeline; for instance, a number of prepositions are tagged as pronouns.

To improve the reliability of the part-of-speech information, an additional normalisation step was introduced for closed-class words. Using the lemma list provided by Morpheus as a reference inventory, closed-class items such as prepositions, conjunctions, particles, and pronouns were identified and used to correct inconsistent part-of-speech assignments in the extracted data. This rule-based correction was applied prior to mapping the resulting PoS labels to the Universal Dependencies tag set.

A further limitation of AGWN is the absence of gender information for nominal lemmas. To address this gap, information from the *Liddell–Scott–Jones* lexicon was integrated. Despite its heterogeneous and lexicographically oriented structure, LSJ consistently characterises nominal entries by means of the definite article, which allows both the unambiguous identification of nouns and the reliable extraction of their grammatical gender. This information was therefore combined with the lemma and part-of-speech data derived from the AGWN.

3.3. Resulting Data Set

The resulting version of the Greek lemma bank comprises *100,280* lemmas. Of these lemmas, *41,173* are nominal, for which grammatical gender could be assigned on the basis of LSJ evidence. The distribution of part-of-speech categories reflects the predominance of open-class items, with verbs and nouns accounting for the majority of entries, while closed-class categories were normalised following the procedure described above. All part-of-speech labels were harmonised to the Universal Dependencies tag set. Table 2 summarises the distribution of lemmas across Universal Dependencies part-of-speech categories in the current version of the Greek lemma bank.

One remaining issue concerns the distinction between common nouns and proper nouns, in particular personal names, which is not systematically captured in the current resources. Addressing this distinction is left for future work and may benefit from the integration of gazetteers or named entity resources.

[6] http://lila-erc.eu/ontologies/lila/hasLemma

[7] http://lila-erc.eu/ontologies/lila/lemmaVariant

[8] http://www.w3.org/ns/lemon/ontolex#writtenRep

UD Part-of-Speech	Number of lemmas
NOUN	41,173
ADJ	28,827
VERB	27,573
ADV	2,509
ADP	73
SCONJ	49
PART	32
PRON	21
NUM	14
CCONJ	5
INTJ	4

Table 2: Distribution of lemmas in the Greek lemma bank across Universal Dependencies part-of-speech categories.

4. Conclusion

This paper presented the initial construction of a lemma bank for Ancient Greek, developed as the core component of the Linking Greek knowledge base and explicitly inspired by the architecture of the LiLa project for Latin. By integrating data from heterogeneous lexical and semantic resources within a shared ontology, the lemma bank provides a unified, lemma-centric layer that supports interoperability across corpora, lexica, and NLP tools, while preserving resource-specific lemmatisation practices. The current version of the lemma bank is populated primarily from the Ancient Greek WordNet and the *Liddell–Scott–Jones* lexicon, combining semantic structure, part-of-speech information, and nominal gender assignment. Additional normalisation steps, including the rule-based correction of closed-class part-of-speech categories and harmonisation to the Universal Dependencies tag set, were introduced to improve consistency and computational usability.

Several directions for future research naturally follow from this work. First, the distinction between common nouns and proper nouns, in particular personal names, remains to be addressed and could benefit from the integration of gazetteers or named entity resources. Second, extending the coverage of dialectal and phonological variation, as well as refining the modelling of inflectional and derivational relations, would further enhance the descriptive power of the lemma bank. Finally, the integration of additional textual resources and the publication of the lemma bank as Linked Open Data will enable broader reuse and facilitate cross-linguistic linking within the wider LOD ecosystem.

5. Acknowledgements

This publication is based upon work from COST Action CA23147 GOBLIN – Global Network on Large-Scale, Cross-domain and Multilingual Open Knowledge Graphs, supported by COST (European Cooperation in Science and Technology, https://www.cost.eu).

6. Bibliographical References

References

David Bamman and Gregory Crane. 2011. The ancient greek and latin dependency treebanks. In *Language Technology for Cultural Heritage*, pages 79–98, Berlin, Heidelberg. Springer Berlin Heidelberg.

Tim Berners-Lee. 2006. Linked data – design issues.

Tim Berners-Lee, James Hendler, and Ora Lassila. 2001. Web semantic. *Scientific American*, 284(5):34–43.

Yuri Bizzoni, Federico Boschetti, Harry Diakoff, Riccardo Del Gratta, Monica Monachini, and Gregory Crane. 2014. The making of Ancient Greek WordNet. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1140–1147, Reykjavik, Iceland. European Language Resources Association (ELRA).

Giuseppe A. Celano. 2024. Opera graeca adnotata: Building a 34m+ token multilayer corpus for ancient greek.

Gregory Crane. 1991. Generating and parsing classical greek. *Literary and Linguistic Computing*, 6(4):243–245.

Michiel de Vaan. 2008. *Etymological Dictionary of Latin: and the other Italic Languages*. Brill, Leiden, The Netherlands.

Theodorus Fransen, Cormac Anderson, Sacha Beniamine, and Marco Passarotti. 2024. The MOLOR lemma bank: a new LLOD resource for Old Irish. In *Proceedings of the 9th Workshop on Linked Data in Linguistics @ LREC-COLING 2024*, pages 37–43, Torino, Italia. ELRA and ICCL.

Alek Keersmaekers, Wouter Mercelis, Colin Swaelens, and Toon Van Hal. 2019. Creating, enriching and valorizing treebanks of Ancient Greek. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 109–117, Paris, France. Association for Computational Linguistics.

Eleonora Litta, Marco Passarotti, Paolo Brasolin, Giovanni Moretti, Valerio Basile, Andrea Di Fabio,

and Cristina Bosco. 2024. The lemma bank of the LiITA knowledge base of interoperable resources for Italian. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 517–522, Pisa, Italy. CEUR Workshop Proceedings.

Marco Passarotti, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, and Rachele Sprugnoli. 2020. Interlinking through lemmas. the lexical collection of the lila knowledge base of linguistic resources for latin. *Studi e Saggi Linguistici*, 58(1):177–212.

Ineke Sluiter, Lucien van Beek, Ton Kessels, and Albert Rijksbaron. 2024. *Woordenboek Grieks/Nederlands*. Amsterdam University Press.

Thea Sommerschield, Yannis Assael, John Pavlopoulos, Vanessa Stefanak, Andrew Senior, Chris Dyer, John Bodel, Jonathan Prag, Ion Androutsopoulos, and Nando de Freitas. 2023. Machine learning for ancient languages: A survey. *Computational Linguistics*, 49(3):703–747.

Colin Swaelens, Pranaydeep Singh, Ilse de Vos, and Els Lefever. 2024. Lemmatisation of medieval Greek: Against the limits of transformer's capabilities? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10293–10302, Torino, Italia. ELRA and ICCL.

A. Vatri and B. McGillivray. 2018. The diorisis ancient greek corpus: Linguistics and literature. *Research Data Journal for the Humanities and Social Sciences*, 3(1):55 – 65.

Across Generations: A Comparative Analysis of NER for Latin Inscriptions from Classical Machine Learning to LLMs

Wenhui Cui, Phillip Benjamin Ströbel
University of Zurich
wenhui.cui@uzh.ch, phillip.stroebel@uzh.ch

Abstract

Latin epigraphic texts are a challenging type of historical data for natural language processing (NLP). They are often fragmentary, contain inconsistent spelling, and follow complex Roman naming conventions. This paper investigates Named Entity Recognition (NER) for this domain by comparing several approaches, including feature-based Support Vector Machines, neural models such as BiLSTM and TreeLSTM, pre-trained language models like LatinBERT, fine-tuned Transformer models based on BERT, and large language models used with prompting and supervised fine-tuning. We introduce a manually annotated dataset of 1,000 inscriptions from the *Epigraphik-Datenbank Clauss-Slaby*, labelled with a fine-grained BIO scheme that captures the internal structure of Roman personal names. Results show that the fine-tuned BERT model achieves the highest performance, with a weighted F1 score of 91.1% and a macro F1 of 68.7%, and clearly outperforms other methods. Additional linguistic features, such as part-of-speech tags and dependency information, yield only limited improvements, likely due to the irregular nature of inscriptional texts. This work provides a new benchmark for NER on Latin inscriptions and offers practical insights into applying modern NLP techniques to historical, non-standardised language.

Keywords: Latin NLP, Named Entity Recognition, Epigraphy, Low-resource Languages, Machine Learning, LLMs

1. Introduction

Named entity recognition (NER) is a key natural language processing (NLP) task that identifies references to persons, locations, and other entities (Munnangi, 2024). Due to the high density of named entities in historical inscriptions, these texts constitute an important linguistic resource for the study of people, society, and geography in the ancient world (Bodel, 2012).

In recent years, there has been growing interest within the Digital Humanities and NLP communities in applying advanced language models to historical and low-resource languages (Manjavacas and Fonteyn, 2022). While NER has achieved strong performance on modern high-resource languages using neural and Transformer-based models, increasing attention is now being directed toward extending these approaches to historical languages (Brandsen et al., 2021; Schweter and März, 2020; Konle and Jannidis, 2020). As epigraphic collections become increasingly digitised, there is growing interest in applying NLP to these materials. However, unlike well-edited literary texts, inscriptions are often fragmentary, contain inconsistent spelling, and follow conventionalised formulaic patterns, which makes automatic processing difficult (Heřmánková et al., 2021).

Latin inscriptions are particularly challenging because Roman personal names consist of multiple components, such as *praenomen* (EN *first name*), *nomen gentilicium* (EN *family/clan name*), and *cognomen* (EN *nickname*),[1] which constitute the so-called *tria nomina* and may appear in abbreviated or incomplete forms (Salway, 1994). E.g., in the name *M. Tullius Cicero*, we find the abbreviation for the *praenomen* ***Marcus***, followed by the *nomen gentilicium* ***Tullius*** and the *cognomen* ***Cicero***. Inscriptions also include place names and institutional titles embedded in fixed expressions. These domain-specific features require a more detailed annotation scheme than standard person and location labels.

In this paper, we address three research questions: (RQ1) How do traditional feature-based models, neural sequence labellers, Transformer encoders, and LLMs compare when applied to NER on Latin inscriptions? (RQ2) To what extent do additional linguistic features, such as POS tags and dependency parses, improve performance on highly irregular inscriptional texts? (RQ3) How competitive are large language models, used via prompting or supervised fine-tuning, compared to task-specific fine-tuned Transformers in this low-resource, historical setting?

The contributions of this paper are the following:

- We develop and evaluate a NER system for Latin inscriptions using a manually annotated dataset of 1,000 inscriptions from the

[1] The translations of the different parts of the name are approximate in the sense that these names were not necessarily used how we would use them nowadays. Moreover, the usage of the different names has changed over the centuries (see, e.g., Ayer (2014)).

Proceedings of LT4HALA 2026 @ LREC 2026, pages 112–124.
11 May 2026. *©ELRA Language Resources Association (ELRA), 2026*

Epigraphik-Datenbank Clauss-Slaby (EDCS, Clauss et al. (1985)) as a gold standard.[2]

- We apply existing Latin NER models, especially LatinCy[3] (Burns, 2023), and test large language models (LLMs) in a zero-shot setting to establish baselines to compete against.
- We trained the following models: (1) A linear SVM baseline using handcrafted features, including prefixes, suffixes, POS tags, and dependency relations, and (2) neural models, including a TreeLSTM for hierarchical structure and a BiLSTM CNN model for morphological and contextual features. We fine-tuned (3) Transformer models, i.e, multilingual BERT and a Latin-specific RoBERTa, both of which use token classification for BIO tagging. We (4) applied LLMs using few-shot prompting and supervised fine-tuning, combined with CRF and rule-based post-processing.
- We test all models on a fixed stratified dataset using token-level precision, recall, and micro/macro F1, enabling a systematic comparison across traditional, neural, Transformer, and LLM approaches.[4]

2. Related Work

The development of NER for Latin has been relatively slow, particularly outside literary texts (Erdmann et al., 2016). Early research was mostly part of broader projects like the Classical Language Toolkit (CLTK) (Johnson et al., 2021) and LatinCy (Burns, 2023). While these initiatives focused on establishing core linguistic processing capabilities, most notably part-of-speech (POS) tagging, they created the essential computational infrastructure that later enabled more advanced tasks such as NER. As noted by Erdmann et al. (2016), who achieve a 90% F-score on literary data, performance differences are expected when applying NER tools to texts from different Latin genres and linguistic styles.

Building upon this infrastructure, early dedicated NER efforts for Latin typically used Conditional Random Field (CRF) models with handcrafted linguistic features such as word shapes and POS tags, as exemplified by the recognition of named entities in Medieval Latin charters (Chastang et al., 2021). While these methods achieved high performance on literary texts, with F1 scores up to 0.95 for exact matches, they have not been applied to inscriptional material. Given the fragmented and formulaic nature of inscriptions, the frequent use of abbreviations, and variations in spelling, feature-based models are expected to face substantial challenges when applied to epigraphic data.

Later, the field shifted towards machine learning models tailored for sequence labelling tasks. Neural architectures such as the BiLSTM-CRF model enable models to learn complex patterns, including contextual cues and morphological variation, directly from data, reducing reliance on handcrafted linguistic features (Lample et al., 2016). This class of models has become a standard approach in NER, achieving strong performance across multiple languages and related domains. Research has shown that explicitly incorporating linguistic structures, such as syntactic constituents, into neural models can improve NER by enabling the model to leverage hierarchical phrase structure information and semantic features from syntactic trees (Li et al., 2017). Building on these ideas, researchers have also adapted sequence labelling approaches to Latin, as in the *Herodotos Project*, which developed NER systems for Classical Latin and ancient Greek texts that accurately identify groups, persons, and places, illustrating the feasibility of machine learning–based entity recognition in historical languages (Erdmann et al., 2023). Their model achieved micro-F1 scores of 0.99 on in-domain data and only slightly lower scores on out-of-domain data (i.e., data the model was not trained on).

Following the introduction of Transformer architectures (Devlin et al., 2019), Latin-specific models such as LatinBERT (Bamman and Burns, 2020) demonstrated that contextualised representations can effectively capture linguistic patterns in Latin. Subsequent work explored other Transformer variants, including ELECTRA-based models for lemmatisation and POS tagging on Classical Latin corpora (Mercelis and Keersmaekers, 2022). Building on these foundations, researchers have adapted Transformer approaches to historical language scenarios, including NER (Beersmans et al., 2023). Also, multilingual models such as mBERT and XLM-RoBERTa have been fine-tuned for NER on medieval charter corpora in Latin, French, and Spanish, achieving strong cross-lingual performance without significant degradation compared to monolingual baselines (Torres Aguilar, 2022). Another approach uses cross-lingual annotation projection to transfer NER labels from modern languages to low-resource classical languages, such as Latin and ancient Greek, leveraging parallel corpora and neural word alignment to expand training data (Yousef et al., 2023).

[2] See https://edcs.hist.uzh.ch.

[3] Available within *spaCy*, see https://spacy.io and for LatinCy especially https://spacy.io/universe/project/latincy.

[4] The code and data of this work are available at https://github.com/Wenhui620/latin-inscriptions-ner.

In recent years, LLMs have made substantial progress in NLP due to their scale and contextual learning ability, and have increasingly been applied to NER (Brown et al., 2020). Because NER is a sequence labelling task while LLMs are generative models, two main adaptation strategies have been proposed: few-shot prompting and fine-tuning (Ji et al., 2025). Few-shot prompting incorporates task descriptions and a small number of annotated examples into the input, allowing models to infer labelling patterns through in-context learning without parameter updates (Federiakin et al., 2024). Approaches such as GPT-NER reformulate sequence labelling as text generation and show strong performance in low-resource settings (Wang et al., 2023). In contrast, fine-tuning updates model parameters on task-specific data to improve domain accuracy and consistency, with parameter-efficient methods reducing computational cost (Zhang et al., 2025).

For Latin and other historical languages, NER remains challenging due to limited annotated data and evolving naming conventions (Ehrmann et al., 2021). Recent work shows that LLMs can outperform traditional NLP frameworks such as spaCy and flair on historical NER, achieving gains of 7 to 22 % F1 through context-aware prompting, although increasing the number of examples beyond a certain point brings little improvement (Hiltmann et al., 2025). However, zero-shot experiments on historical sources reveal persistent difficulties, including inconsistent annotation, entity complexity, multilingual code-switching, and sensitivity to prompt design (González-Gallardo et al., 2023). Because naming rules change over time and annotated data remain scarce, historical NER remains difficult (Ehrmann et al., 2021). Consequently, both few-shot prompting and fine-tuning should be viewed as complementary to traditional supervised methods.

This overview of related work shows several desiderata for NER in Latin. First, none of the approaches mentioned addressed NER in Latin inscriptions. Secondly, there is only a distinction between names and other categories, where names *per se* are not annotated with more fine-grained categories. Thirdly, the problem of Latin being a low-resource language is aggravated in inscriptions, which constitute a specific, highly formulaic subgenre. Fourthly, there is no comprehensive comparison of machine learning methods applied to inscriptions, especially one that includes traditional methods, which, due to the hype generated by LLMs, are often overlooked. Our paper addresses all of these issues.

3. Method

Figure 2 shows the experimental framework used in this study. The workflow illustrates the systematic comparison of five distinct modelling paradigms applied to Latin inscription NER.

3.1. Dataset

The dataset used in this research consists of 1,000 Latin inscriptions manually collected from the EDCS. The selection process focused on gathering a representative sample of Roman names spanning from the early Republican to Imperial periods and distributed across multiple provinces throughout the Roman world. Hence, we compiled the dataset as a stratified sample, accounting for metadata such as date, classification, province, and material. This reduced the number of inscriptions from which we sampled from over 500,000 to approximately 200,000.[5] Each inscription was cleaned to remove modern symbols[6] used to mark missing, erased or supplemented text while keeping the original structure of the text. We applied a fine-grained BIO tagging schema to these texts to identify 10 specific entity types.

The annotation is inspired by existing guidelines for named entity annotation in Classics (e.g., Romanello and Najem-Meyer (2022)), particularly in its treatment of entity categories and boundary definitions, while introducing additional fine-grained labels tailored to the structure of Latin personal names in inscriptions. It includes several tags for fine-grained personal name elements: *PERS:PRAE* (e.g., *Publius*) for the first name, *PERS:NOMEN* (e.g., *Iulius*) for the clan name, *PERS:COG* (e.g., *Cicero*) for the family name, and *PERS:AG* (e.g., *Africanus*) for an additional nickname. It also identifies *PERS:FILI* to mark family relations like a father's name (e.g., *Marci filii*, EN *son of Marcus*) and *PERS:TRIBE* (e.g., *Palatina*) for the voting tribe. Beyond personal names, the dataset marks *PERS:TITLE* (e.g., *Augustus*) when a title is part of a name, and *TITLE* (e.g., *pontifex*) for independent positions such as priest or veteran. For coarse-grained annotation, the general *PERS* tag is used, which spans all name parts of a single person. Geographic information is captured by the

[5] We are aware that we limited the selection to certain factors, which might not result in a fully representative sample of the overall data, but we favoured this approach since it is 1) reproducible and 2) still represents a subset of the corpus. The selected inscriptions span the early Republican to the late Imperial period (roughly 3rd century BCE to 4th century CE) and cover provinces across the Italian peninsula, the western provinces (Gallia, Hispania, Africa), and the eastern Mediterranean, reflecting the geographic spread of the EDCS corpus.

[6] Added during the editing process.

LOC tag for cities or provinces. To ensure the quality of the manual annotation, the data was checked for consistency across different inscription types.

We divided the dataset into three parts for the experiments: 700 inscriptions for training, 150 for validation, and 150 for testing. Again, stratified sampling was used to ensure that each set had a similar mix of entity types. This data organisation enables a reliable evaluation of how different machine learning and transformer-based models handle the specific patterns found in Latin inscriptions.

3.2. Inter-Annotator Agreement

Category	A1 vs A2	A1 vs A3	A2 vs A3
AG	0.691	0.000	0.087
COG	0.789	0.685	0.724
FILI	0.556	0.528	0.645
LOC	0.333	0.000	0.000
NOMEN	0.742	0.835	0.755
PERS	0.303	0.214	0.391
PRAE	0.738	0.802	0.855
TITLE	0.510	0.382	0.419
TRIB	0.000	0.000	0.875
Micro-average	0.618	0.540	0.609
Macro-average	0.518	0.383	0.528

Table 1: Pairwise span-level F1-scores per entity category and global averages. In each pair (X vs Y), annotator X serves as the reference (proxy gold standard) against which Y is evaluated.

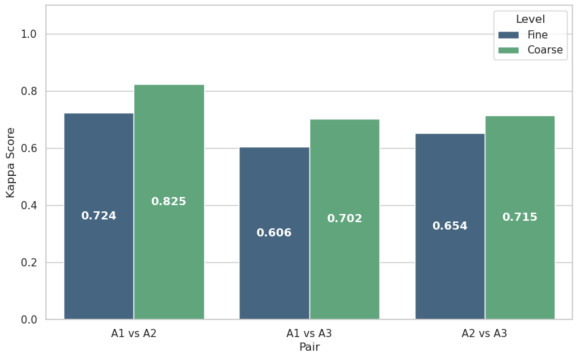


Figure 1: Pairwise comparison between annotators based on Cohen's κ.

Three annotators independently labelled 129 inscriptions from the test set to evaluate inter-annotator agreement. The annotation team comprised a Latin epigraphy expert, a historian in training, and a student responsible for data preparation and annotation coordination. Table 1 shows pairwise F1 scores, while Figure 1 shows Cohen's κ. The overall agreement, according to Fleiss' κ, is 0.658, indicating moderate agreement. However, the results generally show high overlap in annotations of the *tria nomina*, whereas the other parts of the names are more difficult to identify. This is largely due to the limited contextual information available in inscriptions. In many cases, especially in fragmentary or formulaic texts, additional name elements, such as filiations, tribal affiliations, and titles, lack sufficient context to be interpreted reliably. As a result, annotators may differ in both boundary detection and category assignment, leading to lower agreement.

3.3. Models

This study compares several modelling paradigms to evaluate their effectiveness in recognising entities in Latin inscriptions, ranging from traditional machine learning to modern neural and generative approaches.

Existing Latin NER Models We first evaluate an existing Latin NLP system as an external reference point. Specifically, we test the LatinCy pipeline in a zero-shot setting to establish a modern baseline. This model was trained on a mixture of Latin treebanks, web corpora, and historical text collections, but not on epigraphic material. To ensure alignment with our gold dataset, we construct spaCy `Doc` objects directly from the gold tokenisation, preventing the model from re-tokenising the text. The span-level entity predictions produced by LatinCy are converted into token-level BIO tags. Since LatinCy uses a coarse label set, its entity types are mapped to three evaluation categories: PERS, TITLE, and LOC, whereas all other predictions are mapped to O. On the gold-standard side, our fine-grained person name subtypes are collapsed into their coarse classes while preserving the BIO prefixes. Performance is measured using token-level Precision, Recall, and F1 score.

Support Vector Machine As a standard baseline, we use a linear SVM classifier implemented in `scikit-learn`.[7] The task is formulated as token-level BIO classification. We compare two feature settings. The baseline feature set includes the lowercase token form, word shape, prefixes, suffixes, and the previous and next tokens. The linguistic feature set is extended by adding POS tags, dependency relations, and the lowercase form of the syntactic head, which are obtained from the LatinCy parser. All features are converted into sparse vectors using `DictVectorizer`. To address class imbalance, class weights are set to `balanced`. The classifier is further wrapped with

[7] See https://scikit-learn.org/stable.

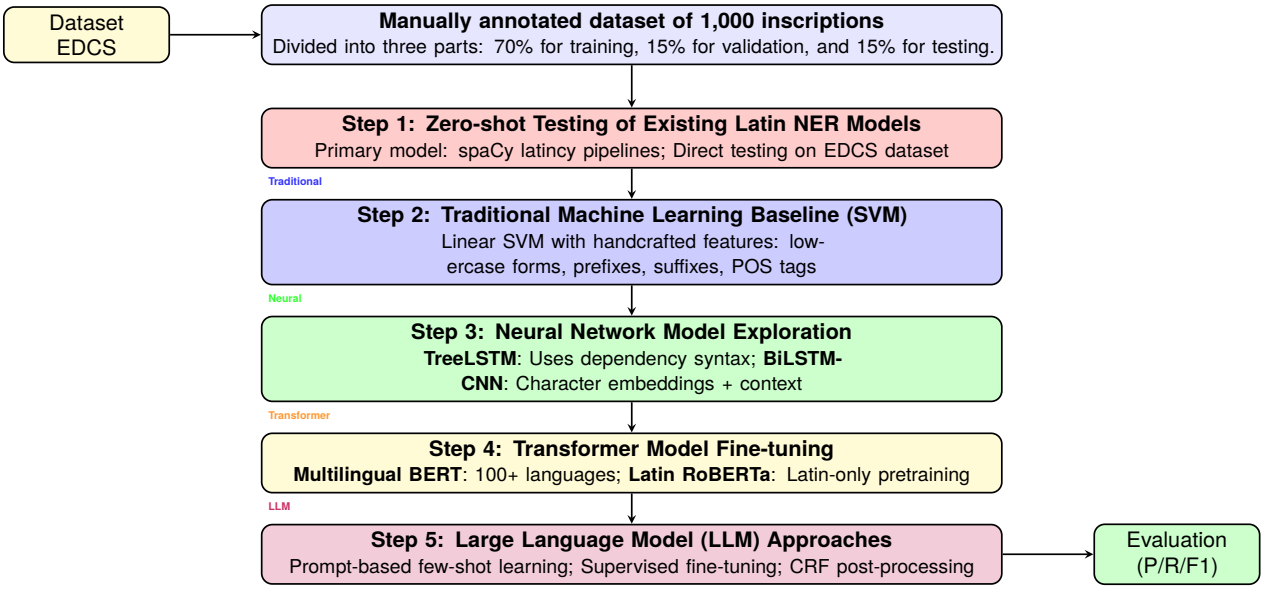


Figure 2: Overview of the experimental framework.

`CalibratedClassifierCV` using sigmoid calibration on the validation set to produce more reliable probability estimates.

Neural Models We then explore two neural network architectures implemented in PyTorch. The first is a TreeLSTM that incorporates dependency structure by aggregating information from child to parent nodes; its input representation concatenates word embeddings, POS and dependency embeddings, and character-level CNN embeddings, which are jointly learned during training, followed by a hidden size of 256 and a linear layer for BIO tag prediction. The second is a BiLSTM-CNN hybrid that processes tokens sequentially; each token representation combines word, character CNN, prefix, suffix, and surrounding token embeddings, which are fed into a bidirectional LSTM with hidden size 256 and a linear classification layer. During training, dropout with a rate of 0.5 is applied to reduce overfitting, and both models are trained using the Adam optimiser with a learning rate of 0.001 for 20 epochs and a batch size of 32; mini-batch gradient descent is employed, and padding tokens are ignored in the loss computation.

BERT-based Models For Transformer-based methods, we fine-tune two pretrained encoders for token classification. The first is `bert-base-multilingual-cased`,[8] which was pretrained on more than 100 languages, including Latin. The second is a Latin-specific RoBERTa-based model pretrained only on Latin corpora.[9] Both encoders are further fine-tuned on our inscription dataset as token-classification models. For multilingual BERT, we use the Hugging Face Transformers library with a standard token classification head and fine-tune for three epochs with a batch size of 16 and a learning rate of 5e-5. For the Latin RoBERTa model, we use the spaCy training framework with a Transformer-plus-NER pipeline, converting the data to spaCy's binary format prior to training.

LLMs Finally, we evaluate LLMs under both prompt-based and fine-tuned settings. In the prompt-based experiments, GPT-4.1-mini[10], Gemini-2.5-flash[11], and Claude-Sonnet-4[12] are tested with zero-shot and few-shot prompting (using 3-, 5-, and 10-shot prompts). The prompts define the BIO tagging task and require the models to output JSON objects containing aligned token and tag sequences.[13] For long inscriptions, the input is split into shorter segments, and the predictions are concatenated. After inference, a CRF model trained on the gold data is applied as a post-processing step, and rule-based constraints enforce common patterns in Roman names.

[8] See https://huggingface.co/google-bert/bert-base-multilingual-cased.

[9] See https://huggingface.co/pstroe/roberta-base-latin-cased.

[10] See https://developers.openai.com/api/docs/models/gpt-4.1-mini.

[11] See https://ai.google.dev/gemini-api/docs/models/gemini-2.5-flash.

[12] https://www.anthropic.com/news/claude-sonnet-4-5.

[13] The prompts are available in the GitHub repository accompanying this paper (see Footnote 4).

In the fine-tuning setting, Gemini and GPT models are further trained on the annotated corpus using supervised learning. The training data are formatted as JSONL files containing token and BIO tag sequences. Gemini fine-tuning is conducted via the Vertex AI interface using the default hyperparameters, such as learning rate, batch size, and optimiser type, which are platform-defined and do not expose fixed numeric values. GPT fine-tuning is performed for 3 epochs with a batch size of 1 and a learning rate multiplier of 2. All models are evaluated on the same fixed test set.

3.4. Evaluation

To evaluate the models' performance on Latin inscriptions, we use standard metrics, including Precision, Recall, and weighted & macro F1 scores. Because the dataset employs a fine-grained BIO tagging schema and exhibits an uneven distribution of entity types, we primarily focus on the weighted F1 score. This ensures that the performance on rare tags is reflected fairly in the final results.[14] For LLM outputs, we apply a format check to convert the generated text into standard BIO format so that it can be compared directly with the other models under the same criteria.

4. Results and Discussion

This section describes the results of our experiments and the subsequent error analysis. Table 2 summarises the performance, specifically weighted F1, macro F1, and accuracy, for all models evaluated in this study.

4.1. Overall Performance Comparison

Clear performance differences emerge across model families. Overall, the fine-tuned multilingual BERT reaches 91.1% weighted F1 and 68.7% macro F1, indicating strong performance even when averaging over all entity types rather than weighting by frequency. Fine-tuned transformer models perform best overall. The fine-tuned multilingual BERT exhibits strong performance on frequent name categories such as *praenomen*, *nomen*, and *filiation* markers, indicating effective adaptation to abbreviation patterns and orthographic variation.

By contrast, zero-shot models such as LatinCy perform substantially worse, often failing to detect complete entity spans. The Latin RoBERTa model, although pretrained exclusively on Latin corpora, achieves a weighted F1 of 88.9% and a macro F1 of 59.9% after fine-tuning on our dataset, indicating that Latin-specific pretraining helps but still falls short of the multilingual BERT baseline.

LLMs show higher variability. Performance improves with more examples, and 10-shot prompting enables Claude Sonnet to reach 90.75% accuracy, close to that of fine-tuned BERT. Fine-tuned LLMs such as Gemini and GPT achieve stable accuracy above 90%, but their macro F1 remains limited due to errors on rare labels and boundary detection.

Traditional SVM models remain strong baselines, with weighted F1 scores of 88.40% and 87.96%, indicating that surface features capture much of the signal. The BiLSTM also performs competitively, with 88.60%, whereas the TreeLSTM performs substantially worse, at 73.45%, likely due to unreliable dependency parses.[15]

Overall, fine-tuned transformers perform best, followed by fine-tuned LLMs, pretrained models, SVM and BiLSTM, and finally TreeLSTM. Models that depend on accurate syntax or zero-shot reasoning are less reliable than those trained on annotated inscription data. Our results show that relying on the latest developments, i.e., LLMs, does not yield the best performance, as LLMs are outperformed by a pre-trained BERT. These representations appear to contain information valuable for the NER task. This is confirmed by the almost equally strong results of LatinBERT.

Recall values follow a broadly similar pattern to weighted F1, though with notable divergences for specific entity classes. Fine-tuned BERT achieves the highest overall weighted recall, closely matched by the fine-tuned LLMs. Among few-shot LLMs, Claude with 10-shot prompting is competitive at the weighted level, but per-class recall reveals a striking weakness on the coarse PERS category, where it almost entirely fails to recover underspecified name spans, preferring to assign a fine-grained subtype or nothing at all. LOC recall is low across all models and architectures, underscoring that location recognition remains the most challenging category. These recall figures are particularly relevant for historical NER, where missed entities may represent irretrievable losses of prosopographic or geographic information. Per-class recall values for all models are available in the project repository (see Footnote 4).

[14] It should be noted that the weighted F1 score is computed over all token labels, including O, which represents the majority class. This inflates the absolute weighted F1 values relative to an entity-only metric. To provide a more entity-focused picture, we also report macro F1, which averages scores across entity categories without weighting by frequency and excludes the dominance of O-class tokens. Readers primarily interested in entity detection performance should therefore weight the macro F1 figures more heavily when comparing models.

[15] For future work, we plan on an evaluation of the parsing results by LatinCy on the inscriptions.

Model Category	Model Name	Weighted F1	Macro F1	Accuracy
Transformer (fine-tuned)	**BERT***	**0.911**	0.687	**0.912**
LLM (few-shot)	Claude-10shot*	0.904	**0.759**	0.908
LLM (fine-tuned)	GPT-fine-tuned	0.896	0.623	0.906
LLM (fine-tuned)	Gemini-fine-tuned*	0.901	0.610	0.905
Transformer (pretrained)	Latin-RoBERTa	0.889	0.599	0.891
Neural Network (sequential)	BiLSTM*	0.886	0.613	0.892
LLM (few-shot)	Gemini-10shot	0.883	0.622	0.890
Traditional (feature-based)	SVM-Linguistic*	0.884	0.625	0.890
Traditional (feature-based)	SVM-nonLinguistic	0.880	0.628	0.888
LLM (few-shot)	Claude-5shot	0.881	0.639	0.880
Neural Network (hierarchical)	TreeLSTM	0.735	0.459	0.746
LLM (baseline)	Claude-0shot*	0.799	0.507	0.814
LLM (baseline)	Gemini-0shot	0.794	0.457	0.803
LLM (baseline)	GPT-0shot	0.712	0.310	0.700
External Pipeline	LatinCy	0.580	0.240	0.580

Table 2: Overall performance of all models. The best performances per measure and the best overall model are in **bold**. The best models per category (feature-based, neural network, Transformer-based, LLM) are marked with an asterisk (*).

4.2. Fine-Grained Entity Performance

Label	BERT	SVM	BiLSTM	Cl-10s	Gem-FT
B-P:PRAE	*0.928*	**0.883**	**0.885**	**0.922**	**0.896**
B-P:NOMEN	0.881	0.789	0.814	*0.895*	0.828
B-P:COG	0.795	0.734	0.771	0.762	***0.813***
B-P:FILI	***0.965***	0.816	0.835	0.882	0.816
B-TITLE	0.711	*0.750*	0.624	0.723	0.714
B-LOC	*0.579*	0.333	0.240	0.480	0.419

Table 3: Fine-grained F1 scores across representative models. Best-performing models are in **bold**. Best performances per category are in *italics*.

As shown in Table 3, model performance differs considerably across entity types. For *praenomen* and *nomen*, supervised models achieve high scores. BERT reaches 92.82% F1 on *praenomen*, and Claude, with 10-shot prompting, performs similarly. For *nomen*, Claude 10-shot gives the best result, followed closely by BERT and the BiLSTM. SVM models perform slightly worse, especially on *nomen*, which shows greater lexical diversity.

Cognomen are more challenging because of their wider range of forms. Fine-tuned Gemini performs best in this category, followed by BERT. Feature-based and few-shot LLM models perform worse, indicating that this category benefits from stronger contextual modelling.

Filiation markers are easier for most systems. Their repetitive structure and fixed abbreviation patterns help all models. BERT achieves the highest score at 96.52%, and several other models also perform strongly.

Titles and location names are the most difficult categories. Titles appear in many syntactic positions and with varied forms. In this category, the SVM slightly outperforms other systems. Location names are sparse and highly variable in spelling. BERT performs best, but overall scores remain much lower than for personal name categories.

These results show that high-frequency and formulaic entity types, i.e., the *tria nomina*, are easier for all models, while rare and lexically diverse categories reveal clearer differences in model capacity and domain adaptation.

4.3. Error Analysis

This section analyses systematic discrepancies between model predictions and gold annotations. Using Claude-Sonnet-4 with 10-shot prompting as an example, Figure 3 provides a confusion matrix summarising category confusions.

Errors in Personal Names Labels such as *PERS:PRAE*, *PERS:NOMEN*, *PERS:COG*, *PERS:FILI*, and *PERS:AG* show high accuracy along the diagonal, indicating that the model broadly captures the semantic domain of Roman personal names. Nevertheless, fine-grained errors are substantial and reveal systematic difficulties in handling the internal structure of the *tria nomina* system.

The most prominent pattern is that *PERS:COG* functions as a default category, absorbing tokens

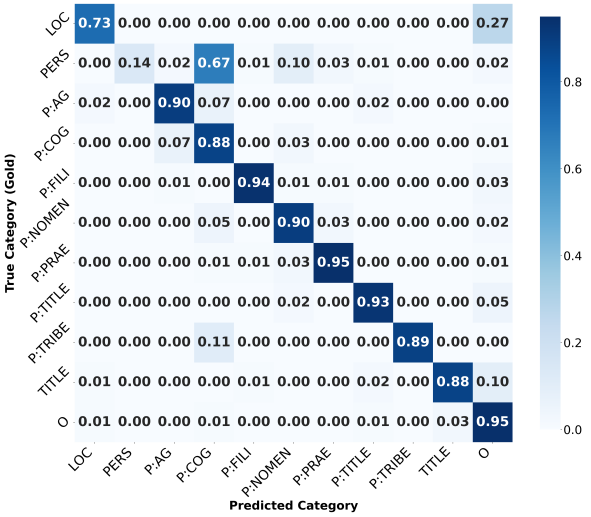


Figure 3: Semantic confusion matrix illustrating the misclassification patterns between entity labels.

from multiple related subtypes. Error rates for *PERS:COG* and the general *PERS* category are particularly high, reaching up to 67%, indicating frequent misassignment between the default *cognomen* label and the coarse-grained person category. Detailed analysis of the confusion matrix reveals specific bidirectional confusions. First, *PERS:COG* and *PERS:AG* exhibit mutual misclassifications (7% in both directions). This confusion is likely due to two factors: *agnomina* appear infrequently in the training data, and both categories occupy a similar structural position within the name. They follow the *nomen* and *cognomen* and serve as additional descriptive or honorific components. Second, *PERS:COG* and *PERS:TRIBE* show substantial bidirectional confusion (11%). This pattern reflects historical developments in Roman naming practices: over time, many tribal designations gradually became absorbed into the nomenclature, sometimes evolving into *nomina* or *cognomina*. As a result, some lexical forms can function as either tribal markers or name components, creating inherent ambiguity that challenges automatic classification.

Two further issues emerge with specific types of name components. First, filiation markers such as *filia*, *filius*, and *filiae* (forms of EN *daughter* or *son*), particularly in formulaic expressions (e.g., *Marci* ***filii***, EN son of Marcus), are annotated as *PERS:FILI* in the gold data because they encode family relationships. The model, however, often labels these tokens as *O*, treating them as ordinary nouns rather than structural components of the name. Second, with fragmented or abbreviated names, although the annotation scheme permits a general *PERS* label when the internal structure is unclear, the model often predicts more specific subcategories such as *PERS:NOMEN* or *PERS:COG* even for short or underspecified tokens. This behaviour reflects an implicit assumption of a complete *tria nomina* structure and overgeneralisation from canonical patterns rather than adaptation to underspecified contexts.

Overall, these patterns demonstrate that while the model captures the semantic notion of personhood, it struggles with the internal hierarchical and relational structure of Roman personal names. The concentration of errors in *PERS:COG* and the high confusion across both general and specific categories highlight the limitations of token-level classification for representing complex onomastic conventions, in which different components encode distinct types of social and familial information.

Errors in Titles and Locations The *TITLE* label is reserved for historically attested Roman administrative or military positions. Terms such as *consularis*, meaning consul, are annotated as *TITLE* when they appear in formal or official contexts, such as in inscriptions recording careers, offices, or honours. The model does not always follow this principle: while most *TITLE* tokens are correctly predicted, with 88% of instances on the diagonal, around 10% are misclassified as *O*. Such errors typically occur when titles resemble common nouns or appear in less explicit contexts, suggesting that the model sometimes treats these items as ordinary nouns rather than recognising their function as name components. Multiword title expressions that function as a single unit are occasionally fragmented, reflecting difficulties in modelling longer, formulaic sequences rather than misunderstandings of individual words.

Location recognition and segmentation exhibit similar challenges. Administrative units and province names are labelled as *LOC* when referring to concrete geopolitical entities. While a majority of *LOC* tokens are correctly identified, accounting for 73% of instances, a substantial proportion, approximately 27%, is misclassified as *O*. This pattern reflects a conservative prediction behaviour when the geographical meaning is ambiguous. Boundary inconsistencies arise when parts of multiword locations are labelled differently, resulting in split or shifted spans. Ethnic adjectives, such as *Gallorum*, meaning *Gallic*, further complicate recognition, as these forms are sometimes interpreted as locations due to their morphological similarity to place names, even when they function as descriptors of people. Overall, these patterns indicate that the model relies heavily on surface-form cues and struggles to distinguish between geographic reference and ethnic attribution in contexts without explicit signals.

4.4. Overall Discussion

Across models and entity types, consistent patterns emerge. Fine-tuned transformers achieve the best performance, particularly on frequent and regular personal name categories such as *PRAE* and *NOMEN*, underscoring the advantage of supervised learning for formulaic inscriptional data. Traditional feature-based models and pretrained transformers handle frequent patterns reasonably well, but struggle with rare or diverse entities. LLMs benefit from few-shot prompting but remain constrained by boundary detection and the scarcity of rare-label data.

Error analysis shows that most errors are systematic. Personal names are generally detected, but fine-grained subtypes are sometimes confused. Titles and locations are often missed or fragmented, especially in long formulaic expressions. Multiword expressions, abbreviations, and ambiguous morphological forms further increase difficulty. For example, ethnic adjectives are sometimes misclassified as locations due to surface similarity.

Overall, model performance depends on the extent to which inscriptional patterns match the learned categories. Frequent and regular entities are reliably recognised, while rare and complex cases remain challenging. These findings suggest that combining semantic modelling with targeted linguistic features and expanding annotated data can further improve Latin NER.

In summary, our experiments show that, in this noisy, low-resource, and highly formulaic historical setting, task-specific fine-tuned Transformer encoders still set the performance ceiling: none of the prompted or fine-tuned LLMs surpasses the fine-tuned BERT baseline, although they come close and exhibit complementary error profiles. This suggests that combining encoder-based models with LLMs, for example, via simple ensembling or post-editing, is a promising direction for future work rather than replacing supervised models outright.

5. Conclusion and Future Work

This study evaluated NER approaches for Latin inscriptions, including feature-based models, neural networks, and large language models. Fine-tuned transformers performed best, with BERT achieving a weighted F1 score of 91.1% and effectively capturing formulaic naming patterns, while few-shot Claude generalised well to rare entities (weighted F1 of 90.4%). Traditional SVMs that incorporated handcrafted linguistic features such as morphology, part-of-speech tags, and dependency relations provided a reasonable baseline, but their contribution was limited, as evidenced by the modest performance of syntax-dependent models such as TreeLSTM. Error analysis revealed systematic challenges: fine-grained subtypes of personal names were sometimes misassigned, titles and locations were often underrecognised or fragmented, and multiword or abbreviated expressions remained difficult to recognise. These patterns highlight the value of domain-specific fine-tuning and suggest that combining LLM semantics with selective linguistic cues may be more effective than relying solely on structural features. Limited overlap in errors between fine-tuned models and few-shot LLMs also indicates potential for ensemble methods. Future work could expand the annotated corpus to cover more regions, periods, and inscription types, thereby improving model generalisation and robustness. Moreover, a logical next step is to link names and locations to external knowledge sources. As such, this work opens the way to re-drawing the social network of the ancient Roman Empire as documented in inscriptions.

6. Limitations

Our study has several limitations. First, the inscription sample, although stratified by date, province, and material, is not fully representative of the entire EDCS corpus and underrepresents some regions and genres. Second, inter-annotator agreement is low for certain labels such as TRIB, TITLE, and LOC, which constrains the achievable upper bound and complicates the evaluation of these categories. Third, syntax-based models depend on LatinCy parses that are not optimised for fragmentary epigraphic Latin, which likely contributes to the poor performance of the TreeLSTM. Finally, we restrict ourselves to surface-level BIO tagging of flat entities; nested or discontinuous entities and document-level co-reference are left for future work.

7. Acknowledgements

We would like to thank Dr. Jens Bartels for his invaluable help in the annotation of the gold standard.

8. Bibliographical References

Meagan Ayer. 2014. *Allen and Greenough's New Latin Grammar for Schools and Colleges*. Dickinson College Commentaries, Carlisle, Pennsylvania.

David Bamman and Patrick J. Burns. 2020. Latin BERT: A Contextual Language Model for Classical Philology. ArXiv:2009.10053 [cs].

Marijke Beersmans, Evelien de Graaf, Tim Van de Cruys, and Margherita Fantoli. 2023. Training

and Evaluation of Named Entity Recognition Models for Classical Latin. In *Proceedings of the Ancient Language Processing Workshop*, pages 1–12, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

John Bodel. 2012. Epigraphic Evidence: Ancient History from Inscriptions.

Alex Brandsen, Suzan Verberne, Karsten Lambers, and Milco Wansleeben. 2021. Can BERT Dig It? – Named Entity Recognition for Information Retrieval in the Archaeology Domain.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners.

Patrick J. Burns. 2023. LatinCy: Synthetic Trained Pipelines for Latin NLP.

Pierre Chastang, Sergio Octavio Torres Aguilar, and Xavier Tannier. 2021. A Named Entity Recognition Model for Medieval Latin Charters. *Digital Humanities Quarterly*, 15(4).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2021. Named Entity Recognition and Classification in Historical Documents: A Survey.

Alexander Erdmann, Christopher Brown, Brian Joseph, Mark Janse, Petra Ajaka, Micha Elsner, and Marie-Catherine de Marneffe. 2016. Challenges and Solutions for Latin Named Entity Recognition. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 85–93, Osaka, Japan. The COLING 2016 Organizing Committee.

Alexander Erdmann, David Joseph Wrisley, Benjamin Allen, Christopher Brown, Sophie Cohen-Bodénès, Micha Elsner, Yukun Feng, Brian Joseph, Béatrice Joyeux-Prunel, and Marie-Catherine de Marneffe. 2023. Herodotos Project Latin NER Tagger Annotation.

Denis Federiakin, Dimitri Molerov, Olga Zlatkin-Troitschanskaia, and Andreas Maur. 2024. Prompt Engineering as a New 21st Century Skill. *Frontiers in Education*.

Carlos-Emiliano González-Gallardo, Emanuela Boros, Nancy Girdhar, Ahmed Hamdi, Jose G. Moreno, and Antoine Doucet. 2023. Yes but.. Can ChatGPT Identify Entities in Historical Documents?

Petra Heřmánková, Vojtěch Kaše, and Adela Sobotkova. 2021. Inscriptions as Data: Digital Epigraphy in Macro-historical Perspective. *Journal of Digital History*, 1(1).

Torsten Hiltmann, Martin Dröge, Nicole Dresselhaus, Till Grallert, Melanie Althage, Paul Bayer, Sophie Eckenstaler, Koray Mendi, Jascha Marijn Schmitz, Philipp Schneider, Wiebke Sczeponik, and Anica Skibba. 2025. NER4all or Context is All You Need: Using LLMs for low-effort, high-performance NER on historical texts. A humanities informed approach.

Lixia Ji, Yiping Dang, Yunlong Du, Wenzhao Gao, and Han Zhang. 2025. Nested Named Entity Recognition: A Survey of Latest Research. *Expert Systems*, 42(7):e70052.

Kyle P. Johnson, Patrick J. Burns, John Stewart, Todd Cook, Clément Besnier, and William J. B. Mattingly. 2021. The Classical Language Toolkit: An NLP framework for pre-modern languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 20–29, Online. Association for Computational Linguistics.

Leonard Konle and Fotis Jannidis. 2020. Domain and Task Adaptive Pretraining for Language Models. In *Workshop on Computational Humanities Research*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition.

Peng-Hsuan Li, Ruo-Ping Dong, Yu-Siang Wang, Ju-Chieh Chou, and Wei-Yun Ma. 2017. Leveraging Linguistic Structures for Named Entity Recognition with Bidirectional Recursive Neural Networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2664–2669, Copenhagen, Denmark. Association for Computational Linguistics.

Enrique Manjavacas and Lauren Fonteyn. 2022. Adapting vs. Pre-training Language Models for

Historical Languages. *Journal of Data Mining & Digital Humanities*, NLP4DH(Digital humanities in languages). Publisher: Episciences.org.

Wouter Mercelis and Alek Keersmaekers. 2022. An ELECTRA Model for Latin Token Tagging Tasks. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 189–192, Marseille, France. European Language Resources Association.

Monica Munnangi. 2024. A Brief History of Named Entity Recognition.

Matteo Romanello and Sven Najem-Meyer. 2022. Guidelines for the annotation of named entities in the domain of classics. Developed in the context of the Ajax Multi-Commentary project and used for the CLEF HIPE 2022 shared task.

Benet Salway. 1994. What's in a Name? A Survey of Roman Onomastic Practice from c. 700 B.C. to A.D. 700. *Journal of Roman Studies*, 84:124–145.

Stefan Schweter and Luisa März. 2020. Triple E - Effective Ensembling of Embeddings and Language Models for NER of Historical German. In *Conference and Labs of the Evaluation Forum*.

Sergio Torres Aguilar. 2022. Multilingual Named Entity Recognition for Medieval Charters Using Stacked Embeddings and Bert-based Models. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 119–128, Marseille, France. European Language Resources Association.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. GPT-NER: Named Entity Recognition via Large Language Models.

Tariq Yousef, Chiara Palladino, Gerhard Heyer, and Stefan Jänicke. 2023. Named Entity Annotation Projection Applied to Classical Languages. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 175–182, Dubrovnik, Croatia. Association for Computational Linguistics.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2025. Instruction Tuning for Large Language Models: A Survey.

9. Language Resource References

Clauss et al. 1985. *Epigraphic Database Clauss Slaby (EDCS)*. Epigraphik-Datenbank Clauss/Slaby Project, University of Zurich; Catholic University of Eichstätt-Ingolstadt. Online database of Latin inscriptions.

A. Prompt for zero-shot experiments

B. Prompt for few-shot experiments

```
You are a Latin epigraphy Named Entity Recognition (NER) expert.
Your task is to assign BIO tags to each token of Latin funerary
inscriptions.

Use only the following entity tags:
- B-PERS:PRAE - Praenomen (personal name)
- B-PERS:NOMEN - Nomen (clan/family name)
- B-PERS:COG - Cognomen (family branch or nickname)
- B-PERS:FILI - Filiational terms (e.g., filius, libertus)
- B-PERS:AG - Agnomen (honorific)
- B-PERS:TITLE - Personal title like consul, pontifex
- B-TITLE - Official state/military/religious title
- B-LOC - Geographical names only (e.g., Roma, Tiberis)
- I-PERS, I-PERS:AG, I-PERS:TITLE - Inside-tag variants
- B-PERS - Only use this if:
    - The token is a standalone name, and
    - It cannot be confidently classified as PRAE, NOMEN, or COG.
    - Do NOT use B-PERS for names that belong to known Roman name
    structures.

Use the BIO format:
- "B-" means beginning of an entity
- "I-" means continuation
- "O" means not an entity

Output Format:
Return your output as exactly one valid JSON object, on a single line,
structured as:
{"tokens": [...], "tags": [...]}

Constraints:
- The number of tags MUST match exactly the number of tokens.
- Do not include explanations or comments.
- Do not output Markdown, ellipses, or formatting.
- Only return the final JSON.

You will be given a list of tokens:
<tokens>
{{TOKENS}}
</tokens>

Label each token with its BIO tag. Be strict with format and tag rules.
```

Figure 4: Zero-shot prompt used for LLM-based Latin NER.

```
You are a Latin epigraphy Named Entity Recognition (NER) expert.
Your task is to assign BIO tags to each token of Latin funerary
inscriptions.

Use only the following entity tags:
- B-PERS:PRAE - Praenomen (personal name)
- B-PERS:NOMEN - Nomen (clan/family name)
- B-PERS:COG - Cognomen (family branch or nickname)
- B-PERS:FILI - Filiational terms (e.g., filius, libertus)
- B-PERS:AG - Agnomen (honorific)
- B-PERS:TITLE - Personal title like consul, pontifex
- B-TITLE - Official state/military/religious title
- B-LOC - Geographical names only (e.g., Roma, Tiberis)
- I-PERS, I-PERS:AG, I-PERS:TITLE - Inside-tag variants
- B-PERS - Only use this if:
  - The token is a standalone name, and
  - It cannot be confidently classified as PRAE, NOMEN, or COG.
  - Do NOT use B-PERS for names that belong to known Roman name
  structures.

Use the BIO format:
- "B-" means beginning of an entity
- "I-" means continuation
- "O" means not an entity

Output Format:
Return your output as exactly one valid JSON object, on a single line,
structured as:
{"tokens": [...], "tags": [...]}

Constraints:
- The number of tags **must match exactly** the number of tokens.
- Do not include explanations or comments.
- Do not output Markdown, ellipses, or formatting.
- Only return the final JSON.

Here are some examples of correctly labeled Latin inscriptions in the
JSON line format:
{"tokens": ["qui", "et", "gni", "in", "pace", "positus"], "tags": ["O",
"O", "O", "O", "O", "O"]}
{"tokens": ["Dis", "Manibus", "Anthusae", "HiYppolytus", "coniugi",
"bene", "merenti", "fecit"], "tags": ["O", "O", "B-PERS", "O", "O",
"O", "O", "O"]}
{"tokens": ["Caius", "Vibius", "Polycarpus", "Caius", "Vibius",
"Dorus", "Halus", "Tiberi", "Claudi", "Caesaris", "aedituus", "de",
"aede", "Iovis", "porticus", "Octaviae"], "tags": ["B-PERS:PRAE",
"B-PERS:NOMEN", "B-PERS:COG", "B-PERS:PRAE", "B-PERS:NOMEN",
"B-PERS:COG", "B-PERS:AG", "B-PERS:PRAE", "B-PERS:NOMEN", "B-PERS:COG",
"O", "O", "O", "O", "O", "O"]}

Label each token with its BIO tag. Be strict with format and tag rules.
```

Figure 5: Three-shot prompt used for LLM-based Latin NER.

POS Tagging with Generative LLMs for Historical Germanic Low-Resource Languages: An Evaluation Against Fine-Tuned BERT

Irene Miani[1], Gregory Darwin[2], Sara Stymne[3]
Department of Linguistics and Philology[1, 3], Department of English[2]
Uppsala University
(irene.miani[1], sara.stymne[3])@lingfil.uu.se, gregory.darwin@engelska.uu.se[2]

Abstract

Part-of-Speech (POS) tagging is a fundamental task in Natural Language Processing, yet its performance on historical low-resource languages is still underexplored, particularly in the context of large generative models. While recent studies have demonstrated strong results for Large Language Models (LLMs) on modern languages and contemporary low-resource settings, their effectiveness for historical varieties remains unclear. Moreover, genre-specific structural variation, which may substantially affect tagging performance, has received limited attention. This study evaluates the zero- and few-shot POS tagging performance of two generative models on four historical Germanic low-resource languages across two literary genres. Their performance is benchmarked against fine-tuned BERT models. To contextualize the performance on historical data, the models are also evaluated on two modern languages. The results show that fine-tuned encoder models consistently outperform generative models across all settings. The performance of the LLMs on historical languages is substantially lower compared to that on modern languages, suggesting limited representation of these varieties in pretraining data. Furthermore, error analysis reveals structural output inconsistencies in LLM predictions that require additional post-processing. These findings highlight the limitations of zero- and few-shot generative models for historical low-resource POS tagging and underline the importance of task-specific fine-tuning.

Keywords: POS tagging, Historical low-resource languages, Large Language Models

1. Introduction

Part-of-Speech (POS) tagging, a fundamental Natural Language Processing (NLP) task, has been extensively investigated over the years. While Large Language Models (LLMs) have demonstrated remarkable abilities across various NLP tasks (Qin et al., 2025), including strong performance on POS tagging for modern languages (Machado and Ruiz, 2024), and promising results for contemporary low-resource languages (Chang et al., 2024), their effectiveness on historical low-resource languages remains unexplored.

Many historical low-resource languages lack reliable NLP tools, or their accuracy is uncertain. This hinders the investigation of several languages that could benefit from the implementation of these tools. Additionally, existing LLM research has primarily concentrated on modern low-resource languages or higher-level tasks, often treating fundamental problems like POS tagging as solved, pushing these languages and their challenges even more into the background. Furthermore, recent studies have highlighted the influence of genre-specific structural patterns on model performance, which has been largely overlooked, despite evidence that low-resource genres can yield degraded results (Miani et al., 2026).

This study evaluates zero- and few-shot POS tagging performance of two generative models—Qwen2.5-3B-Instruct (Qwen Team, 2024), and Apertus-8B-Instruct-2509 (Apertus et al., 2025)—on four historical Germanic low-resource languages. Performance is assessed across both poetry and prose genres and benchmarked against a fine-tuned BERT baseline. To contextualize the performance on historical data, results are further compared with performance on two modern Germanic languages. The results revealed that fine-tuned encoder models consistently outperformed generative models across all languages and genres. The performance of the LLMs was substantially lower for the historical languages compared to modern ones, suggesting limited linguistic representation of these languages in the pretraining data. Additionally, error analysis highlighted structural inconsistencies in the LLM outputs, which required extra post-processing. Overall, the study demonstrated the superior effectiveness of fine-tuned encoder models for tasks such as POS tagging in historical low-resource languages. In contrast, generative models performed poorly on these languages, deviating from their generally strong performance trends observed in modern languages or with other tasks, where they showed comparatively good results.[1]

[1]All code used in the experiments is publicly available at https://github.com/irenemiani/historical-germanic.

Proceedings of LT4HALA 2026 @ LREC 2026, pages 125–138.
11 May 2026. *©ELRA Language Resources Association (ELRA), 2026*

2. Related Work

The abilities of LLMs on POS tagging have been investigated in several studies (Blevins et al., 2023; Lai et al., 2023; Chang et al., 2023), primarily as part of broader surveys examining the linguistic knowledge of generative models. Other studies focused only on POS tagging, such as Machado and Ruiz (2024), who focused on evaluating three LLMs on the Universal Dependencies POS tagging for Brazilian Portuguese using few-shot prompting with only ten examples. Their work stressed the abilities of multilingual LLMs to leverage prior linguistic knowledge to perform POS tagging in zero- and few-shot settings, outperforming monolingual models.

A smaller number of studies focused on LLMs' abilities with low-resource historical languages. Stüssi and Ströbel (2024) examines the performance of LLMs on POS tagging for 16th-century Latin, which poses particular challenges, as most taggers have been trained on Classical Latin. They benchmark several tagging models and treebanks, evaluating accuracy across different Latin varieties. The authors find that fine-tuned GPT models can achieve competitive accuracies, with their best fine-tuned variant reaching an average accuracy of nearly 89% on curated treebank data, demonstrating that LLMs — when properly adapted — can bridge gaps left by traditional rule-based taggers on historical text.

Schöffel et al. (2025a) studied what factors influence LLMs' performance on POS tagging in low-resource Medieval Romance languages — specifically Occitan, Medieval French, and Medieval Spanish. By conducting a broad set of experiments, the authors showed how each design choice impacts tagging accuracy across corpora. They demonstrate that, despite carefully designed prompts and transfer methods, historical language variation still presents a challenge difficult to overcome for LLM-based POS taggers. In the same year, Schöffel et al. (2025b) further investigated the abilities and limitations of LLMs when applied historical low-resource languages without standardized orthography. In the study, the authors evaluated several open-source LLMs across two distinct Old Occitan corpora using a range of prompting strategies. The results show significant drops in model performance when confronted with diachronic language changes and non-standard orthography, highlighting biases and failure patterns not typically observed in high-resource languages.

Fang et al. (2025) presents a comparative analysis of word segmentation, POS tagging, and named entity recognition (NER) on historical Chinese sources dating from 1900–1950, contrasting LLM–based approaches with traditional NLP toolkits such as Jieba and spaCy. Because of the unique challenges of historical Chinese—including logographic script, the absence of explicit word boundaries, and diachronic linguistic variation—standard NLP pipelines have been proven to be ineffective. The results showed that LLM-based methods consistently outperform conventional tools across segmentation, POS tagging, and NER metrics, albeit at significantly higher computational costs. The study highlights an important performance–efficiency trade-off and suggests that contextual learning in LLMs can better accommodate genre and temporal variation in historical texts.

Despite growing interest in LLM-based tagging, we identified no studies that evaluate such approaches for the four historical low-resource languages investigated in this work; however, this task has been investigated in different works using traditional approaches. An Old English POS tagging tool is available as part of the CLTK library (Johnson et al., 2021), which was trained on the data from the ISWOC Treebank (Bech and Eide, 2014). The accuracy of the tool has not been tested. Recently, Miani et al. (2025) investigated cross-domain transfer learning for Old English poetry and prose, highlighting the impact of in-domain data on the performance of the models. Old High German was rarely investigated directly, but some studies evaluated existing POS taggers for Early Modern German corpora (Scheible et al., 2011; Ferreri Hanberry, 2015). Koleva et al. (2017) developed a POS tagger for Middle Low German. In addition, the four languages have also been the subject of investigation in Miani et al. (2026), who addressed cross-genre and cross-domain transfer learning on POS tagging with traditional methods, highlighting the impact of structural differences on models' performance.

3. Datasets

3.1. Historical Corpora

The Old English (henceforth OE) poetry data were retrieved from the York-Helsinki Parsed Corpus of Old English Poetry (YCOEP)[2]. Part of the same project is the York Toronto Helsinki Parsed Corpus of Old English (YCOE)[3] which includes approximately 1.5 million words of Old English prose annotated for syntax and morphology. Only the YCOE documentation was available; for this reason, it was adopted for both datasets. The texts are segmented into *tokens* representing main verbs with arguments and adjuncts, matrix inflectional

[2] https://www-users.york.ac.uk/~lang18/pcorpus.html

[3] https://penn-historical-corpora.uni-mannheim.de/ycoe/YCOEHomepage.html

phrases, complementizer phrases, or independent non-clausal utterances.

Both Old High German (henceforth OHG) poetry and prose were extracted from the Deutsch Diachron Digital, Referenzkorpus Altdeutsch (Version 1.2) dataset (Zeige Lars Erik, 2025). The texts, amounting to approximately 650.000 words, are structurally and linguistically annotated. The smallest unit is the *token*, which corresponds to both the edited and normalized versions of each word in the text. In addition to the Old High German manuscripts, some Old Saxon poetry and prose texts are also present; these were extracted and used alongside HeliPaD[4], an Old Saxon (henceforth OS) corpus containing 5.968 lines from the C manuscript of the poem *Heliand*. The annotation follows the same guidelines as the Old English corpora YCOEP and YCOE; therefore, the information included, the segmentation, and the POS tags are consistent across the three corpora.

Both Old Norse (henceforth ON) poetry and prose data were extracted from the Menotec collection[5] available on the INESS platform (Rosén et al., 2012). From the collection, five Old Norse manuscripts were selected: *Edda Regius*, containing 3665 lines, was used for poetry data; while the prose data were retrieved from the following texts, amounting to 10.318 lines: *Pamphilus*, *The Old Norwegian Homily Book*, *the Legendary Saga of St Olaf*, and *Strengleikar*. The texts present morphological and syntactical annotations that follow the guidelines for the annotation of Old Norwegian[6] texts by Haugen and Øverland (2014). [7]

From each corpus, the original textual forms and their POS tags were extracted. As the corpora employ different annotation schemes, all tags were mapped to the Universal Dependencies UPOS tag set (de Marneffe et al., 2021) following the mapping scheme proposed by Miani et al. (2026).

3.2. Modern Corpora

To compare the impact of historical languages on the performance of the generative models, two modern Germanic languages, English and German, were included in the experiment.

Modern English (henceforth ME) prose data were retrieved from the English part of the LinES Parallel Treebank[8](Ahrenberg, 2015). The dataset contains segments from nine different sources, totaling 106.305 tokens. Seven sub-corpora comprise literary works ranging from 1902 to 2014. These literary texts were extracted and used to assess LLM performance on Modern English prose.

To the extent of our knowledge, there is no modern German (henceforth MG) prose corpus with UD annotations available; for this reason, the closest dataset that seemed reasonable to be used to evaluate models' performance is UD German LIT[9], a treebank hosting texts from German literary history. It consists mainly of short, aphorism-like texts from the early Romantic period, roughly dated to the end of the 18th century, which is considered linguistically modern German (Besch and Wolf, 2009).

For both modern English and modern German poetry, the data were retrieved from the PoeTee dataset (Plecháč et al., 2025), which comprises approximately 335.000 poems and 90.000.000 tokens in 11 languages in JSON format with complete Universal Dependencies analysis; covering a period of time from the 13th to the 20th century. For English, approximately 40.000 poems are available from the Project Gutenberg, covering roughly 1650 to 1925, while for German, 74.000 from Metricalizer and Deutsches Lyrik Korpus, with a relatively small number of poems from roughly 1200 to 1600 and a much larger concentration from 1600 to 1925. Given the broad time span covered by the poetry data, we randomly sampled the majority of sentences from 1900 onward.

4. Experimental Setup

Fine-tuning BERT-based models has demonstrated strong performance on POS tagging tasks and has proven particularly effective for historical Germanic languages (Miani et al., 2026). This study adopts the same approach as a benchmark to compare against the generative models. In addition, contrastive models were trained for modern English and German to assess LLM performance on modern languages and genres, and to determine whether the observed difficulties persist beyond historical data.

4.1. Fine-Tuned Encoder Model

For each language, two models were trained: one only with poetry data and one only with prose data. The dataset sizes used to train the models were equal between poetry and prose and were the same as those used by Miani et al. (2026), except for Old Saxon. Due to the scarcity of Old Saxon prose data compared to poetry, we chose to use different dataset sizes for each genre to prevent the limited

[4] https://zenodo.org/records/4395040

[5] https://clarino.uib.no/iness

[6] Old Norwegian and Old Icelandic are collectively known as Old Norse in the Scandinavian languages (Haugen and Øverland, 2014).

[7] Documentation for the POS annotation was kindly provided by Paul Meurer and Odd Einar Haugen.

[8] https://github.com/UniversalDependencies/UD_English-LinES

[9] https://github.com/UniversalDependencies/UD_German-LIT

Lang.	Genre	Train	Dev.	Test	Tokens
OHG	poetry	5,327	665	667	63,203
	prose	5,327	665	667	101,830
OE	poetry	5,039	629	631	85,970
	prose	5,039	629	631	93,645
ON	poetry	3,029	378	380	39,779
	prose	3,029	378	380	60,082
OS	poetry	3,108	388	389	56,532
	prose	444	55	57	7,454
ME	poetry	4,080	510	511	44,770
	prose	4,080	510	511	94,627
MG	poetry	1,536	192	192	13,857
	prose	1,536	192	192	40,340

Table 1: Dataset statistics showing the number of sentences (train, development, and test splits) and tokens per genre for each language. All languages except Old Saxon contain equal amounts of prose and poetry data.

prose data from negatively affecting poetry results. All corpora were divided into 80% for training, 10% for development, and 10% for testing. The resulting dataset sizes are listed in Table 1.

Multilingual BERT (henceforth mBERT) was fine-tuned using MaChAmp (van der Goot et al., 2021), a toolkit designed for multi-task learning across various NLP tasks. Although MaChAmp supports multi-task learning, we trained separate models for each dataset, allowing for direct comparison of performance across individual languages and genres. For POS tagging, we used the `seq` task type, which applies a greedy softmax classification layer over the contextualized token embeddings provided by the encoder. All experiments were conducted using MaChAmp's default hyperparameter settings: a learning rate of 1e-4, a batch size of 32, an AdamW optimizer, and a dropout rate of 0.2. All models were trained for 20 epochs using three random seeds.

4.2. Generative Models

Two generative models were compared: Qwen2.5-3B-Instruct (Qwen Team, 2024) (henceforth Qwen), a multilingual model supporting over 29 languages, including English and German, and Apertus-8B-Instruct-2509 (Apertus et al., 2025) (henceforth Apertus), an open multilingual model designed to support more than 1,000 languages. These models were specifically chosen because of their pretraining data which include languages related to the historical ones — English, German, and Swedish — following the findings of Schöffel et al. (2025b), which demonstrated that such relatedness improves LLM performance. Both models were obtained from the HuggingFace model hub and used in their default instruction-tuned configurations without any additional fine-tuning. They were evaluated on the same test sets described in Table 1 that were used for the encoder models.

For both models, zero- and few-shot prompting strategies were adopted. The prompt design was based on the ones proposed in Schöffel et al. (2025b) and Machado and Ruiz (2024). Each prompt contained a single sentence, tokenized according to the structure present in the corpora. A system-level instruction specifying the task was adopted: POS tag the pre-tokenized input and output in valid JSON format. The models were informed of the language and genre of the data. In the few-shot setting, three examples per language and genre were randomly selected from the training sets and included in the prompt. Inference was performed on a single NVIDIA L40 GPU with greedy decoding. The maximum number of generated tokens was set to 2048. This setup ensured that results reflect the models' out-of-the-box instruction-following capabilities, without any fine-tuning or prompt optimization. The zero-shot prompt structure was the following:

```
You are a linguist performing UD
POS tagging on {language_genre}
data.
RULES:
- The sentences are ALREADY TOK-
ENIZED, so PRESERVE THE ORIGINAL
TOKENIZATION and just ADD POS
TAGS.
- NEVER modify, remove, correct,
normalize the text.
- NEVER add additional words or
punctuation.
- NEVER change the order of the
words or punctuation.
- NEVER drop punctuation.  ALWAYS
keep ALL punctuation marks, even
with multiple occurrences.
- Use ONLY these UD tags:  ADJ,
ADP, ADV, AUX, CCONJ, DET, INTJ,
NOUN, NUM, PART, PRON, PROPN,
PUNCT, SCONJ, VERB, X.
- NEVER add or change tags.
- ALL words need a tag.  NEVER
skip a word or a tag.
- Do NOT add or remove text.
- Do NOT add explanations.
- The OUTPUT FORMAT MUST BE JSON:
["word":  TOKEN, "upos":  TAG,
...]
- The JSON MUST BE VALID, with
commas between all key-value
pairs, no extra quotes, and
closed arrays.
- Close the JSON array properly.
Tag these tokens:
```

In the few-shot setting, the instruction `Follow the examples below:` was appended to the zero-shot prompt along with three examples, before

Original Tag(s)	Mapped Tag
WHADJP, ADJP	ADJ
WHADP	ADP
ADVERB, WHADVP, ADVP	ADV
MD	AUX
CONJ, ACONJ, ADCONJ, CC	CCONJ
WH_DET	DET
IN	INTJ
NEG, TO	PART
VERB:Past, VERB:Present, VERB:Infinitive, INF, MORPH, Mood:Ind|VerbForm:Fut	VERB
LOC, AComp, REL, AD, MWE, WDT	WRONG_UPOS
Missing UPOS	MISS_UPOS
Missing word and tag	MISS_WORD_UPOS

Table 2: POS tag mappings and placeholder substitutions applied during post-processing, based on the analysis of the development sets.

the final instruction `Tag these tokens:`.

Based on observations from the development set, a post-processing step was applied to the model outputs before evaluation. To ensure structural correctness, the output JSON files were validated using the Python package jsonschema[10]; all format errors were manually corrected. We also observed that some POS tags predicted by the LLMs were not valid UD tags; therefore, where the mapping was straightforward, the tags were corrected to valid ones, while the others were categorized as invalid. The resulting mappings are presented in Table 2.

After structural validation and annotation mapping, the predicted sentences were aligned with the gold data. Since the LLMs occasionally modified the input data in terms of context and length, a heuristic alignment strategy was employed. When the predicted and gold sentences match in content and length, structural alignment is not applied, preserving the predicted sentence. In all other cases, structural discrepancies are assumed, and alignment is performed. To address tokenization inconsistencies, the method generates multiple alignment candidates, including single-token matches, right-side merges (to combine consecutive predicted tokens), and left-side merges (to revise previous alignment decisions). Each candidate is evaluated using a similarity score against the gold token, and candidates exceeding a predefined threshold are selected. We tested threshold values of 0.7, 0.8, and 0.9, and found that 0.8 produced the most accurate alignments, as lower values introduced spurious matches while higher values caused valid matches to be missed. The UPOS tag assigned to the merged token corresponds to the last UPOS tag in the merged sequence. When no candidate meets the predefined similarity threshold, the algorithm employs bounded lookahead strategies in both the gold and predicted sequences to detect insertions and deletions. Missing tokens or those that can't be aligned through any strategy are marked with the placeholder tag, `MISS_WORD_UPOS`. As a result, all predicted sentences have the same number of tokens as the corresponding gold sentences.

[10] https://pypi.org/project/jsonschema/

5. Results

The results are evaluated using accuracy, macro F1 score, precision, recall, and per-tag F1 score. For mBERT, all metrics are averaged across seeds. Punctuation marks are excluded from all evaluations.

5.1. Overall Analysis

Tables 3 and 4 compare the performance of mBERT, Qwen, and Apertus across the six languages, in two genres: poetry and prose.

mBERT constantly achieves the highest scores on every language and metric in both genres, with accuracy typically exceeding 0.90 and F1 scores ranging from roughly 0.73 to 0.96. Despite being fine-tuned on considerably limited dataset sizes, with both genres, the encoder architecture demonstrates substantially stronger performance than the two generative models.

Apertus generally achieves slightly higher scores than Qwen in most configurations. The differences between poetry and prose are not particularly pronounced, suggesting that both models struggle regardless of genre-specific features. Across languages, however, the variation is much more evident: all four historical languages yield lower scores than the two modern ones. Old English and Old Norse produce the highest scores among the historical languages. In poetry, Old English achieves similar scores with both generative models, around 0.64 for accuracy and around 0.36 for F1 score. In prose, Apertus outperforms Qwen, particularly in the F1 score, which increases by 8% points. For Old Norse in both poetry and prose data, Apertus similarly achieves considerably higher scores than Qwen. In contrast, Old High German and Old Saxon produce the lowest scores in both genres, indicating that the models struggle significantly to accurately tag data in these languages. These significant struggles with historical languages are much more evident compared to the performance on modern languages. Despite the still substantially lower scores compared to the encoder architecture, both generative models achieve considerably

Lang.	Model	Acc.	F1	Prec.	Rec.
OHG	mBERT	**0.927**	**0.831**	**0.850**	**0.820**
	Q (0)	0.371	0.187	0.218	0.215
	Q (F)	0.444	0.268	0.300	0.305
	A (0)	0.382	0.213	0.302	0.281
	A (F)	*0.456*	*0.301*	*0.402*	*0.314*
OE	mBERT	**0.950**	**0.957**	**0.955**	**0.959**
	Q (0)	0.640	0.364	0.399	0.429
	Q (F)	0.673	0.401	0.412	0.458
	A (0)	0.637	0.368	0.413	0.430
	A (F)	*0.683*	*0.447*	*0.539*	*0.509*
ON	mBERT	**0.912**	**0.922**	**0.915**	**0.930**
	Q (0)	0.602	0.321	0.376	0.342
	Q (F)	0.632	0.406	0.437	0.437
	A (0)	0.663	0.422	0.462	0.421
	A (F)	*0.712*	*0.543*	*0.570*	*0.544*
OS	mBERT	**0.961**	**0.893**	**0.897**	**0.890**
	Q (0)	0.425	0.215	0.238	0.219
	Q (F)	0.551	0.352	0.372	0.351
	A (0)	0.446	0.238	0.266	0.251
	A (F)	*0.522*	*0.386*	*0.449*	*0.436*
ME	mBERT	**0.950**	**0.836**	**0.848**	**0.833**
	Q (0)	0.807	0.521	0.569	0.517
	Q (F)	0.817	0.503	0.545	0.501
	A (0)	0.850	0.622	0.673	0.608
	A (F)	*0.870*	*0.639*	*0.674*	*0.633*
MG	mBERT	**0.944**	**0.833**	**0.824**	**0.850**
	Q (0)	0.767	0.472	0.479	0.531
	Q (F)	0.782	0.475	0.475	0.510
	A (0)	0.805	0.536	*0.612*	0.556
	A (F)	*0.837*	*0.560*	0.604	*0.588*

Table 3: Evaluation results for the POS tagging performance of mBERT, Qwen, and Apertus on the **poetry** test sets. **(0)** and **(F)** indicate respectively the zero- and few-shot settings. **Bold** values indicate the best-performing model per language. *Italics* indicates the highest scores for the generative models.

Lang.	Model	Acc.	F1	Prec.	Rec.
OHG	mBERT	**0.931**	**0.845**	**0.849**	**0.842**
	Q (0)	0.397	0.204	0.271	0.216
	Q (F)	*0.484*	*0.287*	0.339	*0.270*
	A (0)	0.369	0.200	0.299	0.215
	A (F)	0.464	0.273	*0.387*	0.269
OE	mBERT	**0.966**	**0.926**	**0.928**	**0.923**
	Q (0)	0.611	0.310	0.366	0.305
	Q (F)	0.714	0.368	0.402	0.368
	A (0)	0.653	0.390	0.450	0.401
	A (F)	*0.746*	*0.498*	*0.547*	*0.513*
ON	mBERT	**0.923**	**0.791**	**0.793**	**0.790**
	Q (0)	0.548	0.264	0.328	0.273
	Q (F)	0.606	0.309	0.354	0.314
	A (0)	0.592	0.355	0.380	0.372
	A (F)	*0.679*	*0.419*	*0.425*	*0.480*
OS	mBERT	**0.875**	**0.730**	**0.744**	**0.722**
	Q (0)	0.367	0.199	0.230	0.226
	Q (F)	0.423	0.233	0.287	0.241
	A (0)	0.377	0.267	*0.381*	*0.297*
	A (F)	*0.466*	*0.286*	0.374	0.277
ME	mBERT	**0.978**	**0.900**	**0.913**	**0.892**
	Q (0)	0.803	0.514	0.570	0.555
	Q (F)	0.808	0.394	0.427	0.420
	A (0)	0.878	*0.645*	*0.649*	*0.683*
	A (F)	*0.891*	0.583	0.577	0.627
MG	mBERT	**0.958**	**0.859**	**0.871**	**0.850**
	Q (0)	0.735	0.434	0.444	0.478
	Q (F)	0.770	0.482	0.509	0.523
	A (0)	0.821	0.567	0.623	0.568
	A (F)	*0.859*	*0.677*	*0.719*	*0.718*

Table 4: Evaluation results for the POS tagging performance of mBERT, Qwen, and Apertus on the **prose** test sets. **(0)** and **(F)** indicate respectively the zero- and few-shot settings. **Bold** values indicate the best-performing model per language. *Italics* indicates the highest scores for the generative models.

higher accuracy scores in both genres, especially Apertus. This significant drop in performance from modern to historical languages underscores the severe lack of linguistic knowledge for the historical Germanic languages in these models. Despite the improvement in modern languages, scores remain lower than the encoder architecture, and the gap between accuracy and F1 scores is still more pronounced than for mBERT, suggesting particularly poor performance on rare tags.

The behaviors of the generative models in the zero-shot setting are consistent with the few-shot one; however, overall, the performance improves considerably. Particularly pronounced are the increments of the F1 score values for poetry in the historical languages. The performance of the models with the modern languages is not really influenced by the different prompting strategies; in fact, in some cases, such as Modern English prose, few-shot prompting even worsens the F1 score by 12 percentage points. In addition, despite the improvement in results with the few-shot setting, the gap between accuracy and F1 scores remains far more pronounced compared to the encoder models, indicating persistent struggles with rare tags.

5.2. Per-Tag Analysis

Tables 5 and 6 present the per-tag F1 scores for poetry and prose, respectively, for each model and language in the few-shot setting. The zero-shot results follow similar patterns and are reported in Appendix A, in Tables 8 and 9. Appendix A also includes Figures 1 and 2, showing the relative frequency distributions for all datasets in poetry and prose, respectively.

The per-tag analysis confirms the poor performance of the generative models compared to the encoder architecture, as suggested by the overall metrics in Tables 3 and 4. With both genres, both

	OHG				OE				ON				OS				ME				MG			
POS	*mBERT*	*Qwen*	*Apertus*	*Support*	*mBERT*	*Qwen*	*Apertus*	*Support*	*mBERT*	*Qwen*	*Apertus*	*Support*	*mBERT*	*Qwen*	*Apertus*	*Support*	*mBERT*	*Qwen*	*Apertus*	*Support*	*mBERT*	*Qwen*	*Apertus*	*Support*
ADJ	.804	.301	.196	250	.888	.526	.565	711	.816	.423	.607	265	.908	.335	.320	368	.910	.785	.810	332	.884	.712	.793	116
ADP	.967	.603	.569	338	.972	.795	.803	673	.930	.707	.782	233	.984	.661	.645	358	.953	.796	.867	405	.985	.777	.911	108
ADV	.912	.430	.380	694	.917	.435	.526	537	.886	.406	.639	283	.967	.409	.450	470	.891	.667	.665	203	.871	.638	.694	84
AUX	.891	.282	.227	236	.935	.531	.246	180	.854	.483	.265	81	.913	.233	.165	125	.954	.571	.879	174	.965	.611	.737	45
CCONJ	.930	.099	.088	149	.980	.863	.447	233	.967	.825	.866	124	.992	.847	.834	101	.989	.888	.977	198	.973	.843	.872	66
DET	.949	.160	.430	713	.964	.254	.566	505	.933	.163	.502	150	.974	.305	.372	400	.986	.895	.920	430	.997	.829	.893	150
INTJ	.743	.027	.024	9	1.000	.353	.387	6	.000	.000	.000	0	1.000	.000	.091	1	.778	.571	.465	14	.000	.000	.000	0
NOUN	.916	.495	.501	873	.957	.785	.800	2116	.900	.710	.814	770	.960	.655	.593	1035	.964	.941	.942	831	.957	.895	.910	267
NUM	.691	.143	.222	13	.900	.468	.462	37	.936	.415	.800	12	.856	.000	.222	16	.929	.615	.571	14	.889	.500	.500	1
PART	.959	.248	.479	106	.985	.000	.102	56	1.000	.000	.000	1	.986	.000	.000	70	.978	.183	.598	84	.894	.118	.250	7
PRON	.969	.556	.646	854	.991	.762	.835	543	.970	.694	.711	347	.985	.622	.614	648	.982	.910	.940	428	.980	.848	.862	162
PROPN	.930	.559	.364	74	.963	.630	.659	180	.942	.846	.883	166	.981	.849	.796	141	.852	.733	.781	75	.750	.471	.432	12
SCONJ	.878	.399	.437	183	.956	.399	.403	117	.935	.513	.508	161	.947	.554	.561	136	.827	.145	.564	97	.915	.000	.788	16
VERB	.931	.525	.546	843	.948	.733	.721	1143	.915	.712	.775	692	.948	.512	.505	733	.943	.850	.891	526	.924	.832	.876	146
X	.000	.000	.000	2	1.000	.476	.075	6	.000	.000	.000	0	.000	.000	.000	0	.444	.000	.000	2	.510	.000	.000	7

Table 5: Per-tag F1 scores and corresponding support values for POS tagging with fine-tuned mBERT and generative models (Qwen, Apertus) in the **few-shot** setting on **poetry** test sets.

	OHG				OE				ON				OS				ME				MG			
POS	*mBERT*	*Qwen*	*Apertus*	*Support*	*mBERT*	*Qwen*	*Apertus*	*Support*	*mBERT*	*Qwen*	*Apertus*	*Support*	*mBERT*	*Qwen*	*Apertus*	*Support*	*mBERT*	*Qwen*	*Apertus*	*Support*	*mBERT*	*Qwen*	*Apertus*	*Support*
ADJ	.818	.258	.235	403	.887	.501	.569	504	.778	.463	.543	279	.644	.130	.241	24	.944	.829	.867	603	.919	.855	.859	345
ADP	.965	.648	.573	702	.980	.796	.826	875	.975	.665	.824	476	.918	.471	.611	63	.981	.816	.898	878	.967	.728	.843	259
ADV	.870	.257	.241	599	.945	.518	.651	634	.928	.549	.673	461	.620	.109	.185	29	.954	.661	.767	483	.921	.748	.793	333
AUX	.921	.383	.493	386	.944	.310	.191	189	.909	.358	.078	114	.891	.478	.576	47	.975	.409	.869	476	.478	.413	.849	231
CCONJ	.961	.602	.402	522	.996	.881	.923	598	.985	.801	.827	410	.968	.400	.211	51	.990	.927	.971	264	.933	.828	.898	162
DET	.956	.298	.514	1165	.976	.486	.632	886	.926	.243	.467	583	.916	.245	.324	75	.998	.938	.938	777	.987	.806	.904	522
INTJ	.814	.082	.083	31	1.000	.000	.105	4	.000	.000	.154	1	.000	.250	.000	3	.884	.533	.323	6	.000	.000	.000	0
NOUN	.930	.570	.520	1593	.958	.802	.841	1436	.895	.713	.774	1024	.846	.504	.542	123	.982	.926	.943	1570	.983	.952	.964	699
NUM	.792	.096	.062	49	.962	.670	.480	94	.000	.000	.000	0	.000	.000	.000	0	.955	.782	.839	59	.909	.556	1.000	6
PART	.950	.385	.074	181	.978	.000	.021	92	.929	.000	.000	35	.899	.000	.000	16	.996	.082	.706	190	.967	.102	.113	100
PRON	.965	.486	.524	837	.991	.848	.886	1040	.958	.667	.736	468	.957	.691	.676	74	.993	.880	.897	1009	.988	.792	.831	283
PROPN	.941	.597	.484	226	.980	.879	.920	346	.852	.611	.632	144	.943	.318	.400	17	.977	.741	.853	239	.774	.658	.716	35
SCONJ	.846	.065	.135	268	.957	.573	.500	295	.954	.289	.587	315	.746	.000	.273	14	.959	.055	.733	139	.931	.254	.776	55
VERB	.944	.559	.572	1418	.969	.786	.777	1270	.925	.725	.740	986	.869	.368	.539	116	.971	.824	.927	1071	.942	.741	.852	311
X	.000	.000	.000	0	.968	.043	.140	84	.848	.098	.098	29	.000	.000	.000	0	.841	.060	.129	4	.711	.250	.429	3

Table 6: Per-tag F1 scores and corresponding support values for POS tagging with fine-tuned mBERT and generative models (Qwen, Apertus) in the **few-shot** setting on **prose** test sets.

Qwen and Apertus struggle more with historical languages compared to modern ones. High-frequency categories, such as NOUN, VERB, PRON, and DET, which achieve the highest scores with mBERT, typically exceeding 0.85 and frequently surpassing 0.95, have the worst scores with the generative models, especially with the historical languages. Rare classes such as NUM, INTJ, and X, are penalized by their low frequency, which leads to low performance with mBERT, but influences the LLMs negatively. Interestingly, also ADJ appears to be a challenging tag: in both genres, but especially in the poetry, mBERT manages to barely surpass 0.91, and the LLMs achieve comparable scores only in the modern languages. This behavior could be justified by the overlaps with other classes, such as nouns, participles, or adverbs. For instance, Old English has the highest number of ADJ in the test set, and the same words can also be tagged as PROPN, NOUN, ADV, VERB, DET, and AUX.

Comparing the performance of the two LLMs reveals that Apertus tends to achieve slightly higher scores compared to Qwen in the poetry data, but their behaviors are similar. In the prose data, Apertus performs better, with higher scores in almost all languages. As highlighted by the relative frequency distributions presented in Figures 1 and 2 in Appendix A, and also by previous studies (Miani et al., 2026), the structural differences of the two genres are consistent and may influence the model's performance. In the poetry data, the label distribution appears unbalanced, with a smaller number of high-frequency classes accounting for large proportions of tokens; in contrast to the prose data, which exhibits a more balanced distribution across labels. Although rare classes are usually more prone to misclassification, the more balanced nature of the prose data can help support more stable performance. Despite the fact that neither of the generative models is able to surpass the encoder architecture, Apertus appears to be able to address this class imbalance better than the other LLM.

5.3. Structural Output Errors and Schema Violations

Table 7 presents the error rates for JSON formatting issues and incorrect additional tags in the test sets with both prompting strategies. The JSON errors column indicates the proportion of malformed outputs generated by Qwen and Apertus. The files were validated with a JSON checker, and the resulting percentage was based on the number of mistakes identified with the tool. The percentage

Lang.	Genre	JSON errors		MAP_UPOS		WRONG_UPOS		MISS_UPOS		MISS_WORD_UPOS	
		Q	A	Q	A	Q	A	Q	A	Q	A
Zero-shot											
OHG	poe	0.75	3.60	0.28	-	0.04	-	-	0.41	0.36	1.14
	pro	6.60	4.50	0.33	-	0.02	-	-	0.89	0.73	1.55
OE	poe	1.43	7.45	1.59	0.07	0.14	-	-	0.17	0.26	1.76
	pro	2.54	8.56	1.57	-	0.32	0.01	0.06	0.84	1.16	2.67
ON	poe	1.05	3.68	0.97	-	0.27	-	-	-	0.15	0.37
	pro	3.16	6.84	1.07	-	0.28	-	-	-	1.24	0.56
OS	poe	5.40	7.71	0.15	-	0.17	-	-	0.78	0.04	2.39
	pro	5.26	10.53	0.77	-	-	-	-	5.83	0.15	9.66
ME	poe	2.15	6.26	2.41	0.03	0.05	-	-	1.29	1.15	1.05
	pro	9.59	34.44	0.89	0.28	0.40	-	-	0.88	0.50	3.36
MG	poe	1.04	3.65	2.78	-	-	-	-	0.51	0.42	0.76
	pro	4.69	30.73	1.08	-	0.12	-	-	0.21	1.41	3.14
Few-shot											
OHG	poe	1.35	1.95	0.15	-	0.02	-	-	0.13	0.58	0.69
	pro	16.04	4.20	0.11	-	0.05	-	-	0.79	0.44	1.19
OE	poe	1.27	2.69	0.94	0.07	0.11	-	-	-	0.18	0.47
	pro	3.33	4.12	0.08	0.04	0.31	-	-	0.49	0.54	0.26
ON	poe	0.79	0.79	0.37	-	-	-	-	-	0.27	0.03
	pro	4.47	3.95	0.30	-	0.56	-	-	-	0.36	0.19
OS	poe	7.71	0.26	-	-	-	-	-	-	0.13	0.02
	pro	15.79	5.26	0.31	-	-	-	-	5.83	0.46	7.82
ME	poe	1.17	2.35	2.18	0.29	0.18	-	-	-	1.15	0.16
	pro	18.98	27.79	0.30	0.05	0.54	0.04	-	0.05	0.57	2.09
MG	poe	3.12	3.12	0.25	-	-	-	-	-	0.93	0.34
	pro	44.27	23.44	0.33	-	0.03	-	-	-	1.73	1.32

Table 7: Percentage of structural and tagging errors produced by Qwen (Q) and Apertus (A) across languages and genres in the zero- and few-shot settings. JSON errors correspond to malformed outputs or schema violations, while tag-related errors include mapped erroneous UPOS tags (MAP_UPOS), invalid UPOS labels (WRONG_UPOS), missing POS tags (MISS_UPOS), and missing tokens introduced during alignment (MISS_WORD_UPOS).

can include different errors, such as structurally invalid JSON (e.g., missing brackets, omitted required schema elements), and incomplete or improperly formatted sentences. The errors tend to decrease from the zero-shot to the few-shot setting, especially for Apertus, suggesting that the inclusion of examples helps the model mitigate structural issues. Qwen shows improvements with Old Saxon, but error rates generally increase across the other languages, indicating that the few-shot examples are less beneficial for this model than for Apertus. In both settings, the models exhibit higher error rates on prose than on poetry, likely due to the longer sentence lengths in prose. Overall, Apertus presents higher percentages of JSON errors, especially with modern English and German prose. These two test sets are also the ones with higher average sentence lengths, respectively 17 and 20 tokens; this increased token count could potentially lead to interruptions in the generation process and consequent output degradation, which would explain the high error rates.

MAP_UPOS reports the percentage of invalid UD tags that were mapped to valid ones following Table 2. Qwen produces more invalid tags than Apertus, with both prompting strategies, particularly with Old English, Old Norse, modern English and modern German. Apertus rarely deviates from the UD tagset, with few instances limited to Old and modern English, suggesting that it adheres more closely to the prompt guidelines and requires less post-processing.

The placeholder WRONG_UPOS was used to mark tags that do not belong to the Universal Dependencies tag set and could not be mapped. The reported percentage includes both the tags listed in Table 2 and additional invalid ones generated in the test data. As for MAP_UPOS, Apertus adheres more closely to the prompt instructions and the UD tag inventory, producing only one extra tag in Old English with the zero-shot setting and in modern English with the few-shot one. In contrast, Qwen introduces extra tags in all test sets with the zero-shot setting, but slightly improves the performance

with the few-shot approach.

The MISS_UPOS placeholder was introduced to mark tokens for which no POS tag was provided. Unlike the MAP_UPOS and WRONG_UPOS errors, Qwen successfully assigned tags to nearly all tokens across languages, with only a few omissions in Old English prose in the zero-shot setting. These errors are absent in the few-shot setting. Apertus, however, failed to tag multiple token constructions in most test sets, except for Old Norse. The highest percentages are observed in the Old Saxon prose dataset, likely due to the relatively small size of the test set that amplifies the impact of the error on the overall rate. For the other languages, the scores do not exceed the 1% threshold. The model improves with the few-shot setting, reducing error rates across most languages and eliminating them completely in modern German.

The MISS_WORD_UPOS placeholder was used during the alignment procedure to represent missing tokens and their corresponding POS tags. With both prompting strategies, both LLMs exhibit errors of this type, although Apertus shows higher rates than Qwen, particularly in Old Saxon and the modern language datasets. With the few-shot prompting strategy, Qwen presents worse performance, slightly increasing the error rates, while Apertus, still presents higher scores than the other LLM, but shows a general improvement compared to the zero-shot setting. The results imply that both models were unable to consistently preserve the required input structure, often modifying it in ways that necessitated manual correction.

6. Discussion

The evaluation of the generative models' performance on zero- and few-shot POS tagging on historical Germanic languages revealed the absence of linguistic knowledge for these languages in the LLMs' pretraining data. Despite improved results in the few-shot setting, neither model was able to outperform the fine-tuned mBERT models, which achieved high scores across all languages, genres, and prompting strategies. This interpretation is further supported by the markedly better performance of the generative models on modern English and German data: although their scores still fell short of mBERT, both models obtained substantially stronger results on the modern datasets, suggesting that they possess sufficient linguistic knowledge for these contemporary languages to perform the task more successfully, but critically lack it for their historical counterparts.

The per-tag analysis reinforces these findings. mBERT clearly outperforms both LLMs across all tags, and both generative models achieve higher scores for modern languages than for historical ones. Apertus generally performs slightly better than Qwen with both poetry and prose. The results also underline the impact of the substantial differences between the genres: the unbalanced nature of the poetry data results in fewer high-frequency classes compared to the prose data, which presents a more balanced distribution. Such distributional characteristics may increase the difficulty of correctly predicting rare classes, leading to lower overall scores. While mBERT appears more robust to these effects, both LLMs are facing issues in raising the scores for low-frequency tags.

The error analysis further reveals a fundamental limitation of generative models: their persistent tendency to alter or restructure the required output format. While the few-shot setting mitigates structural issues for one generative model, it does not eliminate the errors, and substantial post-processing remains necessary — an issue that does not arise with fine-tuned encoder-based architectures. Although post-processing may be acceptable when model performance is competitive, the consistently low scores achieved by the LLMs suggest that the additional effort is not justified, particularly when fine-tuned encoder models deliver substantially superior results without the need for post-processing.

7. Conclusions

The study evaluates zero- and few-shot POS tagging performance of two generative models—Qwen2.5-3B-Instruct and Apertus-8B-Instruct-2509—on four historical Germanic low-resource languages and two literary genres. The results are benchmarked against a fine-tuned Multilingual BERT baseline. Two modern languages are used to investigate the impact of historical data on the models' performance.

The results demonstrate that fine-tuned encoder models consistently outperform the generative models on both historical and modern datasets. While, with both prompting strategies, the generative models achieve relatively weak performance on the historical languages, they obtain noticeably better — though still inferior — results on the modern languages, suggesting limited exposure to or representation of the historical varieties in their training data. Error analysis further highlights the tendency of LLMs to generate malformed outputs requiring manual correction and post-processing, not needed with encoder models.

Future work will extend this investigation to additional NLP tasks and explore the fine-tuning of generative LLMs to assess whether task-specific adaptation can improve performance.

8. Limitations

This study evaluates two generative models against a benchmark on the POS tagging task for four historical low-resource Germanic languages.

All four languages are morphologically rich, and their original POS tag sets were highly fine-grained. Despite the UPOS adopted mapping being the most suitable approach for comparison of model performance, it may lead to linguistic information losses. In addition, original annotation errors may influence the quality of the UPOS labels.

Due to the limited availability of UD-annotated corpora covering both genres for the modern languages, the data spans an extended time period, potentially introducing diachronic variation that may have negatively affected model performance.

The generative models tend to modify the original input through normalization, paraphrasing, tokenization, or lexical substitution, rendering traditional string-matching alignment methods ineffective. The adopted custom alignment function attempts to address these modifications, but it relies on heuristic assumptions derived from the development sets. Consequently, it may fail to account for novel textual characteristics not encountered during development, potentially affecting evaluation accuracy.

To maintain comparability across experiments, dataset sizes were substantially reduced, which may have constrained model performance and obscured potential improvements achievable with larger training sets. Additionally, all models were run with default hyperparameters and greedy decoding to ensure deterministic and reproducible outputs, though different prompts or decoding settings could lead to different results, and fine-tuning the models on the target task might improve performance substantially.

9. Acknowledgments

This work has been supported by the Swedish Graduate School of Digital Philology, funded by the Swedish Research Council (grant 2022-06343). Computations were enabled by resources at the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX).

10. Bibliographical References

Lars Ahrenberg. 2015. Converting an English-Swedish parallel treebank to Universal Dependencies. In *Proceedings of the Third International Conference on Dependency Linguistics (DepLing 2015)*, pages 10–19, Uppsala, Sweden. Association for Computational Linguistics.

Project Apertus, Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pasztor, Bettina Messmer, Dhia Garbaya, Eduard Frank Ďurech, Ido Hakimi, Juan García Giraldo, Mete Ismayilzada, Negar Foroutan, Skander Moalla, Tiancheng Chen, Vinko Sabolčec, Yixuan Xu, Michael Aerni, Badr AlKhamissi, Inés Altemir Mariñas, Mohammad Hossein Amani, Matin Ansaripour, Ilia Badanin, Harold Benoit, Emanuela Boros, Nicholas Browning, Fabian Bösch, Maximilian Böther, Niklas Canova, Camille Challier, Clement Charmillot, Jonathan Coles, Jan Deriu, Arnout Devos, Lukas Drescher, Daniil Dzenhaliou, Maud Ehrmann, Dongyang Fan, Simin Fan, Silin Gao, Miguel Gila, María Grandury, Diba Hashemi, Alexander Hoyle, Jiaming Jiang, Mark Klein, Andrei Kucharavy, Anastasiia Kucherenko, Frederike Lübeck, Roman Machacek, Theofilos Manitaras, Andreas Marfurt, Kyle Matoba, Simon Matrenok, Henrique Mendonça, Fawzi Roberto Mohamed, Syrielle Montariol, Luca Mouchel, Sven Najem-Meyer, Jingwei Ni, Gennaro Oliva, Matteo Pagliardini, Elia Palme, Andrei Panferov, Léo Paoletti, Marco Passerini, Ivan Pavlov, Auguste Poiroux, Kaustubh Ponkshe, Nathan Ranchin, Javi Rando, Mathieu Sauser, Jakhongir Saydaliev, Muhammad Ali Sayfiddinov, Marian Schneider, Stefano Schuppli, Marco Scialanga, Andrei Semenov, Kumar Shridhar, Raghav Singhal, Anna Sotnikova, Alexander Sternfeld, Ayush Kumar Tarun, Paul Teiletche, Jannis Vamvas, Xiaozhe Yao, Hao Zhao, Alexander Ilic, Ana Klimovic, Andreas Krause, Caglar Gulcehre, David Rosenthal, Elliott Ash, Florian Tramèr, Joost VandeVondele, Livio Veraldi, Martin Rajman, Thomas Schulthess, Torsten Hoefler, Antoine Bosselut, Martin Jaggi, and Imanol Schlag. 2025. Apertus: Democratizing open and compliant llms for global language environments.

Kristin Bech and Kristine Eide. 2014. The ISWOC corpus. Department of Literature, Area Studies and European Languages, University of Oslo.

Werner Besch and Norbert Richard Wolf. 2009. *Geschichte der deutschen Sprache*. Erich Schmidt, Berlin.

Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2023. Prompting language models for linguistic structure. *arXiv preprint arXiv:2211.07830*.

Chao-Yang Chang, Yan-Ming Lin, Chih-Chung Kuo, Yen-Chun Lai, Chao-Shih Huang, Yuan-Fu Liao, and Tsun-guan Thiann. 2024. A preliminary study on Taiwanese POS taggers: Leveraging Chinese in the absence of Taiwanese POS annotation datasets. In *2024 27th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Zhao Fang, Liang-Chun Wu, Xuening Kong, and Spencer Dean Stewart. 2025. A comparative analysis of word segmentation, part-of-speech tagging, and named entity recognition for historical Chinese sources, 1900–1950. *arXiv preprint arXiv:2503.19844*.

Brendan Ferreri Hanberry. 2015. Application of a POS tagger to a novel chronological division of Early Modern German text. Master's thesis, University of North Carolina at Chapel Hill.

Odd Einar Haugen and Fartein Th. Øverland. 2014. Guidelines for morphological and syntactic annotation of Old Norwegian texts. *Bergen Language and Linguistics Studies (BeLLS)*, 4(2).

Kyle P. Johnson, Patrick J. Burns, John Stewart, Todd Cook, Clément Besnier, and William J. B. Mattingly. 2021. The Classical Language Toolkit: An NLP framework for pre-modern languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 20–29, Online. Association for Computational Linguistics.

Mariya Koleva, Melissa Farasyn, Bart Desmet, Anne Breitbarth, and Veronique Hoste. 2017. An automatic part-of-speech tagger for Middle Low German. *International Journal of Corpus Linguistics*, 22(1):108–141.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint arXiv:2304.05613*.

Mateus Machado and Evandro Ruiz. 2024. Evaluating large language models for the tasks of PoS tagging within the Universal Dependency framework. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese – Vol. 1*, pages 454–460, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.

Irene Miani, Sara Stymne, and Gregory R. Darwin. 2025. Cross-genre learning for Old English poetry POS tagging. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 708–724, Vienna, Austria. Association for Computational Linguistics.

Irene Miani, Sara Stymne, and Gregory R. Darwin. 2026. Cross-lingual and cross-domain transfer learning for POS tagging in historical Germanic low-resource languages. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026)*, pages 542–558, Rabat, Morocco. Association for Computational Linguistics.

Pavel Plecháč, Andris Šeļa, Héctor Bermúdez Sabel, Katharina Bobenhausen, Soňa Cinková, I. L. Dale, Émilie Delente, Marco De Sisto, Thomas Haider, Benjamin Hammerich, Péter Horváth, Randi Kvinnsland, Niko Kočnik, Radek Kolár, Kirill Korchagin, Anna Martynenko, Alexander Mittmann, Bence Nagy, Begoña Navarro Colorado, and Diana Sitchinava. 2025. PoeTree: Poetry corpora in Czech, English, French, German, Hungarian, Italian, Norwegian, Portuguese, Russian, Slovenian, and Spanish (1.0.0) [data set].

Libo Qin, Qiguang Chen, Xiachong Feng, Yang Wu, Yongheng Zhang, Yinghui Li, Min Li, Wanxiang Che, and Philip S. Yu. 2025. Large language models meet NLP: A survey. *arXiv preprint arXiv:2405.12819*.

Qwen Team. 2024. Qwen2.5: A party of foundation models. Qwen Blog.

Victoria Rosén, Koenraad De Smedt, Paul Meurer, and Helge Dyvik. 2012. An open infrastructure for advanced treebanking. In *META-RESEARCH Workshop on Advanced Treebanking at LREC 2012*, pages 22–29, Istanbul, Turkey.

Silke Scheible, Richard J. Whitt, Martin Durrell, and Paul Bennett. 2011. Evaluating an 'off-the-shelf' POS-tagger on Early Modern German text. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 19–23, Portland, OR, USA. Association for Computational Linguistics.

Matthias Schöffel, Esteban Garcés Arias, Marinus Wiedner, Paula Ruppert, Meimingwei Li, Christian Heumann, and Matthias Aßenmacher. 2025a. Unveiling factors for enhanced POS tagging: A study of low-resource medieval romance languages. *arXiv preprint arXiv:2506.17715*.

Matthias Schöffel, Marinus Wiedner, Esteban Garces Arias, Paula Ruppert, Christian Heumann, and Matthias Aßenmacher. 2025b. Modern models, medieval texts: A POS tagging study of Old Occitan. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 334–349, Albuquerque, USA. Association for Computational Linguistics.

Elina Stüssi and Phillip Ströbel. 2024. Part-of-speech tagging of 16th-century Latin with GPT. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 196–206, St. Julians, Malta. Association for Computational Linguistics.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

Klotz Martin Donhauser Karin Gippert Jost Lühr Rosemarie Zeige Lars Erik, Schnelle Gohar. 2025. Deutsch Diachron Digital – Referenzkorpus Altdeutsch (Version 1.2). Humboldt-Universität zu Berlin.

A. Appendix

A.1. Relative POS Tag Frequency

The percentages for the relative frequency of POS tags in the two genres are presented in Figures 1 and 2 for poetry and prose, respectively.

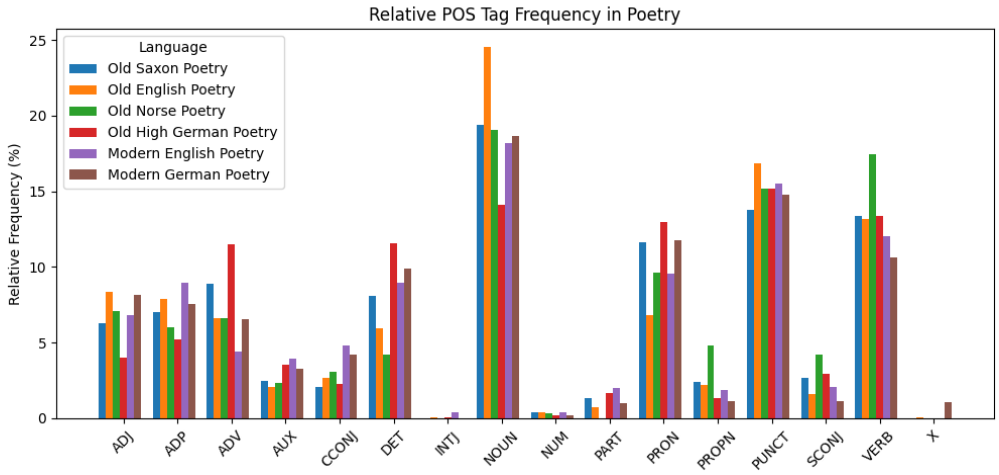


Figure 1: Relative POS tag frequency in the **poetry** datasets.

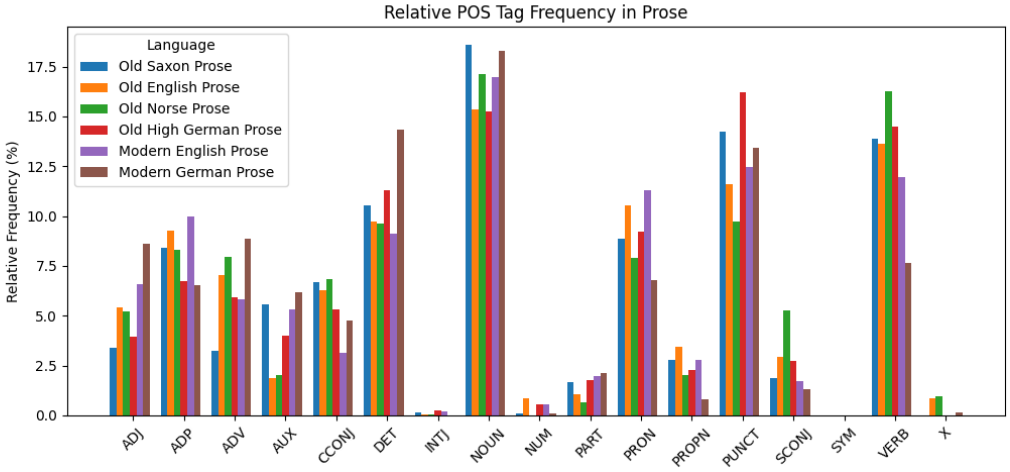


Figure 2: Relative POS tag frequency in the **prose** datasets.

A.2. Additional Results

Tables 8 and 9 present the F1 scores in the zero-shot setting for poetry and prose.

	OHG				OE				ON				OS				ME				MG			
POS	*mBERT*	*Qwen*	*Apertus*	*Support*	*mBERT*	*Qwen*	*Apertus*	*Support*	*mBERT*	*Qwen*	*Apertus*	*Support*	*mBERT*	*Qwen*	*Apertus*	*Support*	*mBERT*	*Qwen*	*Apertus*	*Support*	*mBERT*	*Qwen*	*Apertus*	*Support*
ADJ	.804	.208	.183	250	.888	.468	.524	711	.816	.399	.573	265	.908	.285	.299	368	.910	.778	.784	332	.884	.717	.758	116
ADP	.967	.527	.521	338	.972	.753	.778	673	.930	.674	.689	233	.984	.388	.604	358	.953	.797	.864	405	.985	.766	.889	108
ADV	.912	.268	.271	694	.917	.367	.526	537	.886	.400	.594	283	.967	.271	.458	470	.891	.600	.639	203	.871	.632	.725	84
AUX	.891	.096	.088	236	.935	.271	.155	180	.854	.311	.080	81	.913	.078	.027	125	.954	.414	.830	174	.965	.400	.433	45
CCONJ	.930	.091	.052	149	.980	.886	.407	233	.967	.780	.639	124	.992	.118	.000	101	.989	.946	.969	198	.973	.832	.872	66
DET	.949	.091	.310	713	.964	.315	.561	505	.933	.085	.297	150	.974	.140	.373	400	.986	.899	.906	430	.997	.853	.894	150
INTJ	.743	.042	.048	9	1.000	.308	.375	6	.000	.000	.000	0	1.000	.000	.000	1	.778	.600	.390	14	.000	.000	.000	0
NOUN	.916	.469	.447	873	.957	.765	.748	2116	.900	.698	.757	770	.960	.611	.540	1035	.964	.927	.941	831	.957	.868	.898	267
NUM	.691	.095	.261	13	.900	.480	.440	37	.936	.375	.769	12	.856	.062	.182	16	.929	.667	.727	14	.889	.500	.500	1
PART	.959	.130	.018	106	.985	.064	.000	56	1.000	.000	.000	1	.986	.101	.000	70	.978	.200	.483	84	.894	.194	.250	7
PRON	.969	.549	.601	854	.991	.713	.814	543	.970	.689	.714	347	.985	.453	.509	648	.982	.916	.932	428	.980	.866	.823	162
PROPN	.930	.497	.254	74	.963	.598	.541	180	.942	.870	.855	166	.981	.607	.454	141	.852	.705	.786	75	.750	.383	.421	12
SCONJ	.878	.036	.029	183	.956	.037	.026	117	.935	.126	.485	161	.947	.045	.105	136	.827	.093	.444	97	.915	.000	.720	16
VERB	.931	.461	.521	843	.948	.699	.671	1143	.915	.697	.729	692	.948	.500	.500	733	.943	.829	.872	526	.924	.806	.814	146
X	.000	.000	.008	2	1.000	.190	.053	6	.000	.000	.000	0	.000	.000	.000	0	.444	.000	.000	2	.510	.200	.118	7

Table 8: Per-tag F1 scores and corresponding support values for POS tagging with fine-tuned mBERT and generative models (Qwen, Apertus) in the **zero-shot** setting on **poetry** test sets.

	OHG				OE				ON				OS				ME				MG			
POS	*mBERT*	*Qwen*	*Apertus*	*Support*	*mBERT*	*Qwen*	*Apertus*	*Support*	*mBERT*	*Qwen*	*Apertus*	*Support*	*mBERT*	*Qwen*	*Apertus*	*Support*	*mBERT*	*Qwen*	*Apertus*	*Support*	*mBERT*	*Qwen*	*Apertus*	*Support*
ADJ	.818	.245	.234	403	.887	.434	.483	504	.778	.445	.500	279	.644	.089	.230	24	.944	.827	.857	603	.919	.819	.837	345
ADP	.965	.545	.498	702	.980	.711	.768	875	.975	.610	.695	476	.918	.397	.534	63	.981	.814	.895	878	.967	.749	.841	259
ADV	.870	.239	.266	599	.945	.453	.613	634	.928	.520	.560	461	.620	.172	.271	29	.954	.657	.744	483	.921	.731	.780	333
AUX	.921	.068	.147	386	.944	.290	.237	189	.909	.199	.056	114	.891	.109	.192	47	.975	.320	.853	476	.957	.177	.586	231
CCONJ	.961	.121	.123	522	.996	.431	.613	598	.985	.614	.542	410	.968	.136	.145	51	.990	.965	.975	264	.933	.828	.889	162
DET	.956	.201	.401	1165	.976	.358	.596	886	.926	.211	.334	583	.916	.089	.141	75	.998	.924	.927	777	.987	.759	.887	522
INTJ	.814	.048	.060	31	1.000	.000	.093	4	.000	.000	.000	1	.000	.222	.444	3	.884	.769	.345	6	.000	.000	.000	0
NOUN	.930	.522	.453	1593	.958	.767	.772	1436	.895	.666	.724	1024	.846	.501	.514	123	.982	.925	.936	1570	.983	.934	.946	699
NUM	.792	.156	.033	49	.962	.651	.571	94	.000	.000	.000	0	.000	.000	.000	0	.955	.791	.852	59	.909	.667	.909	6
PART	.950	.040	.000	181	.978	.039	.132	92	.929	.000	.000	35	.899	.186	.000	16	.996	.195	.603	190	.967	.123	.230	100
PRON	.965	.464	.481	837	.991	.736	.842	1040	.958	.646	.696	468	.957	.624	.613	74	.993	.878	.888	1009	.988	.761	.781	283
PROPN	.941	.500	.348	226	.980	.805	.863	346	.852	.543	.596	144	.943	.245	.154	17	.977	.729	.844	239	.774	.571	.691	35
SCONJ	.846	.014	.048	268	.957	.066	.064	295	.954	.100	.564	315	.746	.000	.261	14	.959	.095	.767	139	.931	.171	.804	55
VERB	.944	.502	.507	1418	.969	.760	.713	1270	.925	.691	.703	986	.869	.411	.511	116	.971	.805	.917	1071	.942	.699	.771	311
X	.000	.000	.000	0	.968	.000	.046	84	.848	.040	.071	29	.000	.000	.000	0	.841	.078	.200	4	.711	.267	.250	3

Table 9: Per-tag F1 scores and corresponding support values for POS tagging with fine-tuned mBERT and generative models (Qwen, Apertus) in the **zero-shot** setting on **prose** test sets.

From Manuscript to Model: Developing HTR for Medieval Greek

Nicklas Sindlev Andersen[1], **Byron MacDougall**[1], **Tariq Yousef**[1], **Aglae Pizzone**[1]
[1] University of Southern Denmark, Odense, Denmark
sindlev@imada.sdu.dk, bmacdoug@risd.edu, yousef@imada.sdu.dk, pizzone@sdu.dk

Abstract

We develop and evaluate manuscript-specific text line detection (TLD) and handwritten text recognition (HTR) models for two 14th-century Medieval Greek manuscripts, *Vat. gr. 2228* and *Phil. gr. 130*, comprising 1,356 handwritten pages. From these, we curate and document 36 pages with complete, manually curated text line annotations, together with 10 additional pages with layout annotations only for TLD, forming two manuscript-specific ground truth (GT) datasets. To ensure representative evaluation despite limited annotations, validation splits are optimized for character coverage and distributional similarity using Jensen-Shannon divergence. Using the Transkribus platform, we train manuscript-specific TLD models from scratch and manuscript-specific HTR models, comparing HTR training from scratch with fine-tuning of a publicly available Medieval Greek base model. TLD achieves validation pixel-wise misclassification rates of 5.42% for Vat. gr. 2228 and 8.76% for the more layout-variable Phil. gr. 130. For HTR, fine-tuning consistently outperforms training from scratch. On validation pages with manually curated text line annotations, Vat. gr. 2228 reaches 5.13% character error rate (CER) and 23.66% word error rate (WER), while Phil. gr. 130 reaches 27.13% CER and 65.72% WER after continued training. A supplementary held-out evaluation on Vat. gr. 2228 shows that the fine-tuned model reaches 5.97% CER and 23.52% WER on test pages with manually corrected line polygons, degrading to 12.12% CER and 44.21% WER under automatic TLD-based segmentation. The study also provides a reproducible workflow and evaluation protocol for Medieval Greek HTR under low-resource conditions.

Keywords: Medieval Greek, Handwritten Text Recognition, HTR, Cultural Heritage Digitization, Datasets

1. Introduction

The digitization and transcription of historical documents have become increasingly important for both preserving cultural heritage and enabling new research. By converting physical materials into digital formats, these efforts not only safeguard unique historical sources but also broaden access. Manual transcription, however, is a time-consuming and highly specialized task, making automated approaches, particularly handwritten text recognition (HTR), increasingly valuable.

Medieval Greek manuscripts, in particular those featuring minuscule scripts penned by informal scholarly hands, present a unique set of difficulties. On top of the rich system of diacritics characterizing ancient and medieval Greek, those scripts, designed for rapid execution, present a high degree of cursivity, frequent ligatures, and extensive use of tachygraphic symbols and abbreviations. They also exhibit considerable variation in scribal execution and, depending on the textual typology, complex layouts. Such features further increase transcription complexity. While HTR systems have advanced considerably in recent years, progress has been uneven across languages and scripts. Compared to Latin-based languages, non-Latin scripts such as Medieval Greek still lag behind in terms of available datasets, models, and tools, often requiring researchers to develop resources in relative isolation.

Recently, however, the situation has begun to change. Collaborative communities have started to emerge, bringing together datasets, annotation guidelines, and trained models for Medieval Greek transcription (see, e.g., Table 1) (Verstraete et al., 2025). At the same time, user-friendly platforms such as Transkribus (READ-COOP SCE, 2025) have democratized access to advanced machine learning techniques, enabling broader participation in the transcription process. These developments are transforming archives that were previously hardly accessible to the non-specialist into analyzable data and streamlining research workflows.

In this work, we (1) curate and document two Medieval Greek ground truth (GT) datasets consisting of annotated manuscript pages with line-level text line annotations (used here as a catch-all term for baselines, enclosing line polygons, and transcriptions), covering pages from *Vat. gr. 2228*[1] and *Phil. gr. 130*[2]; (2) develop and evaluate manuscript-specific text line detection (TLD) models trained from scratch and manuscript-specific HTR models in Transkribus, comparing HTR training from scratch with fine-tuning of a publicly available Medieval Greek base model (including continued training where applicable), and reporting Transkribus validation metrics for TLD and externally computed recognition scores, namely character error rate (CER) and word error rate (WER), for HTR; and

[1] Vatican Library: Vat. gr. 2228, part 1 + part 2
[2] Austrian National Library: Phil. gr. 130

Proceedings of LT4HALA 2026 @ LREC 2026, pages 139–151.
11 May 2026. *©ELRA Language Resources Association (ELRA), 2026*

Table 1: Datasets and models for Ancient and Medieval Greek HTR, including the two datasets presented in this paper (Vat. gr. 2228 and Phil. gr. 130). In the table, D = dataset, and M = model. For Vat. gr. 2228, numbers in parentheses indicate additional pages and lines annotated only for TLD, not transcription.

Name	Type	Period	Pages	Lines	Characters (total / unique)
Zenon Papyri (Marthot-Santaniello and Hodel, 2022)	D	3rd BCE	27	321	6,059 / 47
Méléagre (Guénette et al., 2024)	D + M	10th CE	70	3374	108,231 / 108
Chrysostomicus I (Perdiki, 2022)	M	10-14th CE	–	–	– / –
Stavronikita (no. 53) (Pratikakis et al., 2021b)	D	14th CE	54	1038	63,228 / 147
Stavronikita (no. 114) (Pratikakis et al., 2021a)	D	15th CE	44	1006	66,086 / 159
Eparchos (Papazoglou et al., 2020)	D	16th CE	120	2272	203,338 / 156
Stavronikita (no. 79) (Pratikakis et al., 2021c)	D	16th CE	40	803	50,422 / 133
HPGTR (Platanou et al., 2024)	D	8-16th CE	70	1698	66,739 / 159
Vat. gr. 2228	D	14th CE	18 (+10)	731 (+439)	55,637 / 144
Phil. gr. 130	D	14th CE	18	847	123,908 / 143

(3) systematically document the end-to-end workflow from GT preparation to evaluation to support reproducibility and comparative analysis in future studies.

2. Related Work

2.1. HTR Platforms

HTR platforms integrate the core stages of the HTR workflow, typically combining layout analysis, often focusing on text line detection (TLD), with handwritten text recognition. Many platforms also support model training and fine-tuning through graphical or web-based interfaces (Boros et al., 2024).

These systems differ primarily in their trade-off between usability and experimental control. Hosted, closed platforms reduce setup overhead but may limit automation, transparency, and reproducibility. In contrast, open-source systems with application programming interface (API) access enable customizable and fully automated pipelines, at the cost of local infrastructure requirements and technical effort.

In this work, we use Transkribus (READ-COOP SCE, 2025), a widely adopted hosted platform whose HTR stack is based on the open-source PyLaia engine (Teklia, 2022; Tarride et al., 2024). While Transkribus facilitates rapid iteration and collaborative annotation, its closed environment constrains hyperparameter control and full reproducibility. We therefore account for these limitations in our evaluation protocol by performing repeated training runs and carefully documenting configuration settings.

Open-source alternatives such as eScriptorium (Kraken) (eScriptorium, 2025; Kiessling et al., 2019), Arkindex (PyLaia) (Teklia, 2025), and OCR4all (Centre for Philology and Digitality, 2025) provide greater transparency and automation, but typically require institutional hosting or dedicated compute resources for training at scale.

Together, these platforms form the technical foundation upon which most recent work in Ancient and Medieval Greek HTR is built. Understanding their respective strengths and limitations helps contextualize the methodological choices in the current study.

2.2. Ancient & Medieval Greek HTR

Recent efforts in Ancient and Medieval Greek HTR have produced a range of datasets and trained models across different platforms and architectural approaches.

Platanou et al. (2024) introduced the HPGTR dataset, constructed from digitized images of Bodleian Library Greek manuscripts dating from the 10th to the 16th century, and trained recognition models using Transkribus to assess HTR performance. Based on their findings, they argue that paleographic style and manuscript dating strongly influence recognition performance.

In particular, they correlated variation to chronology: manuscripts from the 10th-13th centuries achieved lower character error rates, whereas later manuscripts (14th-16th CE) proved more challenging due to an increasingly cursive style, ligatures, and stylistic complexity. These findings seem to underscore the importance of accounting for temporal and stylistic variation when developing manuscript-specific HTR models, although the question remains whether variations due to genres and uses might not be equally, if not more important.

Addressing scalability, Perdiki (2023) investigated how to minimize training requirements for large collections. Focusing on John Chrysostom's manuscript tradition, they showed that usable Transkribus models can be trained with as little as 1,000 words of GT. Their results highlight the value of careful manuscript selection and transfer learning strategies, particularly when annotated data are limited.

At the level of individual manuscripts, Guénette et al. (2024) studied Codex Palatinus Graecus 23 using eScriptorium. They trained recognition models both from scratch and via fine-tuning (initializing from the CREMMA Medieval Latin model (Pinche, 2022)), achieving approximately 90-91% accuracy. Their work demonstrates both the feasibility of

cross-script transfer learning and the importance of capturing scribal diversity within a manuscript.

From a methodological perspective, Markou et al. (2021) proposed a CRNN-FCNN architecture and introduced the EPARCHOS dataset (16th CE), which captures key challenges of Greek manuscripts, including abbreviations, floating characters, and polytonic orthography. Their ablation study showed that data augmentation and regularization significantly improved recognition performance. Building on convolutional-recurrent architectures, Tsochatzidis et al. (2021) introduced an OctCNN-BGRU model and released additional benchmark datasets from the Stavronikita Monastery (nos. 53, 79, and 114).

Extending beyond medieval manuscripts, and moving backward in time, the Zenon Papyri dataset developed by Marthot-Santaniello and Hodel (2022) within Transkribus as part of the D-Scribes project (D-Scribes Project, 2023) represents a key contribution to Ancient Greek HTR. As one of the few publicly available resources for Greek scripts preserved on papyri, it provides a foundation for the study of significantly earlier material.

Taken together, these studies highlight three recurring observations: (i) recognition performance is strongly influenced by paleographic and layout variation; (ii) transfer learning and fine-tuning consistently outperform training from scratch, particularly when annotated data are limited; and (iii) carefully curated, manuscript-specific GT subsets can be sufficient to train usable HTR models.

Our work builds directly on these observations by adopting a manuscript-specific modeling strategy, comparing training from scratch with fine-tuning of a publicly available Medieval Greek base model, and relying on carefully curated but relatively small annotated subsets.

3. Data

3.1. Source Manuscripts

We develop workflows and models for an overall corpus of approximately 1,356 handwritten pages drawn from two 14th-century Medieval Greek manuscripts: Vatican, Biblioteca Apostolica Vaticana, Vat. gr. 2228 and Vienna, Österreichische Nationalbibliothek, Phil. gr. 130 (see Figure 1). Within the broader manuscript project motivating this study, however, the current phase of the HTR work reported here focuses on specific folio ranges within these manuscripts.

Vat. gr. 2228 is a paper manuscript measuring 231×156 mm and consisting of 508 folia (1,016 pages), currently divided into two volumes. Each folio, with the exception of folia 1r-v and 16r-v, contains 35 to 45 lines of writing on a writing surface of mostly 174×113.118 mm or 166×114.115 mm, according to the full description provided by Salvator Lilla in the catalog of the Greek Vatican manuscripts (Lilla, 1985, pp. 307–313). The layout follows the ruling cataloged by Leroy as 20 D 1 (Leroy, 1976, p. 6) (see Figure 2 for the ruling pattern). The manuscript, copied by nine different hands, offers a selection of rhetorical treatises featuring prominently the Graeco-Roman handbooks authored by Hermogenes of Tarsus (2nd-3rd century CE) and his medieval exegeses. For the current phase, our primary focus is folia 194–313, copied by Hand 2 (according to Lilla's description) and containing the 11th-century commentary by John Doxapatres on Hermogenes' treatise *On Invention*, a commentary that remains unpublished to date. At the same time, the availability of published sections from an earlier portion of the manuscript makes it possible to construct additional GT for model development before targeting this main range of interest.

Phil. gr. 130 is also a paper manuscript dated to the first half of the 14th century. It consists of 170 folia (340 pages) measuring 240/245×155/160 mm. Each folio contains 38 to 56 or 55 to 62 lines of writing, depending on the dimensions of the folio. The manuscript is described by Herbert Hunger (Hunger, 1961, pp. 238–239) in the catalogue of the Greek manuscripts preserved by the Austrian National Library. Like the Vatican manuscript, the Viennese manuscript contains Graeco-Roman rhetorical treatises accompanied by medieval Byzantine commentaries. For the purpose of this study, we focused on folia 85v-170v, containing Hermogenes' treatise *On Issues*, with the exegesis of John Doxapatres. The commentary material is organized in a frame layout on the entirety of the written surface, whereby the commentary forms an open frame around small blocks of main text (Maniaci, 2022).

The particular reproductions of Vat. gr. 2228 we have been annotating are high-resolution grayscale images obtained from the Vatican Library, each measuring 2174×2996 px, while for Phil. gr. 130, the corresponding reproductions consist of high-resolution color images, ranging from 4032-4179 px in width and 6040-6086 px in height.

The broader underlying aim of this work is to facilitate access to the currently unpublished commentary of John Doxapatres, relevant to the history of medieval rhetoric. The material considered here reflects variation in textual density, layout complexity, and handwriting. In particular, Phil. gr. 130 exhibits greater variability in both layout and handwriting, whereas Vat. gr. 2228 is comparatively consistent and visually regular. This difference between the two manuscripts, despite their closeness in time, is reflected in model performance, as discussed in Section 5.

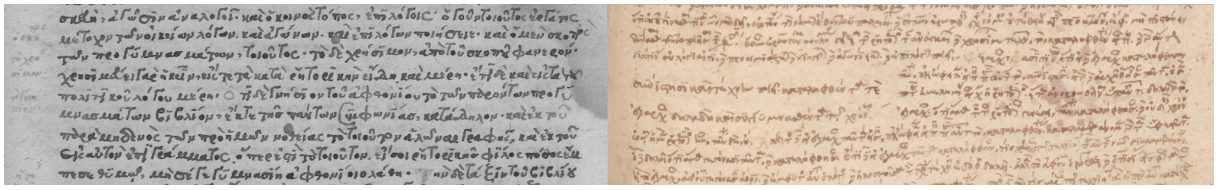


Figure 1: Excerpts from Vat. gr. 2228 (left) and Phil. gr. 130 (right). The image from Vat. gr. 2228 is generally representative due to the manuscript's consistency, while Phil. gr. 130 exhibits more variability.

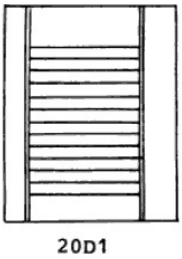


Figure 2: The ruling pattern of Vat. gr. 2228.

3.2. The Ground Truth & Splits

The two GT datasets consist of manuscript pages with line-level text line annotations.

Pages selected for annotation were chosen based on editorial needs (for Phil. gr. 130) and the availability of published sections of the manuscript text (for Vat. gr. 2228).

The GT datasets are distributed as PAGE-XML and split-definition files[3] with no manuscript images, since image rights remain with the holding institutions. In total, the GT release contains 46 PAGE-XML files: 18 for Phil. gr. 130 and 28 for Vat. gr. 2228.

The GT for Vat. gr. 2228 covers folia 12r–20v and 194r–199v. Of these, 18 unique pages are fully annotated at the line level. The GT was established by comparing folia 12r–20v with the existing edition by Christian Walz (Walz, 1835) and folia 194r-v by applying the workflow described in Section 4.1. An additional 10 pages (folia 195r–199v) contain layout annotations only (baselines and enclosing line polygons) and are used exclusively for TLD, contributing 439 lines but no transcriptions (these are shown in parentheses in Table 1).

For Phil. gr. 130, the GT covers folia 86r–94v, all fully annotated from scratch at the line level.

Because annotated data are limited, we partition the pages used for model development into a training set (≈ 90%) and a validation set (≈ 10%). For Phil. gr. 130, this corresponds to 16 training pages and 2 validation pages; for Vat. gr. 2228, 14 training pages and 2 validation pages are used, with 2 additional pages held out for test evaluation. For Vat. gr. 2228, these held-out test pages became available only in a later iteration of the workflow described in Section 4.1. They were therefore excluded when selecting the training/validation split. Accordingly, all splits are performed at the page level to prevent leakage across partitions.

All transcriptions are normalized prior to defining the character vocabulary used for training and evaluation. Normalization removes zero-width characters, maps Unicode whitespace to a single ASCII space, strips leading and trailing whitespace, and collapses consecutive spaces. In this context, let $\mathcal{C}$ denote the resulting normalized character vocabulary of a manuscript corpus, i.e., the set of characters occurring in the GT transcriptions.

Given the small number of annotated pages, evaluation may vary depending on the specific pages selected for validation. To mitigate this effect and promote representativeness, validation pages are selected according to two complementary criteria:

Character coverage. The validation set should contain as many unique characters present in the corpus as possible. For a subset S of pages, let $\mathcal{C}(S) \subseteq \mathcal{C}$ denote the set of characters occurring in S. We define the coverage of S as

$$\text{Coverage}(S) = \frac{|\mathcal{C}(S)|}{|\mathcal{C}|}.$$

A value of 1 indicates that all character types in $\mathcal{C}$ are represented in the validation set.

Distributional similarity. Beyond presence or absence, the relative character frequencies in the validation set should approximate those of the full corpus to avoid biased evaluation. Let P and Q denote the empirical character distributions over $\mathcal{C}$ for the subset and the full corpus, respectively. We then measure similarity using the symmetric Jensen-Shannon divergence:

$$D_{\text{JS}}(P \,\|\, Q) = \tfrac{1}{2} D_{\text{KL}}(P \,\|\, M) + \tfrac{1}{2} D_{\text{KL}}(Q \,\|\, M),$$

where D_{KL} is the Kullback-Leibler divergence (measuring the difference between two probability distributions) and $M = \frac{1}{2}(P + Q)$ is their pointwise average. Smaller D_{JS} values indicate closer agreement between the character frequency distributions of the validation set and the full corpus.

[3] https://doi.org/10.5281/zenodo.19425687.

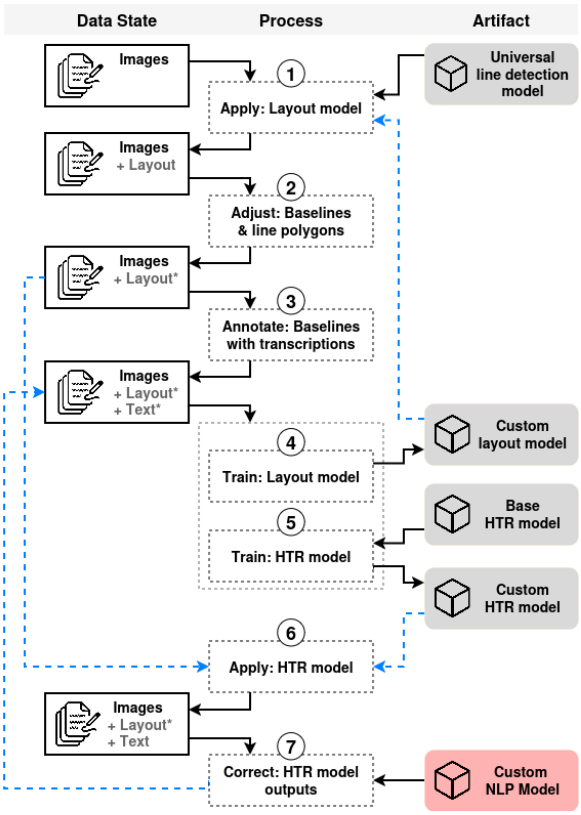


Figure 3: The workflow applied to Vat. gr. 2228 and Phil. gr. 130. Left column: an evolving GT dataset. Middle column: the processing steps. Right column: the resulting artifacts (applied or newly trained models). In the diagram, +Layout/+Layout* denote unadjusted/adjusted baselines and line polygons, and +Text/+Text* unadjusted/adjusted transcriptions. Blue dashed arrows indicate the iterative refinement cycle, where adjusted outputs are reintegrated into a GT dataset and models retrained.

Validation sets are selected using a two-step procedure. We first identify page combinations that maximize character coverage; among those, we select the subset minimizing D_{JS}. This approach yields validation sets that include rare characters while maintaining a character frequency distribution closely aligned with the overall corpus.

4. Workflow & Methods

4.1. Workflow in Transkribus

The workflow in Transkribus is illustrated in Figure 3. We initially employed the workflow consisting of steps ①-⑤, comprising data preparation and model development. The process began with uploading digitized manuscript images to the Transkribus web interface. A critical early step was layout analysis, here focusing specifically on TLD ①, which provided the structural basis for subsequent HTR processing.

For initial TLD, we applied the Transkribus *Universal Lines* model, which predicts baselines and corresponding line polygons of approximately rectangular shape. While generally effective, the automatically generated line polygons occasionally truncated characters, as letters in both manuscripts frequently extended beyond predicted boundaries. We therefore manually adjusted these annotations ②, producing high-quality layout annotations suitable for TLD training and for subsequent pairing with transcriptions in HTR training.

Expert transcriptions were then created on a per-line basis, pairing each annotated text line with its corresponding machine-readable transcription ③. This pairing of precise segmentation and diplomatically accurate transcription formed the foundation of the manuscript-specific GT.

GT preparation was carried out collaboratively by a domain expert in Medieval Greek and a data scientist. The domain expert ensured accurate transcription, including diacritical marks, resolutions of tachygraphic signs, ligatures, floating characters, and other paleographic features, while refining the associated layout annotations. The data scientist oversaw model application, dataset organization, and quality control within Transkribus in preparation for model training ④-⑤. All annotated pages were cross-checked to ensure transcription accuracy and annotation consistency.

In addition to steps ①-⑤, we employed steps ⑥ and ⑦, applying trained TLD and HTR models to previously unannotated pages. Together, steps ①-⑦ form an iterative learning cycle (illustrated in Figure 3 by blue dashed arrows), particularly for Vat. gr. 2228. Automatically segmented and transcribed pages were manually corrected and incorporated into the respective manuscript-specific GT dataset, after which updated models were retrained. This iterative refinement was applied independently to each manuscript, progressively improving TLD and recognition quality while expanding the annotated material, until converging on final manuscript-specific GT datasets (Section 3.2).

4.2. TLD Model Training

In parallel with HTR model training ⑤, we trained manuscript-specific TLD models ④ for each manuscript.

TLD models were trained using all default settings in Transkribus and using the same training and validation partitions as the HTR models (Section 3.2). Their primary purpose was to enable automated segmentation of the remaining unannotated manuscript pages.

Accurate TLD is a prerequisite for reliable HTR performance, as misplaced or truncated line polygons result in incomplete character capture and

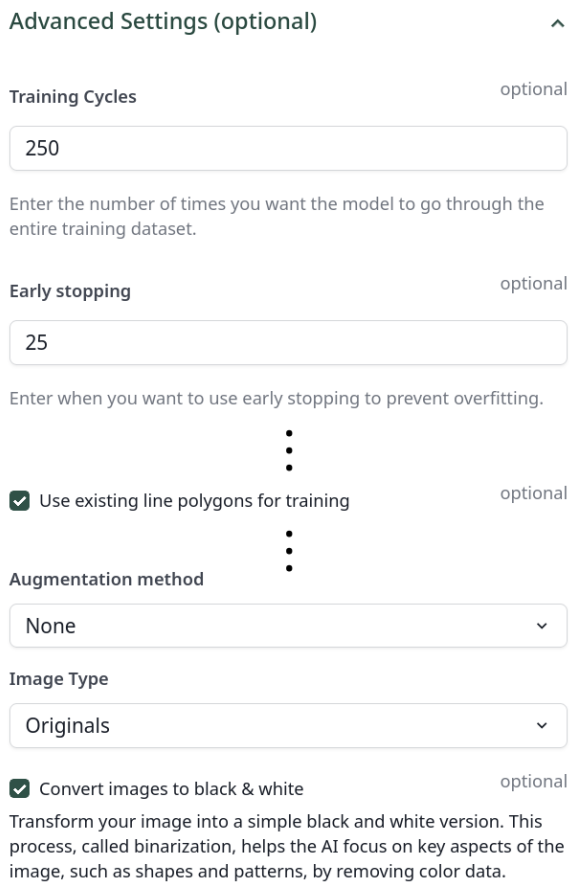


Figure 4: Screenshot of Transkribus HTR "Advanced Settings" used in the various experiments.

subsequent recognition errors (Fizaine et al., 2024). Manuscript-specific training therefore aimed to reduce systematic segmentation artifacts introduced by generic models (Universal Lines) available in Transkribus.

4.3. HTR Model Training

HTR model training in Transkribus ⑤ was carried out using the PyLaia engine, which implements a convolutional-recurrent neural network (CRNN) architecture. The model consists of a convolutional neural network (CNN) backbone for spatial feature extraction, followed by stacked bidirectional long short-term memory (BiLSTM) layers for sequence modeling. Recognition is optimized using the connectionist temporal classification (CTC) loss function, which resolves alignment between image features and character sequences without requiring pre-segmented character boundaries.

For each manuscript, we compared two training strategies: (i) training from scratch using only the manuscript-specific GT, and (ii) fine-tuning a publicly available Medieval Greek base model (transfer learning) (Perdiki, 2022). Where applicable, we further explored continued training to assess any further possible performance gains.

Unless otherwise noted, each training strategy was configured for up to 250 epochs with early stopping after 25 epochs without improvement on the validation set (see Figure 4). Advanced settings included using existing layout annotations during training and converting images to black and white. By binarizing both manuscripts, we aimed to enforce a consistent starting point for model training and evaluation, avoiding differences in color information as an additional source of variation across pipelines. Validation followed the page-level 90/10 split described in Section 3.2, ensuring separation between training and validation pages.

As additional corrected pages were incorporated into the GT datasets, models were retrained to reflect the expanded and refined training data. The results reported in Section 5 correspond to a final training phase conducted after this iterative refinement process had converged on the current GT datasets, described in Section 3.2.

4.4. Evaluation Protocol & Error Metrics

Because Transkribus training is non-deterministic, each training strategy was repeated three times. Evaluation results reported in Tables 2 and 3 therefore correspond to mean and standard deviation across three independent runs, training a model on the training set and evaluating it on the validation set.

TLD Evaluation. TLD performance is reported as the pixel-wise misclassification rate provided by Transkribus and retrieved directly from the platform's web interface. Lower values indicate better segmentation performance[4].

HTR Evaluation. Evaluation was performed externally to Transkribus to ensure reproducible scoring and to enable future, more extensive, line- and character-level error analysis. GT and predicted outputs were exported as PAGE-XML and aligned one-to-one at the text-line level. For all evaluated pages, the number of GT and predicted lines matched exactly.

Let g_i and h_i denote the normalized GT and predicted strings for line $i = 1, \ldots, N$. Before scoring, GT and predictions were normalized as described in Section 3.2, ensuring consistency with the vocabulary construction and dataset splits.

Recognition performance was quantified using character error rate (CER) and word error rate (WER), computed as micro-averaged Levenshtein

[4] https://help.transkribus.org/baselines-models

Table 2: TLD performance reported as a pixel-wise misclassification rate. Results are reported as mean ± std over three training runs.

Manuscript	Miscl. Rate	
	Train	Val
Vat. gr. 2228	6.36% ± 0.07%	5.42% ± 0.21%
Phil. gr. 130	4.86% ± 0.01%	8.76% ± 0.30%

distance with unit insertion, deletion, and substitution costs (no transpositions):

$$\mathrm{CER} = \frac{\sum_{i=1}^{N} d_{\mathrm{Lev}}(g_i, h_i)}{\sum_{i=1}^{N} |g_i|},$$
$$\mathrm{WER} = \frac{\sum_{i=1}^{N} d_{\mathrm{Lev}}(\tau(g_i), \tau(h_i))}{\sum_{i=1}^{N} |\tau(g_i)|}.$$

Here, $|g_i|$ denotes the character length of g_i, $\tau(\cdot)$ denotes whitespace tokenization, and micro-averaging means that edit operations are aggregated across all lines before normalization by total reference length.

In addition to the validation-based comparison used for model development, we report a supplementary held-out evaluation on unseen Vat. gr. 2228 test pages. We restrict this additional analysis to Vat. gr. 2228 because held-out test pages became available for this manuscript through an additional iteration of the workflow described in Section 4.1, and recognition quality had reached a level that made assessment under practical segmentation conditions meaningful. The same three trained HTR models per strategy were reused. Each model was then evaluated under two segmentation settings: manually corrected line polygons, representing a near-best-case recognition condition, and automatically detected line polygons produced by the manuscript-specific TLD model with the lowest validation misclassification rate, representing a more realistic end-to-end condition. Scoring followed exactly the same external PAGE-XML protocol as for validation.

5. Results

5.1. TLD Performance

Table 2 reports overall TLD results, while per-run results are provided in the Appendix (Section 11).

For Vat. gr. 2228, training and validation misclassification rates are comparable (6.36% vs. 5.42%), suggesting similar performance across the selected training and validation pages. This pattern is consistent with the manuscript's comparatively regular layout and homogeneous line structure.

In contrast, Phil. gr. 130 shows a larger gap between training and validation error (4.86% vs. 8.76%), suggesting reduced generalization. This behavior aligns with the manuscript's greater layout variability and scribal complexity.

On qualitative inspection of unseen pages, the manuscript-specific models generally capture the main body of text lines in both manuscripts. Residual errors appear primarily boundary-related, including partial exclusion of diacritics and occasional overlap with adjacent lines. Ultimately, automatically generated layout annotations remain below the quality of the manual layout annotations.

5.2. HTR Performance

Table 3 reports the overall results, while per-run results can be found in the Appendix (Section 11).

For Vat. gr. 2228, fine-tuning the publicly available Medieval Greek base model yields the best validation performance (5.13% CER, 23.66% WER), improving over training from scratch (6.23% CER, 28.26% WER). Training error remains low overall, though the from-scratch runs show noticeably larger variance than the fine-tuned runs.

Recognition of Phil. gr. 130 is substantially more challenging. Training from scratch results in 38.20% CER and 77.95% WER. Fine-tuning reduces validation CER to 28.62% (WER: 68.18%), and continued training further lowers it to 27.13% (WER: 65.72%). Although overall error remains high, transfer learning consistently improves performance, with the largest absolute gains observed for the more complex manuscript, indicating that a pretrained model can provide a beneficial initialization even under paleographic and structural variation.

Table 4 extends the validation results with a supplementary held-out evaluation for Vat. gr. 2228 under two segmentation conditions. On the held-out test pages with manually corrected line polygons, the fine-tuned base model achieved 5.97% CER and 23.52% WER, compared with 8.77% CER and 31.14% WER for training from scratch. Under automatically detected line polygons, recognition performance dropped to 12.12% CER and 44.21% WER for the base model, and to 18.47% CER and 54.81% WER for training from scratch. The relative ranking observed on validation is therefore preserved on unseen material, while the end-to-end setting makes explicit the expected propagation of segmentation errors into downstream HTR.

6. Discussion & Limitations

Because annotated material is limited, we use a fixed page-level split optimized for character coverage and distributional similarity. While this ensures textual representativeness, it does not explicitly account for visual variability. Reported validation scores should therefore be interpreted as estimates

Table 3: HTR performance reported as micro-averaged CER/WER. Mean ± std over three training runs.

Manuscript	Strategy	CER		WER	
		Train	Val	Train	Val
Vat. gr. 2228	From Scratch	0.68% ± 0.55%	6.23% ± 0.22%	3.91% ± 2.84%	28.26% ± 1.41%
	Base Model	0.02% ± 0.01%	5.13% ± 0.09%	0.15% ± 0.07%	23.66% ± 0.60%
Phil. gr. 130	From Scratch	37.42% ± 3.45%	38.20% ± 1.98%	66.80% ± 4.01%	77.95% ± 2.78%
	Base Model	32.86% ± 0.53%	28.62% ± 0.33%	57.28% ± 0.41%	68.18% ± 0.77%
	Base Model (Cont. Train.)	27.95% ± 0.05%	27.13% ± 0.43%	47.69% ± 0.13%	65.72% ± 1.64%

Table 4: Supplementary held-out evaluation of HTR performance for Vat. gr. 2228 on two test settings: manually corrected line polygons and automatically detected line polygons from the manuscript-specific TLD model with the lowest validation misclassification rate. Results are reported as micro-averaged CER/WER, mean ± std over three training runs.

Manuscript	Strategy	CER		WER	
		Manual	Automatic (TLD)	Manual	Automatic (TLD)
Vat. gr. 2228	From Scratch	8.77% ± 0.45%	18.47% ± 0.83%	31.14% ± 0.47%	54.81% ± 0.29%
	Base Model	5.97% ± 0.16%	12.12% ± 0.64%	23.52% ± 0.68%	44.21% ± 2.50%

under this controlled setting rather than guarantees for all pages within each manuscript.

A second limitation concerns experimental control in Transkribus. Model training is non-deterministic because random seeds cannot be fixed. Although we repeat each strategy three times and report mean and standard deviation, run-to-run variance remains. In addition, a single fixed validation split increases sensitivity to page selection and does not capture variability that cross-validation would reveal.

Finally, most of the recognition results are computed on validation pages with manually curated text line annotations. We supplement these with a held-out evaluation for Vat. gr. 2228 under both manual and automatic segmentation conditions, but this more realistic end-to-end setting is so far assessed only for one manuscript and on a small number of test pages. Extending and expanding this type of end-to-end evaluation therefore remains a natural next step, in particular for Phil. gr. 130, where layout and handwriting are more variable.

7. Conclusion & Future Work

We documented two manuscript-specific Medieval Greek GT datasets, including fully annotated pages with manually curated text line annotations and additional layout-only pages for TLD, trained manuscript-specific TLD and HTR models in Transkribus, and described an iterative workflow and evaluation protocol for Medieval Greek HTR under low-resource conditions. Across both manuscripts, fine-tuning a publicly available Medieval Greek base model consistently improved recognition performance over training from scratch, with the largest gains observed for the more complex Phil. gr. 130. TLD results likewise reflected the contrast between the two manuscripts, with more comparable training and validation performance for Vat. gr. 2228 than for the more layout-variable Phil. gr. 130.

Future work will focus on four directions. (1) Data and imaging quality: extend the annotated corpus, especially for Phil. gr. 130, and assess the effect of higher-quality scans on both segmentation and recognition. (2) End-to-end evaluation: extend the supplementary held-out evaluation reported here by quantifying the interaction between TLD quality and HTR performance on automatically segmented pages beyond the current Vat. gr. 2228 test-set analysis. (3) Training improvements: explore data augmentation to improve robustness to layout and handwriting variation, like in Markou et al. (2021). (4) Post-correction: benchmark post-correction models that treat error correction as sequence-to-sequence translation. Byte-level models such as ByT5 (Xue et al., 2022) have shown robustness to noise and rare characters in related historical settings (Löfgren and Dannélls, 2024; Momtaz et al., 2025), and recent work suggests similar potential for Medieval Greek (Pavlopoulos et al., 2024, 2023) and for large language model (LLM)-based correction of Medieval Greek HTR output (Evangelatos et al., 2026). We aim to evaluate similar approaches on Vat. gr. 2228 and Phil. gr. 130 to reduce remaining errors.

8. Acknowledgements

The work presented here is supported by the Carlsberg Foundation, grant CF24-1999.

9. Bibliographical References

Emanuela Boros, Maud Ehrmann, Matteo Romanello, Sven Najem-Meyer, and Frédéric Kaplan. 2024. Post-correction of historical text transcripts with large language models: An exploratory study. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, page 133–159. Association for Computational Linguistics.

Centre for Philology and Digitality. 2025. Ocr4all. URL: https://www.ocr4all.org. Accessed: 2025-09-02.

D-Scribes Project. 2023. | Home | Ancient History | D-Scribes | University of Basel. URL: https://d-scribes.philhist.unibas.ch/en. Accessed: 2025-08-25.

eScriptorium. 2025. escriptorium. URL: https://escriptorium.rich.ru.nl. Accessed: 2025-09-02.

Andreas Evangelatos, Konstantinos Palaiologos, Basilis Gatos, Panagiotis Kaddas, Aikaterini Christopoulou, Vassilis Katsouros, and Andreas Kakridis. 2026. Using llms for improving the ocr accuracy of old greek handwritten documents. In *New Trends in Theory and Practice of Digital Libraries*, pages 100–109, Cham. Springer Nature Switzerland.

Florian Côme Fizaine, Patrick Bard, Michel Paindavoine, Cécile Robin, Edouard Bouyé, Raphaël Lefèvre, and Annie Vinter. 2024. Historical text line segmentation using deep learning algorithms: Mask-rcnn against u-net networks. *Journal of Imaging*, 10(3).

Maxime Guénette, Mathilde Verstraete, Alix Chagué, and Marcello Vitali-Rosati. 2024. Codex palatinus graecus 23 - Ground Truth Dataset Medieval Greek Manuscripts. URL: https://gitlab.huma-num.fr/ecrinum/anthologia/htr_cpgr23. Accessed: 2025-08-25.

Herbert Hunger. 1961. *Katalog der griechischen Handschriften der Österreichischen Nationalbibliothek*, volume 1. Prachner–Hollinek, Wien.

Benjamin Kiessling, Robin Tissot, Peter Stokes, and Daniel Stokl Ben Ezra. 2019. escriptorium: An open source platform for historical document analysis. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 19–19, Piscataway. IEEE.

Julien Leroy. 1976. *Les types de réglures des manuscrits grecs*. Centre national de la recherche scientifique, Paris.

Salvatore Lilla. 1985. *Codices Vaticani Graeci*. Bibliotheca Apostolica Vaticana, Vatican City.

Viktoria Löfgren and Dana Dannélls. 2024. Post-OCR correction of digitized Swedish newspapers with ByT5. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 237–242, St. Julians, Malta. Association for Computational Linguistics.

Marilena Maniaci. 2022. *Words within Words : Layout Strategies in Some Glossed Manuscripts of the Iliad*, pages 575–598. De Gruyter, Berlin, Boston.

K. Markou, L. Tsochatzidis, K. Zagoris, A. Papazoglou, X. Karagiannis, S. Symeonidis, and I. Pratikakis. 2021. A convolutional recurrent neural network for the handwritten text recognition of historical greek manuscripts. In *Pattern Recognition. ICPR International Workshops and Challenges*, pages 249–262, Cham. Springer International Publishing.

Isabelle Marthot-Santaniello and Tobias Hodel. 2022. Ground-truthed data set of zenon papyri for handwritten text recognition.

Yahya Momtaz, Lorenza Laccetti, and Guido Russo. 2025. Modular pipeline for text recognition in early printed books using kraken and byt5. *Electronics*, 14(15).

Aleksandros Papazoglou, Ioannis Pratikakis, Kleopatra Markou, and Lazaros Tsochatzidis. 2020. Eparchos - historical greek handwritten document dataset. Zenodo.

John Pavlopoulos, Vasiliki Kougia, Esteban Garces Arias, Paraskevi Platanou, Stepan Shabalin, Konstantina Liagkou, Emmanouil Papadatos, Holger Essler, Jean-Baptiste Camps, and Franz Fischer. 2024. Challenging error correction in recognised byzantine Greek. In *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)*, pages 1–12. Association for Computational Linguistics.

John Pavlopoulos, Vasiliki Kougia, Paraskevi Platanou, and Holger Essler. 2023. Detecting erroneously recognized handwritten byzantine text. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7818–7828, Singapore. Association for Computational Linguistics.

Elpida Perdiki. 2022. Chrysostomicus i | transkibus models. URL: https://app.transkribus.org/models/public/text/chrysostomicus-i. Accessed: 2025-08-25.

Elpida Perdiki. 2023. Preparing Big Manuscript Data for Hierarchical Clustering with Minimal HTR Training. *Journal of Data Mining and Digital Humanities*, Historical Documents and automatic text recognition.

Ariane Pinche. 2022. Cremma Medieval. URL: https://github.com/HTR-United/cremma-medieval.

Paraskevi Platanou, John Pavlopoulos, and Georgios Papaioannou. 2024. HPGTR: Handwritten Paleographic Greek Text Recognition dataset. URL: https://github.com/vivianpl/hpgtr. Accessed: 2025-08-25.

Ioannis Pratikakis, Aleksandros Papazoglou, Symeon Symeonidis, and Lazaros Tsochatzidis. 2021a. Stavronikita monastery greek handwritten document collection no.114.

Ioannis Pratikakis, Aleksandros Papazoglou, Symeon Symeonidis, and Lazaros Tsochatzidis. 2021b. Stavronikita monastery greek handwritten document collection no.53.

Ioannis Pratikakis, Aleksandros Papazoglou, Symeon Symeonidis, and Lazaros Tsochatzidis. 2021c. Stavronikita monastery greek handwritten document collection no.79.

READ-COOP SCE. 2025. Transkribus. URL: https://www.transkribus.org. Accessed: 2025-09-02.

Solène Tarride, Yoann Schneider, Marie Generali-Lince, Mélodie Boillet, Bastien Abadie, and Christopher Kermorvant. 2024. Improving automatic text recognition with language models in the pylaia open-source library. In *Document Analysis and Recognition - ICDAR 2024*, pages 387–404, Cham. Springer Nature Switzerland.

Teklia. 2022. Pylaia. URL: https://gitlab.teklia.com/atr/pylaia. Accessed: 2025-08-25.

Teklia. 2025. Arkindex: A document processing platform. URL: https://doc.teklia.com/arkindex. Accessed: 2025-09-02.

Lazaros Tsochatzidis, Symeon Symeonidis, Alexandros Papazoglou, and Ioannis Pratikakis. 2021. Htr for greek historical handwritten documents. *Journal of Imaging*, 7(12).

Mathilde Verstraete, Maxime Guénette, Malamatenia Vlachou-Efstathiou, Marianne Reboul, and Marcello Vitali-Rosati. 2025. Harmonizing Guidelines for Handwritten Text Recognition of Ancient Greek. URL: https://dhtr25.anthologiagraeca.org. Accessed: 2025-08-25.

Christian Walz. 1835. *Rhetores Graeci*, volume 2. Cotta.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

10. Appendix A: Data Splits

Figure 5 and Figure 6 illustrate the resulting character frequency distributions for the full corpus and the corresponding training and validation splits for Vat. gr. 2228 and Phil. gr. 130, respectively. In each figure, characters are ordered by frequency in the full corpus and split into the 50% most frequent and 50% least frequent characters. Relative frequencies are computed within each subset, enabling direct comparison independent of corpus size. For both manuscripts, the training distribution closely follows the full corpus distribution, while the validation set shows slightly weaker distributional alignment but still achieves good coverage across the full set of characters, including less frequent ones.

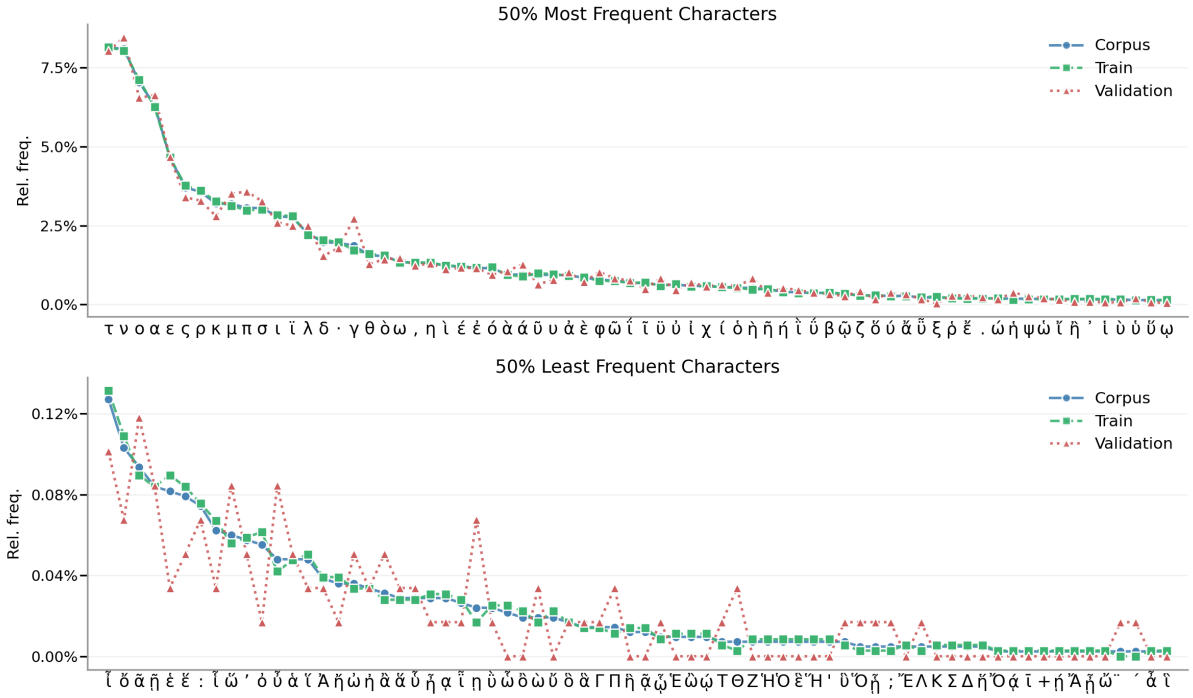


Figure 5: Character frequency distribution in Vat. gr. 2228, split into the 50% most frequent and 50% least frequent characters.

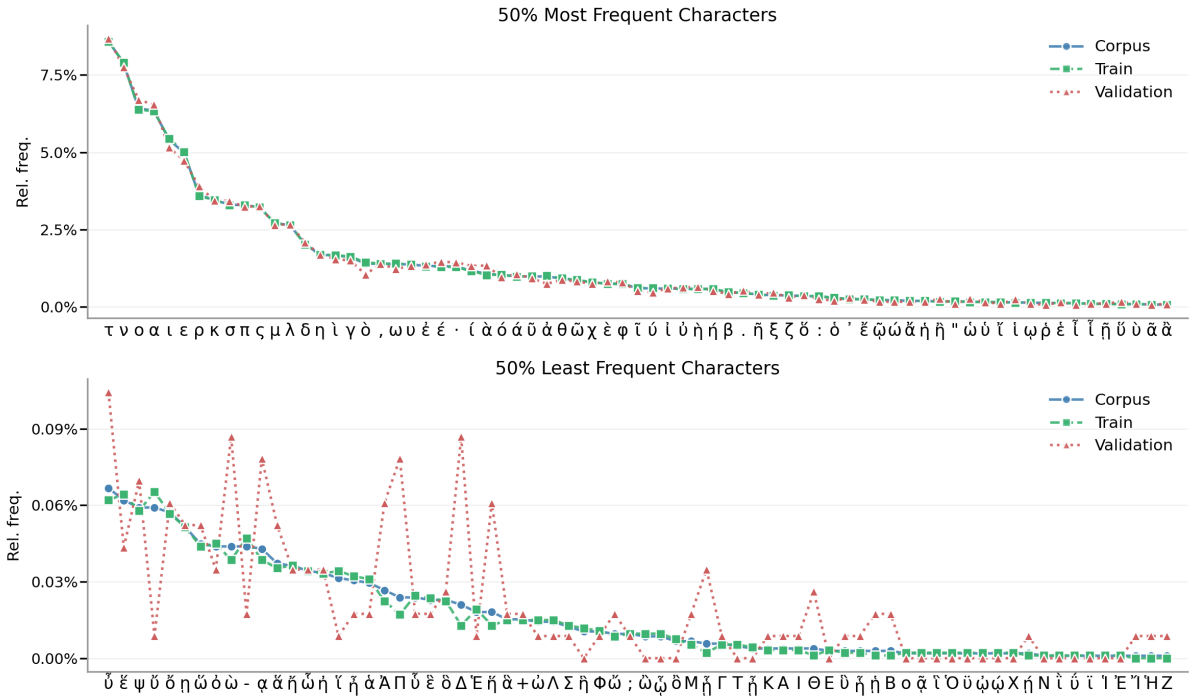


Figure 6: Character frequency distribution in Phil. gr. 130, split into the 50% most frequent and 50% least frequent characters.

11. Appendix B: Per-run Results

This appendix reports per-run results underlying the aggregated metrics presented in Section 5. We present per-run metrics for TLD, HTR on the training and validation splits, and the Vat. gr. 2228 HTR test-set evaluations. The tables reported here make more explicit the run-to-run variability arising from the non-deterministic training process.

Table 5: Per-run TLD performance reported as pixel-wise misclassification rate. Lower values indicate better segmentation performance.

Manuscript	Run	Miscl. Rate	
		Train	Val
Vat. gr. 2228	Run 1	6.32%	5.19%
	Run 2	6.44%	5.59%
	Run 3	6.32%	5.50%
Phil. gr. 130	Run 1	4.87%	8.93%
	Run 2	4.85%	8.41%
	Run 3	4.87%	8.93%

Table 6: Per-run HTR performance on the training and validation splits, reported as micro-averaged CER/WER.

Manuscript	Strategy	Run	CER		WER	
			Train	Val	Train	Val
Vat. gr. 2228	From Scratch	Run 1	0.363%	6.48%	2.21%	29.82%
		Run 2	1.320%	6.07%	7.19%	27.88%
		Run 3	0.367%	6.15%	2.33%	27.08%
	Base Model	Run 1	0.012%	5.05%	0.10%	23.45%
		Run 2	0.026%	5.22%	0.23%	24.34%
		Run 3	0.014%	5.12%	0.11%	23.19%
Phil. gr. 130	From Scratch	Run 1	36.59%	38.17%	65.75%	76.61%
		Run 2	34.46%	36.24%	63.43%	76.10%
		Run 3	41.21%	40.20%	71.23%	81.15%
	Base Model	Run 1	33.48%	28.88%	57.43%	68.69%
		Run 2	32.55%	28.74%	57.58%	68.55%
		Run 3	32.56%	28.25%	56.81%	67.30%
	Base Model (Cont. Train.)	Run 1	28.01%	26.63%	47.63%	64.47%
		Run 2	27.94%	27.41%	47.60%	65.12%
		Run 3	27.90%	27.34%	47.84%	67.58%

Table 7: Per-run HTR test-set performance for Vat. gr. 2228, reported as micro-averaged CER/WER for manually corrected line polygons and automatically detected line polygons from TLD.

Strategy	Run	CER		WER	
		Manual	Automatic (TLD)	Manual	Automatic (TLD)
From Scratch	Run 1	9.27%	19.09%	31.61%	54.75%
	Run 2	8.43%	18.78%	31.14%	54.56%
	Run 3	8.60%	17.53%	30.67%	55.13%
Base Model	Run 1	6.16%	12.27%	24.27%	44.50%
	Run 2	5.86%	12.67%	23.33%	46.57%
	Run 3	5.90%	11.42%	22.95%	41.58%

I, RE:Claudius 256: Towards Linking Classical Latin Person Mentions to a Domain-specific Knowledge Base

Marijke Beersmans, Evelien de Graaf, Julie Nijs, Valeria Irene Boano, Alek Keersmaekers, Mark Depauw, Tim Van de Cruys, Margherita Fantoli
KU Leuven
Blijde Inkomststraat 21, 3000 Leuven, Belgium
{marijke.beersmans, evelien.degraaf, julie.nijs}@kuleuven.be
{valeria.boano, alek.keersmaekers, mark.depauw}@kuleuven.be
{tim.vandecruys, margherita.fantoli}@kuleuven.be

Abstract

This paper considers Named Entity Linking for person mentions from classical Latin texts to a domain-specific, German language knowledge base, namely *Paulys Realencyclopädie*. Following a methodology similar to Beersmans et al. (2025), adapted to classical Latin, we train a transformer-based, retrieval and ranking model (BLINK) for this task. To mitigate data scarcity, we first train the model on a newly created general-purpose dataset derived from Wikipedia. We then fine-tune it on our domain-specific dataset, which is collected from various sources and linked to our target knowledge base. Results show that while BLINK performs well on mention-entity pairs linked to entities seen during training, it performs significantly worse on mention-entity pairs linking to unseen entities. We provide a detailed error analysis, propose possible exploitation strategies for a human-in-the-loop approach, and identify directions for future improvement.

Keywords: Named Entity Linking, BLINK, Classical Latin, Paulys Realencyclopädie, domain specific knowledge base

1. Introduction

Named Entity Linking (NEL) is the task of linking mentions in a running text to their corresponding entity in a predefined knowledge base (KB). As such, the texts are enriched with entity-based information, and the mentions are disambiguated.

As an example, in the passage:

- "After their time the courses of both stars for 600 years were prophesied by **Hipparchus**, whose work embraced the calendar of the nations and the situations of places and aspects of the peoples..." (*Naturalis Historia*, book II, section 9.53, Transl. Harris Rackham)

the mention *Hipparchus* should be linked to:

- Hipparchos 18, Gr. Astronom aus Nikaia in Bithynien (Hipparchos 18, Greek astronomer from Nicaea in Bithynia). (*Paulys Realencyclopädie Wikisource*)

and not to any of the other entries in the chosen KB (here, the Paulys Realencyclopädie, Wissowa et al. 1893–1980).

This paper considers NEL for classical Latin texts. The task is in high demand within the humanities, as a prerequisite for enhancing texts through structured searching or complex analysis, such as social network analysis. Currently, almost all Entity Linking in the context of historical texts is performed manually, which is labour-intensive and expensive (Ehrmann et al., 2023). A sufficiently accurate automatic NEL model could reduce these costs and save time, at least by providing reliable suggestions for a human-in-the-loop approach. We aim to address this gap from the context of the wider NIKAW project[1], which seeks to study the circulation of Knowledge based on the mentions of person names in texts from Greco-Roman antiquity.

Most research on automatic NEL relies on large, openly available, community-curated knowledge bases that cover entities from many different types and domains, most notably Wikipedia and related projects, such as Wikidata or DBPedia (Sevgili et al., 2022). While this approach is well motivated and widely adopted, we instead rely on a domain-specific knowledge base, namely the Wikisource edition of *Paulys Realencyclopädie* (*RE*) (in German). This choice offers more comprehensive and reliable entity coverage for antiquity, but it comes at the cost of lacking pre-existing resources such as Wikipedia's hyperlink structure or Wikidata's rich property annotations. In this paper, we assess the viability of BLINK (BERT entity Linking), a knowledge-base agnostic, transformer-based approach (Wu et al., 2020) for linking entities in classical Latin texts to a German-language knowledge base. We used a methodology similar to our previous work Beersmans et al. (2025), with the workflow adapted to classical Latin and an additional training

[1] https://research.kuleuven.be/portal/en/project/3H220323

Proceedings of LT4HALA 2026 @ LREC 2026, pages 152–162.
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

stage based on a Wikipedia-derived dataset (see Section 4 for more information). This constitutes a challenging, cross-lingual, and domain-specific setting. We train a model tailored to this setting while still relying on a Wikipedia derived dataset in a first training stage. We provide a detailed qualitative error analysis, identify bottlenecks, and suggest potential improvements. Code and data will be provided on Github.

The paper is structured as follows: we first discuss related work in Section 2. We then introduce the Knowledge Base and the enhancements made to it in Section 3. Subsequently, we detail our approach, data and model more thoroughly in Section 4. Afterwards, we report the results and error analysis in Section 5 and discuss our findings in Section 6. Finally, we conclude with Section 7.

2. Related Work

NEL for historical corpora is a particularly challenging task. Its complexity stems both from the long-tail distribution of entities (many appear only once and therefore provide very limited contextual information for disambiguation) and from the difficulty of relying on Wikipedia or Wikidata as knowledge bases. These resources, although widely used in general NEL and Wikification pipelines (i.e. Entity Linking to Wikipedia), often lack sufficient coverage of premodern or highly specialized historical entities.

Although the task of Wikification dates back to the early 2000s, a major shift occurred with the HIPE shared tasks in 2020 (Ehrmann et al., 2020) and 2022 (Ehrmann et al., 2022). These initiatives explicitly included NEL for historical corpora: newspapers from the 19th-20th centuries in 2020, and additionally modern commentaries to classical works in 2022. Already in the 2020 edition, some participants combined transformer models with filtering based on Wikipedia or Wikidata, such as Labusch and Neudecker (2020) and Boros et al. (2020), both discussed by Santini et al. (2026).

Among transformer-based approaches, BLINK has shown potential for NEL on Ancient Greek, though by no means solving the task (see Beersmans et al. 2025). Its key advantage lies in allowing the use of any knowledge bases, extending the task beyond standard Wikification. In this paper, we apply BLINK to a Latin corpus using a KB other than Wikipedia or Wikidata, while still exploiting the availability of the Latin Wikipedia to perform domain-specific pre-training on relevant entities. BLINK has since been extended in various directions: for example Graciotti et al. (2025a) introduced constraints to better handle non-linkable entities, achieving state-of-the-art performance. In parallel, generative models have been explored. GENRE (De Cao et al., 2020), and its multilingual adaptation mGENRE (De Cao et al., 2022), generate the target entity label directly as a sequence prediction task, rather than ranking pre-computed candidates. More recently, Graciotti et al. (2025b) evaluated BELA (Plekhanov et al., 2023), a multilingual bi-encoder based on BLINK (see Section 4.3 for an explanation of the functioning of BLINK), mGENRE, GPT-4o mini and LLaMa3-70B in zero-shot conditions for Wikidata-based NEL, with specific attention for datasets with long tails. Their experiments showed that generic Large Language Models such as GPT and LLaMa perform substantially worse on low-popularity entities. Building on these insights, Santini et al. (2026) proposed a hybrid framework that combines BELA, structured Wikidata metadata, and an instruction-tuned LLM adapted to the language of the corpus. However, this framework is explicitly tailored to Wikidata and cannot be directly applied to arbitrary KBs. Finally, to the best of our knowledge, no previous work on automated NEL has been undertaken for texts written in Latin.

In addition to neural approaches, string similarity methods (notably Levenshtein distance) remain common in NEL pipelines. Both Shen et al. (2015) and Sevgili et al. (2022), two comprehensive surveys of NEL, describe the use of string similarity for candidate generation and ranking, two core subtasks of entity linking. We therefore include a string similarity baseline in our experimental setup.

3. Knowledge Base : Paulys Realencyclopädie

Paulys Realencyclopädie der classischen Altertumswissenschaft, below *RE*, was originally published between 1893 and 1978. As an expansive reference encyclopedia on classical antiquity, it comprises around 100,000 entries contributed by prominent classical philologists. At present, a significant portion has been digitized through the German Wikisource project. De Graaf et al. (2024) have shown that this resource provides sufficiently broad entity coverage for NEL in the target domain.

At the moment of writing, the complete work is available as scanned images, accompanied by access to a manually verified *Volltext* (full text) for over 65,000 articles. A searchable index or *Register* of keywords (*Stichwörter*), including brief summaries for all entries (*Kurztext*), is also provided to support navigation. Figure 1 shows how these components are presented in the entry for the hero Abas (Abas 3).

In de Graaf (forthcoming) a separate instance of the *RE* was created on the basis of the Wikisource project. This local instance was then further preprocessed by the developers of the Trismegistos (TM) database, an expansive metadata platform fo-

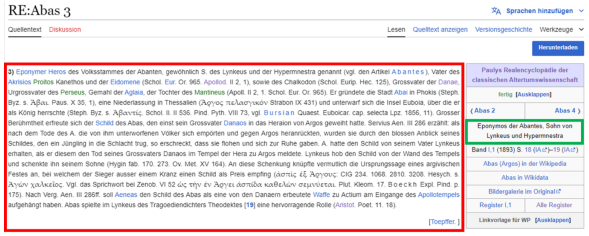


Figure 1: "Abas 3" in the Wikisource edition, highlighting *Volltext* in red and *Kurztext* in green.

cused on the Ancient World (Depauw and Gheldof, 2014). Firstly, all persons were manually identified, with the definition of “person” kept broad (including mythological creatures such as nymphs or muses, and deities). This choice ensures interoperability with NER systems and avoids confusion regarding edge cases such as “Heracles”, a hero with a demigod status. After this filtering, the KB contained a total of 54,180 entries manually identified as persons.

Secondly, for these identified people, the complete name was restored where possible by the TM team and split into its constituent elements. This is particularly important for Latin/Roman names, which often consist of three distinct parts: praenomen, nomen and cognomen. For instance, for the entry *Iulius 131*, the complete name is “C. (Gaius) Iulius Caesar”. Each of these was linked by members of the TM team to, among other things, the corresponding TM Nam ID 9067 (C./Gaius), TM Nam ID 6964 (Iulius) and TM Nam ID 9725 (Caesar) (Broux and Depauw, 2015).

4. Methodology

4.1. Overview

The experimental set-up of this paper is partially based on Beersmans et al. (2025), with two key differences. First, whereas the original pipeline was developed for Ancient Greek, we adapt it to link Latin texts to the *RE*. Second, we introduce an additional Wikipedia-derived pre-training phase that was not part of the original approach. We therefore train BLINK in two stages: we begin with a large, general Wikipedia-derived dataset to improve representations and mitigate data scarcity, and then continue with a smaller, heterogeneous dataset of classical Latin texts linked to the *RE* (the *RE*-dataset).

The *RE*-dataset is divided into "gold" and "silver" subsets. The gold data are manually annotated in the framework of philological projects, whereas the silver data is generated through a rule-based linking procedure that relies on various pre-existing linked identifiers. While the silver data is less time-consuming to create, it is also more error-prone than manual annotation. As such, silver data is used only during training and excluded from testing to ensure the reliability of evaluation results.

To link mentions from Latin texts to the entities of the *RE*, we additionally evaluate two scenarios. In the first, the model is trained to link exclusively to the *Kurztexts* (see Section 3), which are available for all entities (below referred to as *Kurz*). In the second, the model uses the *Volltexts* wherever they exist and falls back on the *Kurztext* otherwise (below referred to as *Voll*).

4.2. Data

This section first introduces the data used for the Wikipedia training phase (Section 4.2.1). It then describes the datasets used in the second, domain-specific phase, namely the silver data (Section 4.2.2) and the gold data (Section 4.2.3).

4.2.1. Wikipedia Data

The data for the Wikipedia-derived training phase were created based on two resources. First, Wikipedia (German and Latin versions) and second, a filtered Wikidata knowledge base, containing 30,000 persons of antiquity, created by Fantoli et al. (2026). To emulate the linguistic situation of the *RE*, only those entities that have a Wikipedia page in German were selected. Our final knowledge base contains 11,652 entities.

Based on scripts provided by De Cao et al. (2022), we derived a dataset of mention-entity pairs where the mention is a hyperlinked phrase in a Latin Wikipedia page and the entity is the corresponding page on German Wikipedia. An example can be found below:

- **Mention**: *Sed anno 1903 Sartus coronatus est Papa Pius X, et haud multum post coronationem, die* [MENTION_START] *Sanctae Caeciliae* [MENTION_END] *(patronae musicae), vulgavit Motu Proprio "Inter Sollicitudines" (cum Perosi conscriptus)* (Latin Wikipedia entry "Laurentius Perosi")
- **Entity**: *Cäcilia von Rom, Cäcilia von Rom wird in mehreren christlichen Konfessionen als Heilige, Jungfrau und Märtyrin verehrt [...])* (German Wikipedia entry "Cäcilia von Rom")

The resulting 29,421 mention-entity pairs were split into train, validation and test partitions. To ensure that certain uncommon entities were not observed during training, we deliberately reserved a small subset of mentions linked to these rare entities exclusively for the validation and test sets. An overview of the final counts per partition can be found in Table 1. The test set is only used for internal validation and results are not reported here.

Subset	Train	Dev	Test
Count	26,665	1,218	1,538
Of which unseen	n/a	122	118

Table 1: Subset counts for the Wikipedia dataset.

4.2.2. Silver Data

The starting point of the silver data pipeline was the description (*Volltext*) of a person entry in the *RE*. These descriptions sometimes cite passages in classical Latin texts, where the *RE* authors derived information about this entity. For instance, for the entity Caecilius 86 (*RE* id: 17476), the *Volltext* references *Hor. carm. II 1*, 1.

To move from the *RE Volltext* to these classical Latin source passages, we made use of various interlinked resources. One such resource is the LASLA corpus, in which all tokens have been manually linked to LiLa Lemma IDs (Fantoli et al., 2022).[2] We also rely on the Trismegistos Nam IDs assigned to each component of the restored full names for all *RE* entries, as described in Section 3. A link between LiLa:lemmas and Trismegistos Nam IDs already exists within the Trismegistos knowledge base, which facilitates this alignment. For more details, consult Beersmans et al. (2025), where a similar pipeline was used, or the provided schematic in Figure 2. Different from the Greek version of this pipeline, we did apply a rule-based procedure to reconstruct multi-token entities, where adjacent tokens assigned the same *RE* identifier are annotated as a single mention.

Since the procedure is fully automated, we evaluated the results using a random sample of 50 instances, of which 39 were correct. We found that one type of error pertained to **plural or generalized name forms**. For example, in the case where *Metellorum*, referring collectively to the Metelli, in *Cic. Planc.* 89 was incorrectly annotated as Caecilius 97 simply because he is mentioned in singular form elsewhere in the same source passage.

To avoid this, we opted to remove all plurals from the final dataset after this analysis was concluded, which was possible thanks to the LASLA morphological annotations. The final silver dataset comprises 620 mention-entity pairs.

4.2.3. Gold Data

The gold data was created in several contexts. In all subsets, person mentions were annotated with their respective corresponding *RE* article. Following Palladino et al. (2024), a person is defined as

[2]LiLa is a knowledge base interlinking linguistic resources, NLP tools, and corpora for Latin (Passarotti et al., 2020)

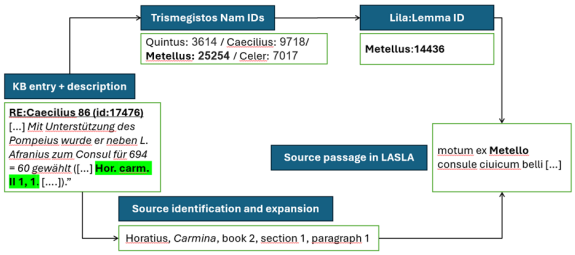


Figure 2: Schematic of the silver data creation process.

"any identifiable single individual, including deities and anthropomorphic mythological figures". The texts are taken from different digital corpora.

Ammianus' *Res Gestae* & Tacitus' *Historiae*. The first subset was created by de Graaf et al. (2024).[3] in the context of evaluating knowledge bases for NEL. It contains Tacitus *Historiae* 1 (annotated based on the LASLA version) and book XIV of the *Res Gestae* of Ammianus Marcellinus from LTA.[4]

Plato case study. The second subset consists of ancient Greek and Latin texts where the figure of Plato plays a relevant role and that are authored in diverse cultural and historical settings (de Graaf, forthcoming). It was developed in the framework of the NIKAW project.[5] The Latin texts of this corpus are annotated based on the linguistically annotated version from the *Corpus Latin Antiquité et Antiquité Tardive lemmatisé* (Clérice, 2020) (henceforth Corpus Latin).[6] The annotated texts used from this resource are:

- *Tusculanae Disputationes* by Cicero,
- *De Anima* by Tertullian,
- *Pro Se De Magia Liber* by Apuleius,
- *Divinarum Institutionum* by Lactantius.

In addition, one text from LASLA is included:

- *Ad Lucilium Epistulae Morales* by Seneca.

For *Divinarum Institutionum* and *Ad Lucilium Epistulae Morales*, only the sections that mention

[3]Data available here: https://github.com/evelien-degraaf/Named-Entity-Linking-Latin-DAAL-2024/tree/main/Gold_data.

[4]https://lta.bbaw.de/text/show/24819722_ammianus_marcellinus_res_gestae.

[5]https://research.kuleuven.be/portal/en/project/3H220323

[6]https://github.com/lascivaroma/latin-lemmatized-texts.

Plato were annotated (respectively chapters and letters).

Pliny NH. The third subset was created in the context of the MECANO doctoral network.[7] It consists of books 2-6 of Pliny the Elder's *Naturalis Historia*, taken from the Corpus Latin and manually corrected from the observed OCR-related errors.

Trismegistos/LASLA gold. A final subset consists of annotations produced by the Trismegistos team during work on LASLA/LiLa. Roughly 5,000 distinct capitalized words in the LASLA corpus were semi-automatically matched against Trismegistos variant datasets for place names (GEOVAR), divine names (GODVAR), and personal names (NAMVAR), yielding about 3,000 matches. The remaining 2,000 items were classified manually.

All 78,919 capitalized words in the corpus had been disambiguated as places (26,199), gods (8,566), or people (44,242). Places and gods were linked to TM identifiers, while only a minority of personal names (3,244) received identifiers. For these identified persons, multi-token entities were reconstructed using the same rule-based procedure described in Section 4.2.2. After removing all mentions overlapping with other annotated corpora, the final set comprised 2,982 mention–entity pairs.

4.2.4. Final *RE* Dataset

For the final dataset, all the gold subsets and the silver data were aggregated. The total counts of the mention-entity pairs for each subset, including silver, are presented in Table 2. The dataset was divided into training, validation, and test partitions. Silver entities were included exclusively in the training set. In addition, a subset of mentions linking to infrequent entities was set aside for the development and test sets to ensure that each contained at least 100 mentions linking to previously unseen entities. An overview is provided in Table 3.

Subset	# Mention-entity pairs
Trismegistos gold	2,982
Plato case study	2,388
Pliny NH	917
Silver data	620
Tacitus	619
Ammianus	202
Total	7,728

Table 2: Overview of different subset sizes

Subset	Train	Dev	Test
Count	6.656	537	535
Of which unseen	n/a	108	102

Table 3: Final train-validation split partitions with the numbers of mention linking to unseen entities.

4.3. Model

BLINK (entity linking with BERT) is a transformer-based entity linking model that makes no assumptions about the nature of the KB, except that a description of each entity should be present. From a technical perspective, the model supports both knowledge-base-specific entity linking and the transfer of entity linking capabilities from one knowledge base (e.g., a large, well-resourced KB) to another that is smaller or more resource-constrained.

BLINK uses a two-stage entity linking approach. In the first stage, both mentions and KB entity descriptions are encoded with two separate transformer models, forming a bi-encoder. The mention embedding is compared with all entity embeddings using dot-product similarity, and the top-k most similar entities (or candidates) are selected. During training, the bi-encoder is optimized to maximize the similarity between the mention and the correct entity compared to the mention and in-batch negatives (incorrect mentions found in the same batch).

In the second stage, each of the top-k entity candidates is concatenated with the mention and jointly encoded using a cross-encoder. A final linear layer assigns a score to each mention–entity pair in order to determine the best match. Although this step is slower and more computationally expensive, it yields higher accuracy because the model can attend the mention and the entity jointly.

While the original BLINK implementation is English-only (although a multilingual implementation of the bi-encoder exists, see Section 2), it can be adapted by replacing the base transformer BERT with a multilingual model. Our new base transformer is XLM-RoBERTa (Conneau et al., 2020), which contains both Latin and German.

4.3.1. Training Parameters

To choose our training hyperparameters, we began by reusing the hyperparameters from the zero-shot setting in Wu et al. (2020), since this setting involved training a model of comparable size to ours. We performed a manual search for the the learning rate and number of epochs for the bi-encoder, monitoring the in-batch accuracy using the validation set. We use a per-GPU batch size of 32, which is a standard and memory-efficient configuration for sequence length 128 and provides 31 in-batch negatives per instance. An overview of the final parameters can be found in Table 4. The hyperparameters for *Voll* and the *Kurz* are kept identical.

[7] https://mecano-dn.eu/

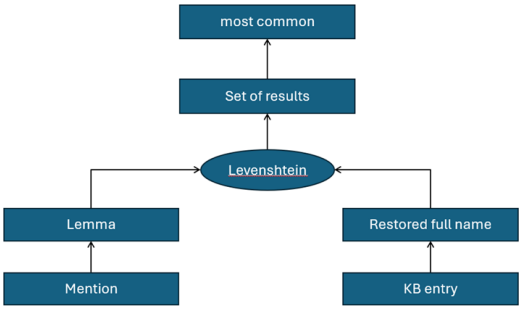


Figure 3: Schematic of the baseline approach

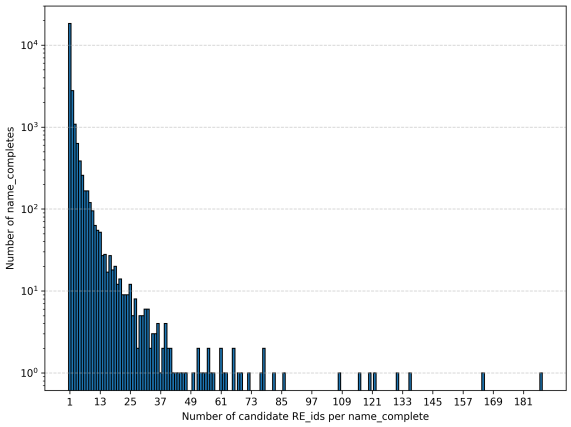


Figure 4: Number of RE_ids per complete name.

4.3.2. Baseline

We compare the performance of BLINK to a two-step approach relying first on string similarity (for the bi-encoder, or "candidate generation" step) and, second, on the frequency of the entity in the training data (for the final ranking step). For clarity, a schematic is provided in Figure 3.

The first part of the baseline implements a lexical candidate selection and identifier expansion procedure using the *RE* KB. Mention data and the *RE* data are first loaded and lowercased. From the KB data, a lookup structure is constructed that maps each available complete name to its corresponding RE_IDs. If a complete name (see Section 3) is present for an entry, it is used as the lexical representation; if not, a cleaned version of the entity's title in the *RE* (e.g. Iulius 131 would be cleaned to "Iulius") serves as a fallback.

For each mention, the baseline pipeline performs exhaustive top-1 Levenshtein matching against all candidate names. Whenever a lemma is available for the mention, the edit-distance computation is performed on the lemmatized form[8]; otherwise, the surface mention is used. For each token in the mention, the minimal Levenshtein distance to any token in the candidate name is computed, and these minimal distances are averaged to yield a similarity score. The candidate name(s) with the lowest average distance are selected, with ties retained if multiple names achieve the same best score.

In a second step, all RE_IDs associated with the selected top-1 name(s) are retrieved and aggregated. A single canonical name may correspond to multiple *RE* entries (as a straightforward example, the canonical name Thespis is associated to *RE* entries Thespis 1 (83642); Thespis 2 (83643) and Thespis 3 (83644)). This expansion can therefore result in one or multiple candidate identifiers (RE_ids) per mention (see Figure 4 for the distribution of candidate entities per name).

[8]As the data are compiled from different sources, the lemmata have differing methods of u/v normalization, similarly to the mentions.

Parameter	Wiki	*RE*
Bi-encoder		
LR	2×10^{-5}	2×10^{-6}
# Epochs	10	5
Batch size	32	32
Cross-encoder		
LR	1×10^{-5}	1×10^{-5}
# Epochs	2	2
Batch size	1	1
Top-k	64	64
Context length	128	128

Table 4: Hyperparameters across training phases.

The entire group of entities retrieved by this setup is compared to the output of the bi-encoder. Although the Levenshtein pipeline is allowed to retrieve an undefined amount of RE_IDs and the bi-encoder retrieves a fixed amount, the comparison provides insights into the effectiveness of the surface matching for the task. The Levenshtein pipeline retrieves 17 RE_IDs on average, with outliers up to 509.

Finally, for the second step of the baseline, we select the most common entity from this set of retrieved entities in the training data. Naturally, this comparison is mostly relevant for the mention-entity pairs where the entity was seen during training. However, we also include in the evaluation of the second step the cases in which the entity is unseen but the Levenshtein pipeline only retrieved one entity, and hence we simply assess whether it is correct or not.

5. Results

5.1. General Overview

The following section details the results of the models after training on the RE. We only report the results on our final KB and not the intermediate results on the Wikipedia test set. The bi-encoder is evaluated on recall@64, whether or not the correct entity is present in the retrieved set of 64 most likely entities. 64 is one of the standard values tested by Wu et al. (2020) and also used in the zero-shot setting. The cross-encoder is then evaluated on its correct predictions (using accuracy) for those instances where the bi-encoder was able to retrieve the correct entity. The overall accuracy, where both bi-encoder retrieval errors and cross-encoder errors are taking into account, is also reported. The column "in training" refers to mention-entity pairs in the test set where entities were seen during training (m=433), and "unseen" refers to mention-entity pairs where the entity has not been seen during training (m=102). We additionally report whether the correct entity was present in the final top 5.

An overview can be found in Table 5. Overall, the Levenshtein baseline is competitive as a candidate generation step compared to the bi-encoder, even outperforming it in the "unseen" setting. Furthermore, the *Voll* outperforms the *Kurz* and both models outperform the most-common baseline in all settings. As expected, performance drops notably when the models are evaluated on unseen entities in both variants. Interestingly, when evaluated on mentions linking to entities unseen in the training data, *Kurz* and *Voll* score comparably. When evaluating whether the correct prediction is among the top 5 entities ranked by the cross-encoder, results improve significantly, up to +20% for unseen entities. We discuss this point in Section 6.

Model	Total	In training	Unseen
Baseline			
Lev	79.44	79.21	80.39
Most common	64.67	75.06	20.58
Kurz			
Bi recall@64	85.42	87.53	76.47
Cross acc.	90.37	94.72	69.23
Total acc.	77.20	82.91	52.94
In top 5	84.30	87.07	72.54
Voll			
Bi recall@64	87.66	91.22	72.55
Cross acc.	93.17	96.96	72.97
Total acc.	81.68	88.45	52.94
In top 5	86.54	90.76	68.63

Table 5: RE dataset results for the total test set, mention linking to entities seen during training and those linking to unseen entities, evaluated on bi-encoder recall@64 (bi recall@64), cross-encoder accuracy (Cross acc.) and total accuracy (Total acc.).

5.2. Error Analysis

Among the 535 entities in the test set, *Kurz* produced 122 incorrect predictions and *Voll* produced 98. Of these errors, 76 were shared: in 31 cases both models made the same incorrect prediction, while in the remaining 45 their incorrect outputs differed. The observations below are based on a closer inspection of 50 incorrect predictions per model.

Firstly, the analysis highlights several aspects of the datasets that could have influenced the models. We observed several issues, including the **inconsistent representation of the phoneme /v/** as <v> or <u> across editions in the test set. For example, both the form *uespasiano* (Tacitus) and *vespasiano* (Pliny NH) occur. Additional errors originate in the training data due to imperfect OCR, such as in Plato case study *omdius* for *ovidius*, *maxumus* for *maximus*, *heropbilus* for *herophilus* and, in NH_Plinius, *vlixis* for *ulixis*.[9]

In addition, we observed **inconsistent annotations across datasets**. These issues occur primarily when multiple entries in the KB refer to the same individual. Examples include separately recorded epithets, aliases, *Verweise* (redirects from one *RE* entry to the other), and Latin and Greek name variants. Such duplications can produce false negatives: *ulixes*, for instance, the Latin name for the hero Odysseus, may be linked to either Ulixes or Odysseus. Epithets for Jupiter illustrate the problem as well: *elicius* in Pliny NH is annotated as Iuppiter, yet *Voll* predicts Elicius, which is counted as an error despite being semantically correct. In the case of *uespasiano*, both models return the *Verweis* in the *RE*, "Vespasianus → Flavius 206" rather than the target entity Flavius 206. Not all predictions are defensible, however: for *elissa*, an alias of queen Dido, *Kurz* predicts Elix 1 and *Voll* predicts Elis 5, both unrelated. Other issues arise when a person is referred to only by alias which is not independently represented in the KB. This is the case for *magne* for Pompeius 31 (Pompeius Magnus), where *Kurz* predicts Magnes 1 and *Voll* predicts Mago 15.

The remaining errors show that both models struggle to disambiguate individuals who share the same name (homonyms). As an example, in the test set, there are 32 instances where a lemma refers to more than one historical individual. These

[9] This represents a case of OCR error missed during the correction process of the Pliny NH data. The error has been corrected in a more recent version of the dataset.

32 occurrences involve six different lemmas.[10] For these lemmas, we evaluated whether the model predictions were correct. Out of the 32 homonymous cases, the Voll model misclassified 18, while the Kurz model misclassified 22. Importantly, this count reflects only homonymy within the test set; other lemmas may be homonymous in the training or development splits.

We can observe three main patterns when models are confronted with homonymity.

1. **Defaulting to one entity**: The lemma *Cinna* refers in the test set to both Helvius 12 (C. Helvius Cinna) and Cornelius 106 (L. Cornelius Cinna): the two models predict Cornelius 106 for all but one instance, where *Kurz* incorrectly selects Helvius 12.

   The case of Caesar is equally illustrative. Correct predictions would include Constantius 5, Iulius 131, Sulpicius 63, and Claudius 256. In seven of nine occurrences of *caesar*, the two models default to Constantius 5; although for *claudius caesar* (correct: Claudius 256), both predict Claudius 92, despite the correct entity being more prevalent in the training data.

2. **Confusion between multiple candidates**: In case of *Claudius*, where three individuals are possible (Claudius 115, 239, and 256). Only for *claudius maximus* (plato_case_study) both models correctly return Claudius 239. Elsewhere, predictions diverge and are uniformly wrong: Claudianus 9, Claudius 328 (*Voll*) or Claudius 109 (*Kurz*) and Claudius 92.

3. **Correct disambiguation (rare)**: *Cato* is the sole example where one model succeeds for multiple individuals: *Voll* correctly links both mentions of Porcius 16 and the single mention of Porcius 9.

Notably, for the examples highlighted in this paragraph, these outcomes cannot be explained simply by frequency in the training data alone, since the models do not consistently choose the most common individual.

The 46 errors made by *Kurz* but not by *Voll* indicate that the *Voll* model handles variation in mentions' surface forms more effectively. These cases suggest that the bi-encoder in *Voll* is less sensitive to variation: *Kurz* correctly predicts Alexandros 10 only for the token *alexandri*, while for the other eight forms (*alexander*, *alexandri*, and *alexandro*) Alexandros 10 does not even appear in *Kurz*'s top-64 candidates. *Voll*'s predictions of these entities, instead, are all correct. A similar improvement appears for the entity Amphitryon, in the form of *amphitruo*, which is correctly predicted by *Voll*. On the other hand, *Kurz* correctly predicts only once the form *amphitruone*, while it never links correctly the form *amphitruo*.

In the 22 errors made by *Voll* but not by *Kurz*, we found no consistent patterns.

6. Discussion

From our results and error analysis, we can investigate whether our trained BLINK is viable for humanities research. While model performance is competitive for entities seen during training, for unseen entities, roughly a fifth of the test set, both *Kurz* and *Voll* were correct in only about half the cases, making them hard to employ without further verification.

The error analysis revealed that KB phenomena such as duplicated entries, redirects, aliases, and parallel Greek/Latin forms, produce systematic false negatives and errors. This indicates that KB selection is crucial for the task: inconsistencies within the KB propagate into both the training and test sets, inevitably affecting model performance and undermining the soundness of the evaluation. The Wikisource *RE* partially suffers, from this standpoint, from its incompleteness with regards to the *Volltext*, due to being a work-in-progress subject to copyright laws, and from the fact that it mirrors the structure of a printed resource without a rigorous data model behind it. However, Wikidata and Wikipedia, being community-based, notably suffer from similar issues (e.g., duplicate entries) and currently offer less extensive coverage (Fantoli et al., 2026). In the future, we plan to compare the results obtained using the *RE* to those with Wikidata.

Moreover, while both *Kurz* and *Voll* models struggle with homonym disambiguation, *Voll* handles variation in surface forms more robustly than *Kurz*, especially for inflected mentions. This may indicate that, when available in the *Volltext*, contextual information is leveraged to disambiguate less straight-forward surface mentions.

We find that string-based top-1 Levenshtein is a competitive method for candidate generation on unseen entities. However, in rare cases it returns an unmanageably large number of candidates (see Section 4.3.2), which limits its direct usability. This can be mitigated by discarding Levenshtein suggestions that correspond to full names mapped to many RE_ids. With such filtering, its outputs can effectively supplement the BLINK bi-encoder's initial selection by adding Levenshtein matches that aren't already in the retrieved set. Finally, the findings are promising for a human-in-the-loop approach. The top 5 suggestions from BLINK have accuracies of 0.9 and 0.7 for seen and unseen entities, respectively. This, together with the first sugges-

[10] Caesar (8), Cato (3), Cinna (6), Claudius (6), Lucilius (2), and Maximus (7)

tion for unseen entities being right in 50% of cases, indicates that it could reduce the verification time.

Secondly, Figure 4 shows that many names correspond to only a small number of RE_IDs. This suggests that a system could reliably propose the correct match either by selecting the RE_ID with a very high surface-matching score when no homonyms are present, or by presenting the few possible candidates. Such an approach could serve as a useful aid for human verification.

7. Conclusions

In this paper, we explored the viability of a transformer-based BLINK model for cross-lingual, domain-specific NEL in a highly specialised humanities context.

The results showed that while BLINK scored competitively for data that was seen during training, generalising to unseen entities remains a challenge. The analysis highlights that issues related to KB structure and annotation choices limit performance. More broadly, the study shows that contextual richness in the KB and data quality and homogeneity are decisive factors in this domain. We also found that both *Kurz* and *Voll* models struggled to disambiguate individuals with the same name. However, the top 5 scores and the performance of the Levenshtein-based baseline, point toward the feasibility of human-in-the-loop workflows, where automated suggestions can reduce verification time without requiring full automation.

Future improvements could include incorporating Levenshtein retrieval. Additionally, we could experiment with adding real-world rule constraints such as the time-based constraints Graciotti et al. (2025a) added, to make sure no temporally unlikely entities are added. Next, KB enrichment could be helpful, for exemple exploiting the already existing interlinks between Wikidata and the RE to resolve ambiguity and add more textual data in the *Kurz* setting. Lastly, we could explore more advanced modeling strategies, such as hard negative mining (Gillick et al., 2019), to improve bi-encoder performance.

8. Funding statement

Part of this work has been carried out under funding from the European Union's Horizon Europe Research and Innovation Programme through the Marie Sklodowska-Curie Actions Doctoral Network MECANO, Grant Agreement No. 101120349. Funded by the European Union. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

9. Bibliographical References

Marijke Beersmans, Evelien De Graaf, Alek Keersmaekers, Mark Depauw, Tim Van De Cruys, and Margherita Fantoli. 2025. Automatic Named Entity Linking for Ancient Greek with a Domain-Specific Knowledge Base. *Anthology of Computers and the Humanities*, 3:540–556.

Emanuela Boros, Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Ahmed Hamdi, José G. Moreno, Nicolas Sidère, and Antoine Doucet. 2020. Robust Named Entity Recognition and Linking on Historical Multilingual Documents. In *Conference and Labs of the Evaluation Forum (CLEF 2020)*, Paper 171. CEUR-WS Working Notes.

Yanne Broux and Mark Depauw. 2015. Developing onomastic gazetteers and prosopographies for the ancient world through named entity recognition and graph visualization: Some examples from trismegistos people. In *Proceedings of the International Conference on Social Informatics (SocInfo)*, volume 8852, pages 304–313, Barcelona. Springer.

Thibault Clérice. 2020. Corpus latin antiquité et antiquité tardive lemmatisé.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2020. Autoregressive Entity Retrieval. Version Number: 3.

Nicola De Cao, Ledell Wu, Kashyap Popat, Mikel Artetxe, Naman Goyal, Mikhail Plekhanov, Luke Zettlemoyer, Nicola Cancedda, Sebastian Riedel, and Fabio Petroni. 2022. Multilingual autoregressive entity linking. *Transactions of the Association for Computational Linguistics*, 10:274–290.

Evelien de Graaf. forthcoming. Challenges in disambiguating ancient individuals with Paulys Realencyclopädie. In *Semantic Annotation for the Ancient World 2025*.

Evelien de Graaf, Mark Depauw, and Margherita Fantoli. 2024. "Nescio Carneades iste qui fuerit": Evaluation of Knowledge Bases for Named Entity Linking for Latin Texts. In *Proceedings of The First Workshop on Data-driven Approaches to Ancient Languages*, pages 1–19. Language & Translation Technology Team (LT3).

Mark Depauw and Tom Gheldof. 2014. Trismegistos: An Interdisciplinary Platform for Ancient World Texts and Related Information. In *Theory and Practice of Digital Libraries – TPDL 2013 Selected Workshops*, Communications in Computer and Information Science, pages 40–52, Cham. Springer International Publishing.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named entity recognition and classification in historical documents: A survey. *ACM Comput. Surv.*, 56(2):1–47.

Maud Ehrmann, Matteo Romanello, Alex Flückiger, and Simon Clematide. 2020. Overview of clef hipe 2020: Named entity recognition and linking on historical newspapers. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 288–310, Cham. Springer International Publishing.

Maud Ehrmann, Matteo Romanello, Sven Najem-Meyer, Antoine Doucet, and Simon Clematide. 2022. Overview of hipe-2022: Named entity recognition and linking in multilingual historical documents. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 423–446, Cham. Springer International Publishing.

Margherita Fantoli, Valeria Irene Boano, Evelien de Graaf, and Camillo Carlo Pellizzari di San Girolamo. 2026. Wikidata as a knowledge base for people of the greco-roman world. *Journal of Open Humanities Data*, 12(1).

Margherita Fantoli, Marco Passarotti, Francesco Mambrini, Giovanni Moretti, and Paolo Ruffolo. 2022. Linking the LASLA Corpus in the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin. In *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference*, pages 26–34. European Language Resources Association.

Daniel Gillick, Sayali Kulkarni, Larry Lansing, Alessandro Presta, Jason Baldridge, Eugene Ie, and Diego Garcia-Olano. 2019. Learning dense representations for entity retrieval. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, page 528–537, Hong Kong, China. Association for Computational Linguistics.

Arianna Graciotti, Nicolas Lazzari, Valentina Presutti, and Rocco Tripodi. 2025a. Musical heritage historical entity linking. *Artificial Intelligence Review*, 58(5):140.

Arianna Graciotti, Leonardo Piano, Nicolas Lazzari, Enrico Daga, Rocco Tripodi, Valentina Presutti, and Livio Pompianu. 2025b. KE-MHISTO: Towards a multilingual historical knowledge extraction benchmark for addressing the long-tail problem. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20316–20339, Vienna, Austria. Association for Computational Linguistics.

Kai Labusch and Clemens Neudecker. 2020. Named entity disambiguation and linking historic newspaper ocr with bert. In *CLEF 2020 Working Notes*, volume 2696 of *CEUR Workshop Proceedings*. CEUR-WS.

Chiara Palladino, Margherita Fantoli, Evelien de Graaf, Monica Berti, Matteo Romanello, Tariq Yousef, Marijke Beersmans, Tom Gheldof, Laura Soffiantini, and Eleonora Litta Modignani Picozzi. 2024. Experience and challenges with named entities – workshop at DHBenelux 2024.

Marco Passarotti, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, and Rachele Sprugnoli. 2020. Interlinking through Lemmas. The Lexical Collection of the LiLa Knowledge Base of Linguistic Resources for Latin. *Studi e Saggi Linguistici*, 58(1).

Mikhail Plekhanov, Nora Kassner, Kashyap Popat, Louis Martin, Simone Merello, Borislav Kozlovskii, Frédéric A. Dreyer, and Nicola Cancedda. 2023. Multilingual End to End Entity Linking.

Cristian Santini, Marieke Van Erp, and Mehwish Alam. 2026. It's All About the Confidence: An Unsupervised Approach for Multilingual Historical Entity Linking using Large Language Models.

Ozge Sevgili, Artem Shelmanov, Mikhail Arkhipov, Alexander Panchenko, and Chris Biemann. 2022. Neural entity linking: A survey of models based on deep learning. *Semantic Web*, 13(3).

Wei Shen, Jianyong Wang, and Jiawei Han. 2015. Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27:443–460.

Georg Wissowa, Wilhelm Kroll, Karl Mittelhaus, Konrat Ziegler, and Hans Gärtner, editors. 1893–1980. *Paulys Realencyclopädie der classischen Altertumswissenschaft*. Metzler, Stuttgart. Multivolume encyclopaedia.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 6397–6407, Online. Association for Computational Linguistics.

Capturing Ancient Chinese Sense Induction with Automatic Pipelines

[1]Guan-Yu Tseng, [2]Chunki Lim, [3]Chih-Han Lin, [4]Tung-Le Pan,
[5]Yu-Chieh Wang, [6]Lang-Ching Yeh, [7]Shu-Kai Hsieh
Graduate Institute of Linguistics, National Taiwan University
{[1]R14142007, [2]R14142001, [3]R13142001, [4]R12142006, [5]R14142006,
[6]R14142008, [7]shukaihsieh}@ntu.edu.tw

Abstract

While the study of diachronic semantic change has advanced alongside recent computational developments, structured lexical resources that reflect semantic evolution remain scarce for many languages, including Ancient Chinese. By systematizing the diachronic transformations within the Chinese Text Project (ctext, a large corpus of Ancient Chinese), we aim to bridge the gap between traditional philological inquiry and contemporary computational linguistics. This study proposes a pipeline that extracts contextualized embeddings from GujiBERT-fan, a language model pre-trained on pre-modern Chinese, and applies dynamic hierarchical clustering to identify distinct senses across historical periods. The pipeline operates at two levels: a global clustering that aggregates data across all periods to capture the full semantic space, and local clustering within each dynasty to reveal period-specific usage patterns. We test the pipeline with a pilot study on the character 手 (*shǒu*, "hand") across eight dynastic periods, covering over 185,000 occurrences. The results show that the pipeline can capture the diachronic shift from concrete to abstract senses, demonstrating its potential as a scalable method for mapping semantic evolution in historical languages.

Keywords: Chinese NLP, Semantic Change, Historical Linguistics

1. Introduction

Lexico-semantic change is fundamental to language evolution. Historically, the diachronic analysis of meaning was a time-consuming and labor-intensive task. Scholarly inquiry lacked a comprehensive overview of the field, as researchers were largely constrained to qualitative case studies involving only one or a small selection of lexemes at a time.

This changed with the introduction of computational methods. Scholars pioneered the application of distributional semantics to large-scale corpora (Hamilton et al., 2016; Schlechtweg et al., 2019), enabling time-aware semantic change studies for languages including English, German, Italian (Basile et al., 2020), and Latin (Perrone et al., 2021). These advancements help researchers track semantic change across thousands of words simultaneously.

However, Ancient Chinese-a language spanning over three millennia with a remarkably rich and continuous literary tradition—has received comparatively little attention in the area. Despite its potential to contribute profound insights into universal patterns of semantic evolution, the language has not yet been subjected to the same level of systematic, large-scale computational scrutiny, unlike Latin, English, or German.

To address this gap, this study proposes a computational framework to automatically map the Diachronic Semantic Landscape of the Chinese lexicon. The primary objective is not merely to rediscover established dictionary definitions, but to visualize the shifting densities and fuzzy boundaries of semantic fields across historical epochs. By leveraging dense vector representations, this research aims to quantify the topography of meaning, revealing how concepts expand, contract, and merge throughout Chinese history.

A key methodological challenge in applying lexical semantic change methods to Ancient Chinese is the severe imbalance between the sparse textual record and the voluminous late imperial corpora. To address this, the proposed pipeline adopts a two-level architecture combining Global and Local clustering: rather than choosing between joint and incremental approaches, each of which faces distinct limitations in this setting, the framework performs independent dynasty-specific clustering while maintaining a cross-period reference space, designed to handle the diachronic data sparsity inherent to historical Chinese corpora.

2. Literature Review

2.1. Lexical Semantic Change and Word Sense Induction

The computational detection of Lexical Semantic Change (LSC) has evolved considerably over the past decade. Early approaches relied on aligning static word embeddings across time pe-

Proceedings of LT4HALA 2026 @ LREC 2026, pages 163–176.
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

riods (Hamilton et al., 2016), while more recent work leverages contextualized representations from pre-trained language models, clustering token-level embeddings to identify distinct word senses. This clustering-based approach is closely related to Word Sense Induction (WSI), which aims to discover the senses of a polysemous word directly from corpus data, without relying on a predefined sense inventory. In the diachronic setting, WSI extends to identifying how a word's set of senses changes over time. The SemEval-2020 Task 1 (Schlechtweg et al., 2020) introduced the first shared task for unsupervised LSC detection, providing manually annotated datasets for English, German, Latin, and Swedish. Many top-performing systems in that evaluation relied on clustering contextualized embeddings —the same general strategy we adopt in this study.

2.2. Joint vs. Incremental Methodologies

Traditional batch or "joint" clustering processes a longitudinal corpus as a single, static entity. While this approach maximizes the data available for each cluster, it faces significant challenges with "contextual saturation" (Periti and Montanelli, 2024), where the overwhelming volume of data from later periods can obscure subtle semantic nuances in earlier, sparser eras.

Conversely, incremental approaches like "What is Done is Done" (WiDiD) treat Semantic Shift Detection (SSD) as an organic, evolutionary process (Periti et al., 2025). By processing data chronologically and using prototypes from preceding periods, incremental models aim to capture "temporal transactions" in meaning. This avoids the risk of anachronistically projecting modern senses onto ancient texts.

2.3. Theoretical Limitations and Structural Recovery

Despite the practical appeal of incremental methods for large-scale corpora, theoretical analysis suggests inherent limitations. Ackerman and Dasgupta provided a formal proof that the incremental setting is "strictly weaker" than the batch model (Ackerman and Dasgupta, 2014). Specifically, certain "nice" cluster structures that are easily detectable when viewing the data in its entirety may be impossible for an incremental method to recover. They suggest, however, that these limitations can be mitigated by allowing for "extra clusters" or hierarchical refinements.

2.4. The Ancient Chinese Context

Computational processing of Ancient Chinese has gained increasing attention in recent years. The EvaHan shared task series, co-located with LT4HALA since 2022, has organized evaluations on word segmentation, POS tagging, and other fundamental NLP tasks for Ancient Chinese (Li et al., 2022). Pre-trained language models tailored to Classical Chinese, such as GujiBERT-fan (Wang et al., 2023), have also been developed. However, computational research on the diachronic semantics of Ancient Chinese —particularly sense induction over time —remains largely unexplored.

Applying LSC methods to this language presents a specific challenge. The literary record of the Pre-Qin period (before 221 BCE) is extremely limited compared to later dynasties. Under an incremental framework, initial cluster prototypes would be derived from this sparse early data, and any inaccuracy would propagate forward through all subsequent periods. Joint clustering avoids this error propagation by pooling all data together, but at the cost of allowing the massive late-imperial corpora (Ming and Qing) to dominate the semantic space, potentially flattening the very evolutionary trajectory the analysis seeks to reveal. Our study takes a different path. Rather than choosing between joint and incremental clustering, we perform independent clustering within each dynasty (Local Clustering) while also providing a cross-period aggregation (Global Clustering). Each dynasty's clustering is self-contained and not dependent on earlier prototypes, avoiding the error propagation of incremental methods. Meanwhile, the global view preserves rare or emergent senses that might be lost in any single period. Combined with a dynamic hierarchical sub-clustering step, this approach aims to trace the semantic evolution of the target lexeme at multiple levels of granularity.

3. Data and Methodology

The historical data are retrieved from *Chinese Text Project* (Sturgeon, 2011), an extensive corpus of Classical Chinese covering literary, philosophical, medical, poetic, and historiographic registers. It provides broad temporal and domain coverage, with bibliographic metadata supporting reliable periodization.

Our proposed methodology consists of a six-stage pipeline: **(1) Contextual Embedding, (2) OCR Cleaning, (3) Stratified Sampling, (4) Over-clustering and Merging, (5) Hierarchical Sub-clustering, (6) Human-in-the-Loop Semantic Annotation**.

To evaluate the feasibility of the proposed pipeline, we conducted a focused pilot analysis on the lexical item **手** (*shǒu*, "hand"), selected for its high token frequency, stable orthography across periods, and rich semantic diversity from physical to metaphorical uses.

3.1. Data Processing and Contextual Embeddings

We extracted all occurrences of *shǒu* from all the accessible texts in the *Chinese Text Project* database. The data was split into eight periods by major historical event such as dynasty changes: **先秦** (*xiān qín* 'pre-Qin', before 221 BCE), **秦漢** (*qín hàn* 'Qin-Han', 221 BCE–220 CE), **魏晉** (*wèi jìn* 'Wei-Jin', 221–509 CE), **隋唐** (*suí táng* 'Sui-Tang', 589–960 CE), **宋元** (*sòng yuán* 'Song Yuan', 960–1368 CE) **明** (*míng* 'Ming', 1368–1644 CE), **清** (*qīng* 'Qing', 1644–1912 CE) and **民國** (*mín guó* 'Republican', after 1912 CE). This sampling strategy covers all major dynastic periods and reflects a range of genres in each dynastic period for comprehensive, diachronic analysis.

After retrieving the raw data from the *Chinese Text Project*, we segmented the texts into individual sentences using standard terminal punctuation marks (i.e., "!", "?", "。", ";").

To capture the specific historical context, each complete sentence containing the target lexeme was processed through GujiBERT-fan (Wang et al., 2023), a transformer modeled after BERT (Devlin et al., 2018) and pretrained on the **四庫全書** *Sì kù quán shū*, an extensive collection of ancient Chinese text. From the resulting sequence outputs, we specifically isolated the contextualized vector representation corresponding to the character **手** (*shǒu*). These target word vectors were then stored with metadata (e.g., dynasty, work title) and original source sentences for subsequent clustering analysis.

Data Cleaning: To ensure data quality, instances containing obvious OCR (Optical Character Recognition) errors, such as abnormally long sentences, excessively repeated characters, and unidentifiable lexical items, were filtered out to yield a cleaner dataset.

Stratified Sampling: A stratified sampling method is implemented to prevent any single text or genre from dominating the semantic distribution. Within each historical period, the extraction was capped at a maximum of 200 instances of **手** (*shǒu*) per individual document. This threshold effectively mitigates the domination of voluminous historical records, ensuring that the resulting semantic landscape reflects a diverse and representative cross-section of language use for each era.

3.2. Clustering and Annotation

Over-clustering and Merging: Within each dynasty, the contextual embeddings of *shǒu* were grouped using **k-means** clustering to identify distinct senses. To capture fine-grained nuances, based on preliminary experiments, we adopted an over-clustering strategy by setting the initial number of clusters (K) to 20. Subsequently, an automated merging step was applied: clusters with a centroid cosine similarity exceeding the threshold of 0.95 were merged to become the same group. Each finalized cluster approximates a distinct contextualized usage pattern or lexical sense.

The choice of $K = 20$ was validated through a one-factor-at-a-time sensitivity analysis, in which we tested $K \in \{10, 15, 20, 25, 30\}$, *merge_ratio* $\in \{0.90, 0.92, 0.95, 0.97\}$, and *split_ratio* $\in \{0.1, 0.3, 0.5\}$ for both global and local clustering, while holding the remaining parameters fixed at their baseline values. Although $K = 25$ yielded a marginally higher global Silhouette Score (0.054 vs. 0.049, $\Delta = +0.005$), its final cluster count after merging (16) was nearly identical to that of $K = 20$ (17), indicating that the additional initial clusters were redundant. $K = 20$ was therefore selected as the setting that achieves comparable clustering quality while avoiding unnecessary over-fragmentation.

Table 1: Hyperparameter configurations for Global and Local clustering.

Parameter	Global	Local
K (initial clusters)	20	20
merge_ratio	0.95	0.95
split_ratio	0.1	0.3
sub-K	6 (fixed)	dynamic

Hierarchical Sub-clustering: Due to the semantic continuity of natural language, the merging phase often produces overly large, coarse-grained clusters. To extract the nuanced senses of *shǒu*, we incorporated a dynamic hierarchical sub-clustering mechanism. This process is governed by two parameters[1]: a *split-ratio* acts as the threshold to trigger further division (e.g., a ratio of 0.3 means any cluster containing over 30% of the

[1]For example, if the Ming dynasty corpus contains 10,000 occurrences of **手**, setting the *split-ratio* to 0.3 means any merged cluster exceeding 3,000 instances will trigger the sub-clustering pipeline. Once triggered, the algorithm evaluates the cluster's exact size to assign a proportional *sub-*K. A cluster of 4,000 instances might be automatically partitioned into 3 sub-groups, whereas a massive cluster of 20,000 instances in another dynasty would be assigned a higher *sub-*K (e.g., 6). This dynamic allocation prevents both under-clustering and over-fragmentation.

dynasty's total data will be sub-clustered). Rather than using a fixed number, the value of *sub-*K is dynamically determined based on the absolute data volume of the target cluster, ensuring that the semantic granularity scales proportionally with the size of the data.

Human-in-the-Loop Semantic Annotation — Seven Major Categories for shǒu: While the automated pipeline efficiently discovers latent semantic structures, assigning meaningful linguistic interpretations to these mathematical clusters requires domain expertise. We adopted a human-in-the-loop approach to annotate the final output. To enhance efficiency, two large language models, Google Gemini Pro and Anthropic Claude Sonnet (both accessed in February 2026), were employed as independent pre-annotators. By examining the representative "core sentences" located closest to the centroid of each cluster, two LLMs mapped the fine-grained clusters into seven macro-categories. Discrepancies between the two models were flagged for human review, with human experts in Chinese linguistics making the final determination.

This coding scheme reflects the diachronic semantic extension of *shǒu*: (1) **Body_Medical** (the literal physical organ, limbs, and traditional medical concepts like meridians and symptoms); (2) **Physical_Action** (concrete body movements, holding, spatial relations, and tool usage); (3) **Social_Interaction** (interpersonal contact, social signaling, and metaphorical relationships); (4) **Text_Culture** (writing, calligraphy, official edicts, and divination); (5) **Power_Skill** (abstract agency, control, subordinates, and titles for experts/masters); (6) **Grammar_Suffix** (lexicalized occupational suffixes); and (7) **NOISE** (OCR errors and unidentifiable strings). This human-validated categorization ensures that our data-driven clusters are robustly grounded in usage-based linguistics. The complete hierarchical coding scheme, detailing the macro-categories, subcategories, and their corresponding core examples, is presented in Table 2.

4. Results and Analysis

In this section, we present the clustering results of the contextual embeddings for **手** (*shǒu*). We first present a unified diachronic projection to present the overall clustering (Global Clustering, Section 4.1), followed by an examination of the dynasty-specific semantic landscapes to capture the semantic distributions within each specific era (Local Clustering, Section 4.2).

All visualizations in this study follow the same encoding scheme. In the subsequent scatter plots, each data point represents a single occurrence of the character **手**, encoded as a 768-dimensional GujiBERT-fan contextual embedding. While the clustering operates in this high-dimensional space, the coordinates are projected onto a 2D plane via Principal Component Analysis (PCA) strictly for visualization purposes.

Data points are color-coded to represent the six overarching macro-categories (along with a grey NOISE category) defined in our semantic annotation scheme. Text boxes indicate the exact centroids of the dynamically generated, fine-grained clusters. To contextualize the semantic shifts quantitatively and mitigate biases from absolute data volume, each visualization panel reports three key metrics: the absolute sample size (n) in the title, the relative proportion of each macro-category in the top-right corner, and the Silhouette Score (Rousseeuw, 1987) in the bottom-left/right to validate the structural cohesion of the clustering.

4.1. Analysis of Global Clustering

The global clustering analysis provides a diachronic overview of the semantic landscape of **手** (*shǒu*) by aggregating data from all eight historical periods. This unified projection (Figure 1) shows the primary senses that persist across all periods. The macro-categories identified—including **Body_Medical**, **Physical_Action**, **Power_Skill**, **Social_Interaction**, **Text_Culture**, and **Grammar_Suffix**—form the foundational semantic infrastructure of the character. For instance, the clustering isolates core anatomical references and medical concepts within the **Body_Medical** sector, while simultaneously capturing abstract extensions in the **Power_Skill** and **Social_Interaction** domains.

However, because the global view is an aggregation of 185,792 occurrences, it may be susceptible to corpus imbalance, particularly the dominance of late imperial texts (Ming and Qing). To discern the "mainstream" semantic senses of **手** within a specific time frame and to avoid being overshadowed by high-volume eras, it is necessary to refer to the dynasty-specific distributions (Local Clustering) to examine how these semantic weights shifted over time.

4.2. Analysis of Local Clustering

By examining the dynasty-specific landscapes (Figure 2), we can observe the fine-grained evolution of the character's semantic priorities. Each panel reveals the dominant usage types of **手** unique to its respective era. In the early stages (Pre-Qin), the semantic gravity is heavily concentrated in the concrete domains of **Body_Medical** and **Physical_Action**, reflecting a period where **手**

Table 2: The Hierarchical Semantic Coding Scheme for the Character **手** (*shǒu*)

Macro Category	**Sub-categories (Senses)**	**Typical Examples (Hanzi + *Pinyin* + Translation)**
Body_Medical	Literal_Hand, Limbs, Anatomy	右手 *yòu shǒu* 'right hand'; 手足 *shǒu zú* 'hands and feet'
	Symptom (Medical state)	手足厥冷 *shǒu zú jué lěng* 'coldness in hands and feet'
	Meridian (TCM concept)	手太陰 *shǒu tài yīn* 'Hand-Taiyin (Lung Meridian)'
Physical_Action	Holding, Combat_Hold	手執 *shǒu zhí* 'to hold in hand'
	Spatial, State	手中 *shǒu zhōng* 'in the hands'
	Gesture, Take_Action, Instrumental	下手 *xià shǒu* 'to take action/strike'; 以手 *yǐ shǒu* 'using hands'
Social_Interaction	Connection, Contact_Sign	攜手 *xié shǒu* 'hand in hand'; 拍手 *pāi shǒu* 'to clap hands'
	Metaphor (Relationships)	猶人手足 *yóu rén shǒu zú* 'like hands and feet (metaphor for brotherhood)'
Text_Culture	Letter_Edict, Writing, Artifact	手詔 *shǒu zhào* 'imperial edict written by the emperor'
	Divination, Etymology	艮手 *gèn shǒu* 'Gen represents the hand (divination)'
	Literary (Poetic expressions)	飛來我手 *fēi lái wǒ shǒu* 'flies into my hand'
Power_Skill	Agency, Control, Possession	出其手 *chū qí shǒu* 'directed by someone'; 權在手 *quán zài shǒu* 'power in hand'
	Subordinate	手下 *shǒu xià* 'subordinates'
	Expertise, Capability	妙手 *miào shǒu* 'highly skilled person / master'
Grammar_Suffix	Occupation_Suffix	水手 *shuǐ shǒu* 'sailor'; 弓箭手 *gōng jiàn shǒu* 'bowman'
NOISE	OCR_Error	差手靈蓋 *chā shǒu líng gài* (Meaningless OCR noise)

was primarily utilized for literal bodily functions and direct physical movements.

Significant diachronic variations emerge during the medieval and late imperial periods. From the Sui–Tang through the Song–Yuan dynasties, we observe a "semantic migration" toward the upper-left quadrants of the vector space. During these eras, the proportions of **Social_Interaction** and **Power_Skill** clusters expand significantly, indicating that the character's role in social connections and expressions of agency was becoming mainstream. By the Ming and Qing periods, the emergence of the **Grammar_Suffix** cluster signifies the final stage of this evolution, where the character became structurally stabilized as an occupational marker. These local shifts highlight that while the global senses remain relatively consistent, the "mainstream" usage of the word has undergone a transformation from the physical to the abstract.

5. Discussions

The Global and Local clustering approaches serve different analytical purposes. The Global Cluster concatenates data across all millennia, providing a macro-historical landscape of the target lexeme. The global space helps preserve rare senses that would be lost in dynasty-specific clustering. For instance, an abstract sense of *shǒu* (e.g., representing power or agency) might be exceedingly rare in the Pre-Qin period. In a dynasty-specific Local Cluster, such sparse occurrences would likely be absorbed into dominant physical action clusters or entirely discarded as statistical noise. However, in the Global space, the explosion of these abstract usages in later eras (e.g., Ming and Qing) constructs a massive, well-defined semantic territory. The sparse Pre-Qin instances can thus find their "home" within this global distribution, allow-

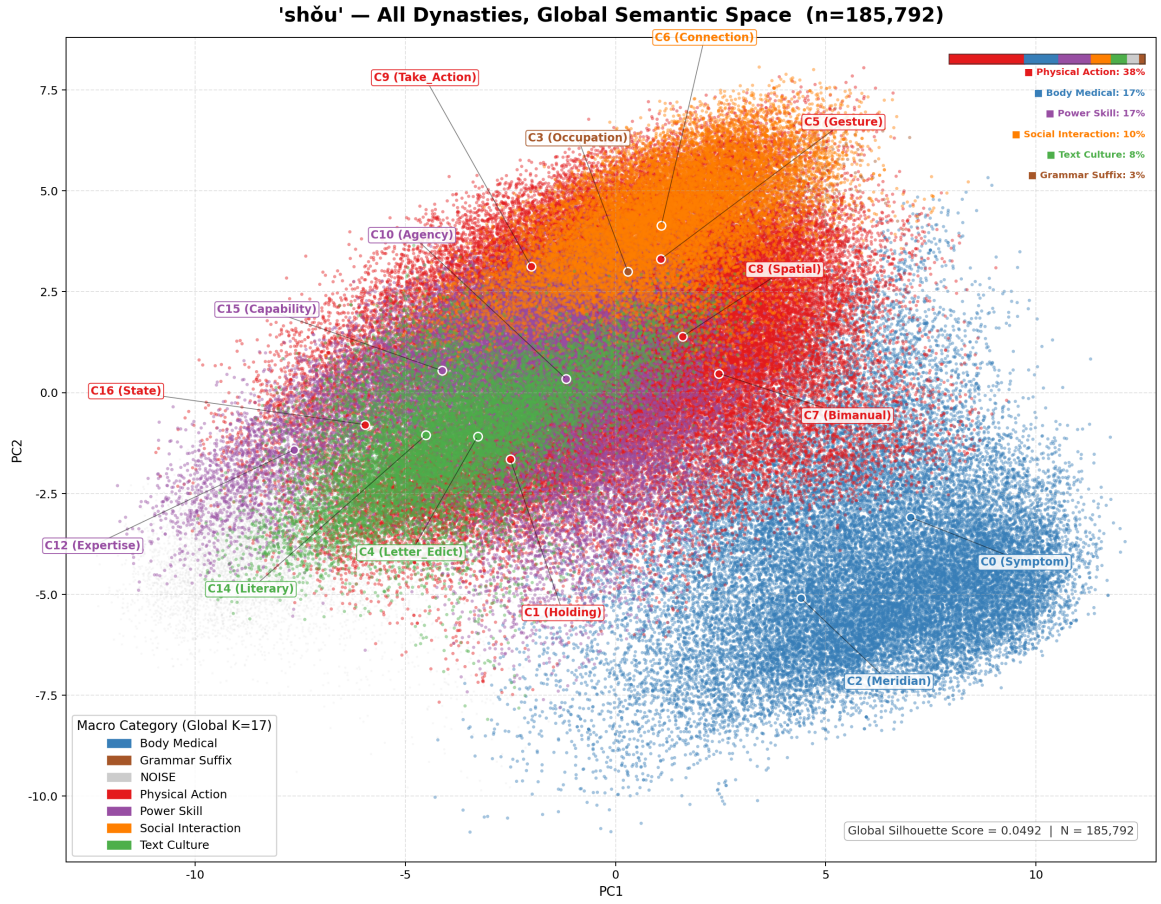


Figure 1: **Unified diachronic semantic space of 手 ($n = 185,792$)**
This plot maps all historical periods onto an absolute PCA coordinate system. Initialized with $K = 20$ and parameterized with a *merge-ratio* of 0.95 and *split-ratio* of 0.1 (*sub-K*=6), the pipeline dynamically yielded 17 semantic clusters (C0–C16).

ing us to trace the earliest seeds of semantic shift. Conversely, Local Clusters are indispensable for providing time-specific snapshots, demonstrating the mainstream semantic priorities and contextual boundaries unique to each individual dynasty.

This approach differs from traditional sense inventory in that it represents meaning as a continuous, multi-dimensional space rather than a discrete list. The character's meaning is not a collection of static definitions, but a dynamic integration of semantic fields such as social and medical fields, grammatical role such as occupational suffix, and prototypical senses.

The clustering output identifies overarching semantic fields (e.g., the medical and social domains), clear syntactic transformations (e.g., the emergence of grammatical suffixes), sense prototypes where usage is highly conventionalized. By moving away from the rigid constraints of a 1D sense list, our approach respects the "semantic continuum" inherent in natural language, where boundaries between capability, agency, and physical action are often fluid and overlapping.

6. Conclusion

In this work, we present a computational framework to automatically map the Diachronic Semantic Landscape of the Chinese lexicon. By integrating contextualized embeddings from a specialized Classical Chinese language model (Wang et al., 2023) with a dynamic, unsupervised clustering pipeline, we demonstrate a methodology that can reveal semantic structure which static dictionary entries do not capture.

The implications of this mapping extend well beyond computational linguistics. The framework can help digital humanities scholars to trace how the senses of key terms shift over time, complementing traditional philological method.

The methodology established in this study offers a scalable framework that could potentially be extended to other languages with extensive diachronic corpora, such as Sanskrit or Ancient

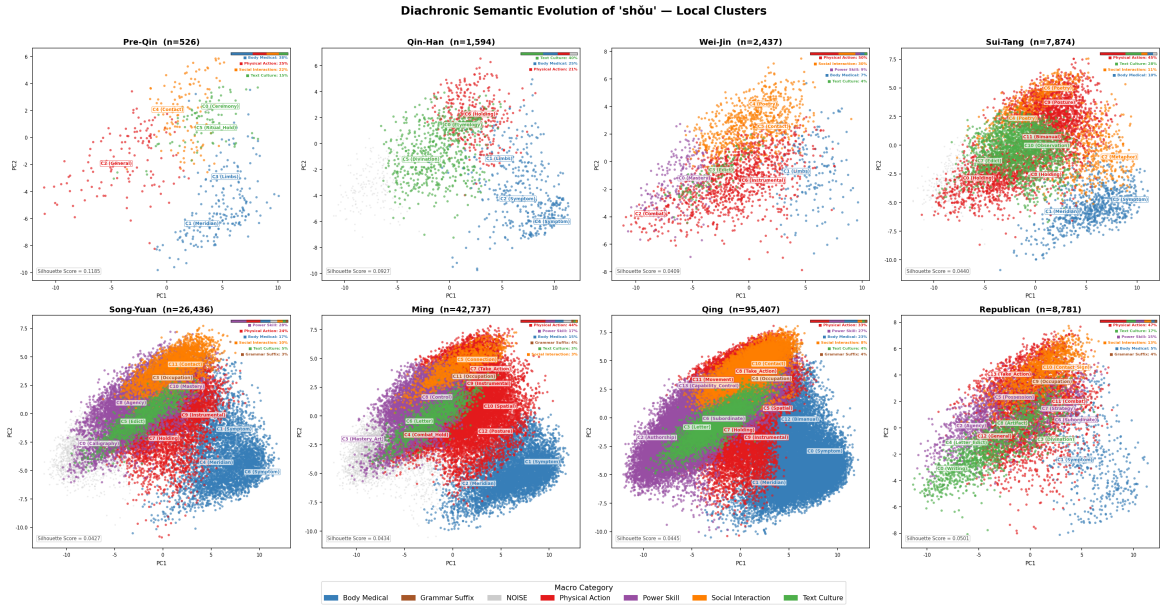


Figure 2: **Dynasty-specific (Local) semantic landscapes of 手**

To capture the semantic distributions within each specific era, the clustering pipeline was initialized with $K = 20$ for every period. We applied a *merge-ratio* of 0.95 and a *split-ratio* of 0.3, with the *sub-*K dynamically determined by the absolute sample size (n) of the respective era. Visual encodings (colors, centroids, and metrics) follow the principles detailed in Section 4.

Greek. By demonstrating the feasibility of unsupervised semantic mapping in a historical context, this research contributes to the ongoing development of dynamic historical lexicons. Ultimately, we hope this pipeline offers a practical, reproducible method for studying lexical change in Chinese and other historical languages.

Furthermore, as the current framework was evaluated through a pilot study of a single high-frequency lexeme (*shǒu*), future work will expand the dynamic clustering analysis to a broader set of target characters. As demonstrated in the generalizability analysis presented in Appendix A, the pipeline produces coherent semantic clusters for two additional lexemes — 水 (*shuǐ*, "water") and 道 (*dào*, "way/road/principle") — suggesting that the approach is not limited to concrete body-part nouns. Extending the analysis to a wider range of lexical types, including abstract concepts and function words, remains an important direction for future work.

7. Limitations

While the integration of GujiBERT-fan and dynamic hierarchical clustering provides a macro-level perspective on the semantic evolution of 手 (*shǒu*), two primary methodological limitations should be acknowledged.

First, regarding the granularity of semantic representation, contextualized embeddings are highly sensitive to broader usage-type distributions and semantic fields (e.g., medical contexts or grammatical functions) rather than fine-grained, dictionary-level senses (Periti and Tahmasebi, 2024). Furthermore, applying a hard-clustering algorithm (K-means) inherently imposes discrete mathematical boundaries upon lexical semantics, which naturally operates on a continuous spectrum. While the visual overlaps in our PCA projections help illustrate transitional "bridging contexts," the generated clusters should be interpreted as semantic prototypes rather than mutually exclusive lexicographic categories.

Second, a persistent challenge in historical corpus linguistics is diachronic data sparsity, particularly the severe corpus imbalance between early eras (e.g., Pre-Qin) and late imperial periods. This sparsity reflects two compounding factors: the historically limited survival of early texts, and the greater susceptibility of archaic orthography to OCR errors, which results in higher data attrition during the cleaning stage. Although our methodology employs stratified sampling to mitigate volume imbalances, it cannot fully compensate for these constraints. Consequently, while the observed spatial expansion from concrete to abstract domains provides strong correlational evidence for metaphorical extension, establishing direct diachronic causality requires future research to supplement this distant-reading approach with qualitative close-reading of historical texts, alongside cor-

pora with denser coverage of underrepresented dynastic periods and improved OCR pipelines.

8. Bibliographical References

Margareta Ackerman and Sanjoy Dasgupta. 2014. Incremental clustering: The case for extra clusters. *arXiv preprint arXiv:1406.6398*.

Pierpaolo Basile, Annalina Caputo, Tommaso Caselli, Pierluigi Cassotti, and Rossella Varvara. 2020. *DIACR-Ita @ EVALITA2020: Overview of the EVALITA2020 Diachronic Lexical Semantics (DIACR-Ita) Task*, pages 411–419.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Bin Li, Yiguo Yuan, Jingya Lu, Minxuan Feng, Chao Xu, Weiguang Qu, and Dongbo Wang. 2022. The first international Ancient Chinese word segmentation and POS tagging bakeoff: Overview of the EvaHan 2022 evaluation campaign. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 135–140, Marseille, France. European Language Resources Association.

Francesco Periti and Stefano Montanelli. 2024. Lexical semantic change through large language models: a survey. *ACM Computing Surveys*, 56(11):1–36.

Francesco Periti, Sergio Picascia, Stefano Montanelli, Alfio Ferrara, and Nina Tahmasebi. 2025. Studying word meaning evolution through incremental semantic shift detection. *Language Resources and Evaluation*, 59:1363–1399.

Francesco Periti and Nina Tahmasebi. 2024. A systematic comparison of contextualized word embeddings for lexical semantic change. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4262–4282, Mexico City, Mexico. Association for Computational Linguistics.

Valerio Perrone, Simon Hengchen, Marco Palma, Alessandro Vatri, Jim Q. Smith, and Barbara McGillivray. 2021. Lexical semantic change for ancient greek and latin. *CoRR*, abs/2101.09069.

Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Dominik Schlechtweg, Anna Hätty, Marco Del Tredici, and Sabine Schulte im Walde. 2019. A wind of change: Detecting and evaluating lexical semantic change across times and domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 732–746, Florence, Italy. Association for Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

9. Language Resource References

Donald Sturgeon. 2011. Chinese text project: a dynamic digital library of premodern chinese. `https://ctext.org/`.

Dongbo Wang, Chang Liu, Zhixiao Zhao, Si Shen, Liu Liu, Bin Li, Haotian Hu, Mengcheng Wu, Litao Lin, Xue Zhao, and Xiyu Wang. 2023. Gujibert and gujigpt: Construction of intelligent information processing foundation language models for ancient texts.

A. Appendix: Generalizability Analysis on 水 (shuǐ) and 道 (dào)

To evaluate the generalizability of the proposed pipeline beyond the pilot lexeme 手 (shǒu), we applied the same methodology to two additional characters: 水 (shuǐ, "water") and 道 (dào, "way/road/principle"). These characters were selected to represent distinct semantic profiles — 水 as a high-frequency concrete noun with rich metaphorical extensions, and 道 as a polysemous term spanning concrete, philosophical, administrative, and grammaticalized senses.

A.1. 水 (shuǐ, 'water')

The global clustering (Figure 3) analysis of 水 ($n = 313,184$) yielded 12 semantic clusters organized into eight macro-categories: Geography, Physical Usage, Literary, Geomancy, Cosmology, Body Medical, Military, and NOISE. Geography (24%) and Physical Usage (21%) dominate the semantic space, reflecting the character's stable concrete core across all dynasties. The Literary category (19%) captures the extensive use of 水 as a poetic image, while Geomancy (12%) reflects its centrality in fengshui traditions.

The dynasty-specific local clustering (Figure 4) reveals a clear diachronic trajectory. In the Pre-Qin period ($n = 1,722$), the semantic space is dominated by Geography and Cosmology, consistent with the cosmological significance of water in early Chinese thought. From the Song-Yuan period onward, the Literary category expands substantially, reflecting the flourishing of water imagery in classical poetry.

A.2. 道 (dào, 'way/road/principle')

The global clustering (Figure 5) of 道 ($n = 331,910$) yielded 22 semantic clusters across nine macro-categories: Philosophy, Speech Act, Geography, Administration, Religion, Onomastics, Cognition, Medical, and NOISE. Philosophy dominates at 41%, reflecting the centrality of 道 as a philosophical concept. Most notably, Speech Act (15%) forms a visually distinct cluster in the embedding space, spatially separated from the philosophical senses, providing strong evidence for a grammaticalization pathway from lexical verb to quotative marker.

The local clustering (Figure 6) results reveal a particularly rich diachronic trajectory. In the Pre-Qin period ($n = 3,226$), the semantic space is almost entirely dominated by Philosophy, with no Speech Act clusters present. The Speech Act category becomes prominent from the Song-Yuan period and expands significantly through the Ming and Qing dynasties. Notably, the Republican period achieves the highest Silhouette Score across all dynasties (0.1718), suggesting that the semantic categories of 道 have reached their highest degree of differentiation in the modern era.

A.3. Cross-lexeme Comparison

Taken together, the results for 水 and 道 demonstrate that the pipeline generalizes effectively across lexemes with distinct semantic profiles. While 手 exhibits a trajectory from concrete bodily meaning toward abstract and grammaticalized senses, 水 shows stable concrete dominance with expanding literary and cosmological extensions, and 道 displays the most complex trajectory, spanning philosophical elaboration, religious extension, administrative specialization, and grammaticalization into a quotative marker. These contrasting patterns confirm that the pipeline is sensitive to the unique semantic histories of individual lexemes while maintaining a consistent methodological framework across all three cases.

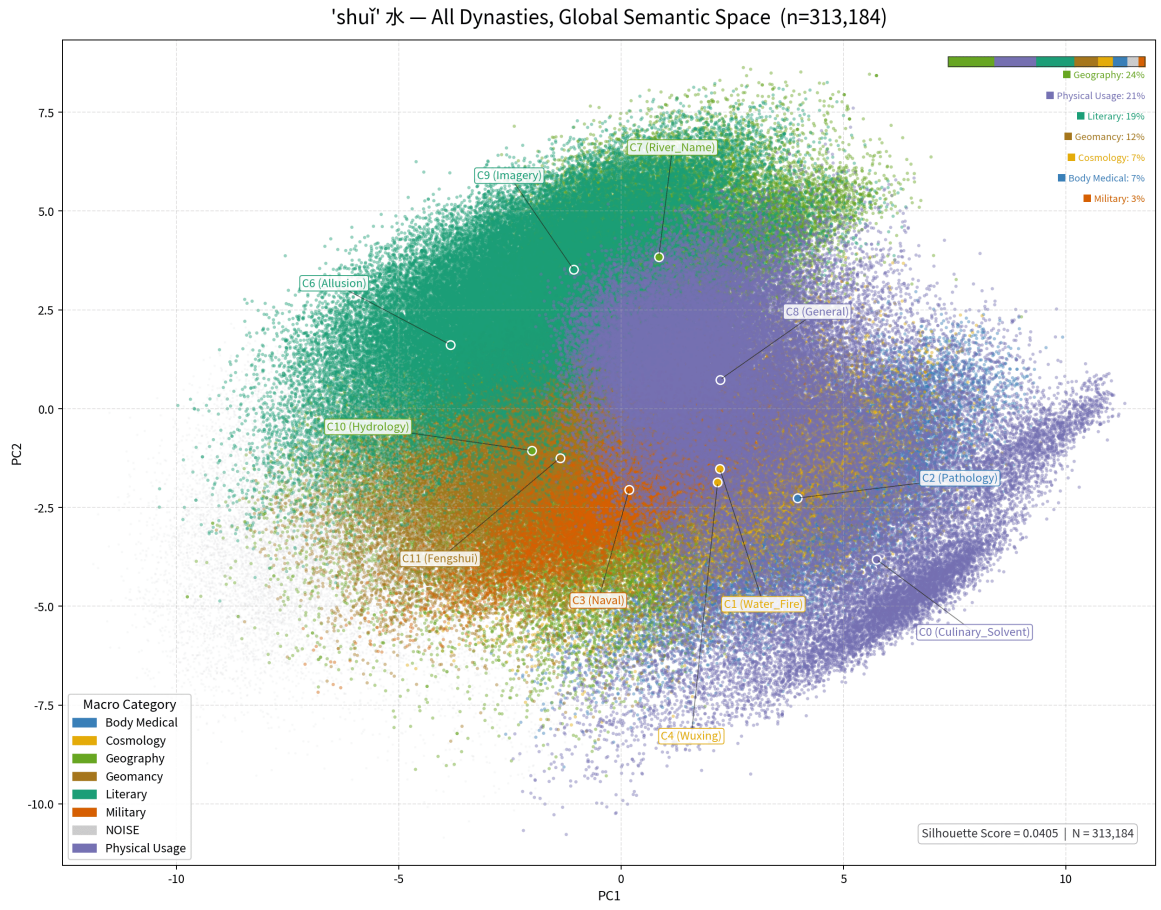


Figure 3: **Unified diachronic semantic space of 水 ($n = 313,184$).** This plot maps all historical periods onto an absolute PCA coordinate system. Initialized with $K = 20$ and parameterized with a *merge-ratio* of 0.95 and *split-ratio* of 0.1 (*sub-K*=6), the pipeline dynamically yielded 12 semantic clusters (C0–C11).

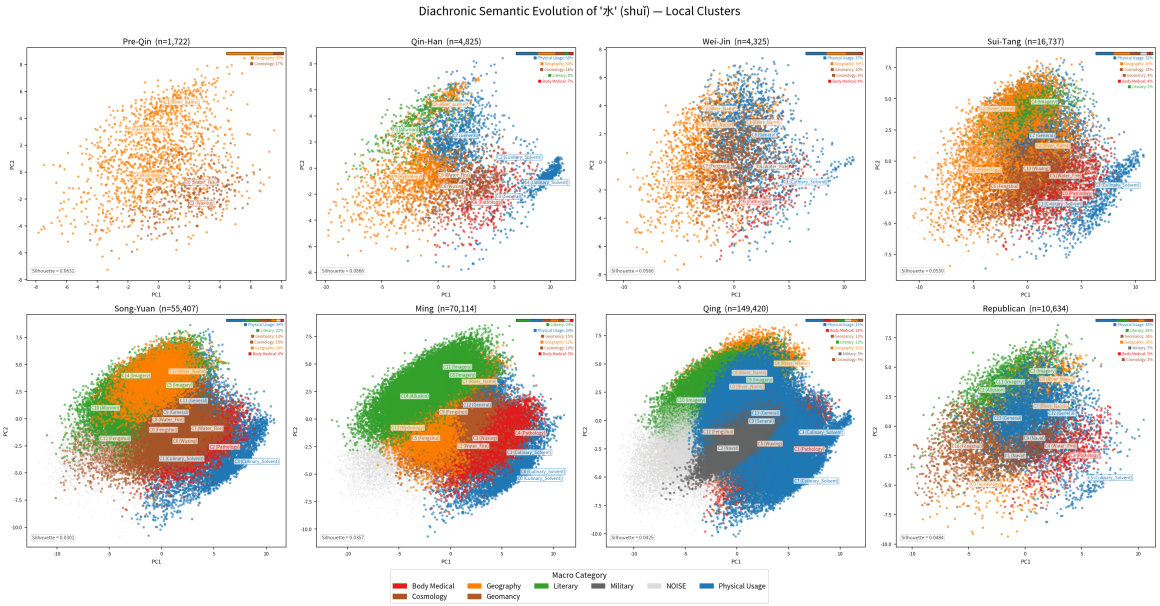


Figure 4: **Dynasty-specific (Local) semantic landscapes of 水**. To capture the semantic distributions within each specific era, the clustering pipeline was initialized with $K = 20$ for every period. We applied a *merge-ratio* of 0.95 and a *split-ratio* of 0.3, with the *sub-*K dynamically determined by the absolute sample size (n) of the respective era.

Table 3: The Hierarchical Semantic Coding Scheme for the Character **水** (*shuǐ*)

Macro Category	**Sub-categories (Senses)**	**Typical Examples (Hanzi + *Pinyin* + Translation)**
Geography	River Name	濼水 *Luòshuǐ* 'Luo River'; 濮水 *Púshuǐ* 'Pu River'
	Hydrology	措置水之術 *cuòzhì shuǐ zhī shù* 'techniques for water management'
Physical_Usage	Culinary Solvent	水煎熟 *shuǐ jiān shú* 'boiled in water'
	General	清水 *qīng shuǐ* 'clear water'
Cosmology	Wuxing (Five Elements)	水乘火 *shuǐ chéng huǒ* 'water overcomes fire'
	Water-Fire Pair	水火之數 *shuǐ huǒ zhī shù* 'the numerology of water and fire'
Body_Medical	Pathology (TCM)	利水 *lì shuǐ* 'to promote diuresis (TCM)'
Military	Naval	水師提督 *shuǐ shī tídū* 'naval commander'
Literary	Allusion	易水 *Yìshuǐ* 'Yi River (allusion to Jing Ke)'
	Imagery	遠水縈紆而來 *yuǎn shuǐ yíng yū* 'distant water winding toward us'
Geomancy	Fengshui	水口關欄 *shuǐkǒu guānlán* 'water mouth closure (feng shui)'
NOISE	OCR Error	近蒙專使至虔遠致時服寢衣之魏尋附居布謝必連 (Meaningless OCR noise)

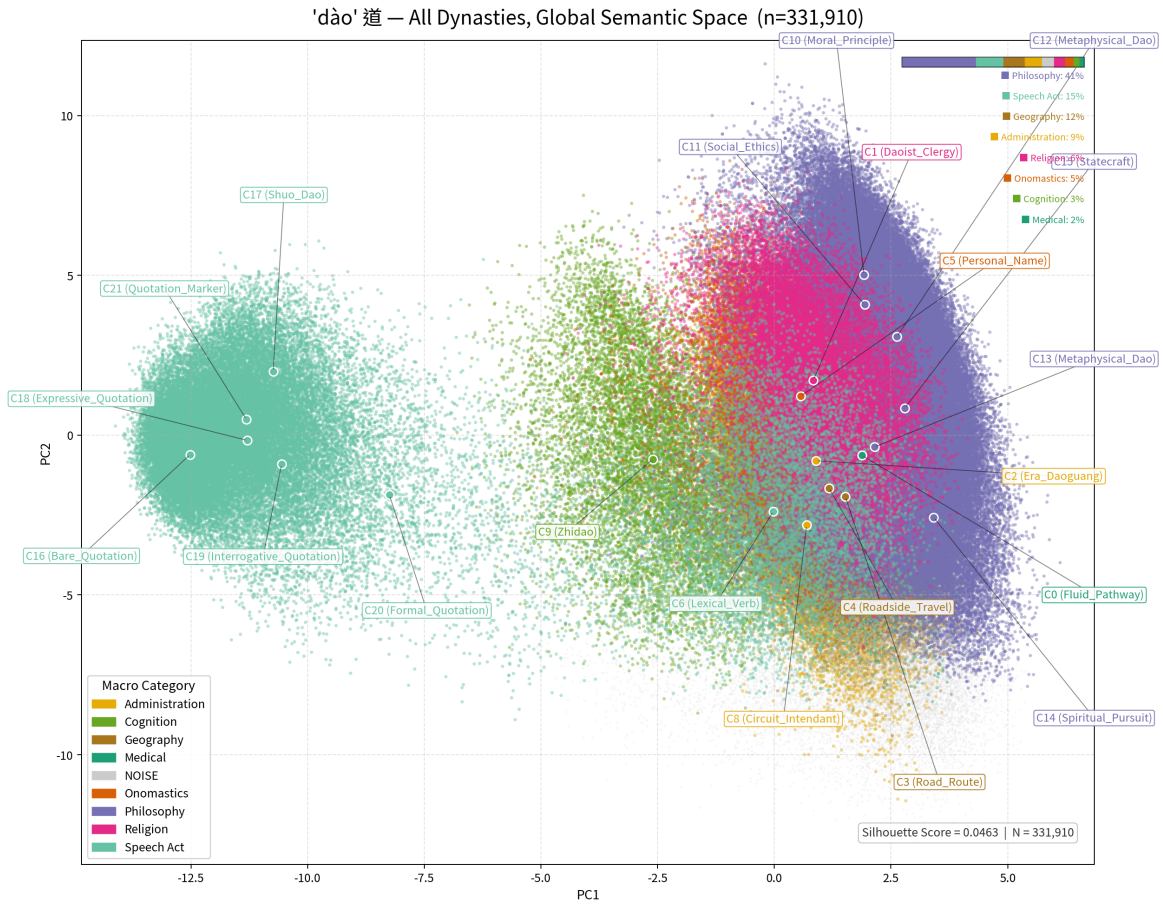


Figure 5: **Unified diachronic semantic space of 道 ($n = 331,910$).** This plot maps all historical periods onto an absolute PCA coordinate system. Initialized with $K = 20$ and parameterized with a *merge-ratio* of 0.95 and *split-ratio* of 0.1 (*sub-K*=6), the pipeline dynamically yielded 22 semantic clusters (C0–C21).

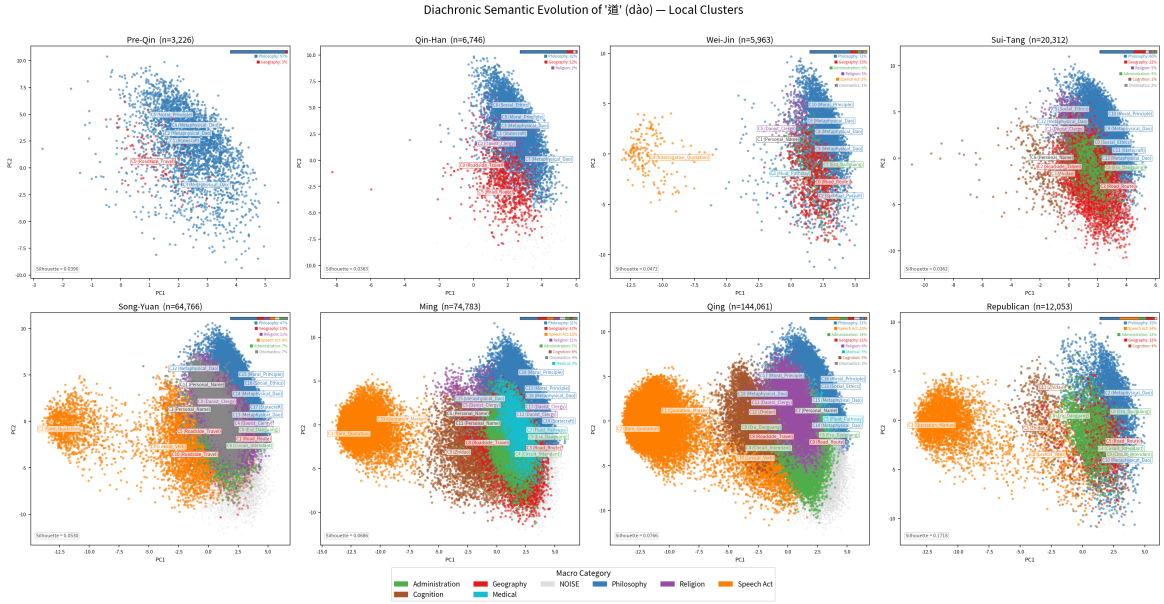


Figure 6: **Dynasty-specific (Local) semantic landscapes of 道**. To capture the semantic distributions within each specific era, the clustering pipeline was initialized with $K = 20$ for every period. We applied a *merge-ratio* of 0.95 and a *split-ratio* of 0.3, with the *sub-*K dynamically determined by the absolute sample size (n) of the respective era.

Table 4: The Hierarchical Semantic Coding Scheme for the Character **道** (*dào*)

Macro Category	**Sub-categories (Senses)**	**Typical Examples (Hanzi + *Pinyin* + Translation)**
Philosophy	Moral Principle, Political Ethics	知生之道 *zhī shēng zhī dào* 'the way of sustaining life'
	Metaphysical Dao	道之精妙 *dào zhī jīng miào* 'the subtlety of the Dao'
	Spiritual Pursuit, Statecraft	求道志 *qiú dào zhì* 'aspiration to seek the Dao'
Speech_Act	Lexical Verb	必能道之 *bì néng dào zhī* 'must be able to speak of it'
	Bare Quotation, Shuo Dao	說道：「弟之苦衷實難告人」 *shuō dào* 'said: …'
	Expressive, Interrogative, Formal	大笑道 *dà xiào dào* 'laughed and said'; 問道 *wèn dào* 'asked'
Geography	Road Route	不敢循驛道 *bù gǎn xún yì dào* 'dared not follow the post road'
	Roadside Travel	伏在道側蓬蒿之內 *fú zài dào cè* 'hid beside the road'
Religion	Daoist Clergy	道士復歸 *dào shì fù guī* 'the Taoist returned'
	Personal Name	杜遵道 *Dù Zūndào* (personal name)
Administration	Circuit Intendant	分巡福寧道 *fēn xún Fúníng dào* 'Circuit Intendant of Funing'
	Era Name (Daoguang)	道光中黃氏後人 *Dàoguāng zhōng* 'during the Daoguang reign'
Onomastics	Personal Name	蜀人吳師道為漢州太守 *Wú Shīdào wéi Hànzhōu tàishǒu* 'Wu Shidao served as Prefect of Hanzhou'
Cognition	Zhidao (Know)	知道今日饑荒苦 *zhī dào jīnrì jīhuāng kǔ* 'knows today's suffering'
Medical	Fluid Pathway (TCM)	通調水道 *tōng tiáo shuǐ dào* 'to regulate the water pathway (TCM)'
NOISE	OCR Error	栗難曉氣通南者謝班中絕景 (Meaningless OCR noise)

A Computational Evaluation of Syllabic Hypotheses for Rongorongo: Evidence from N-gram Analysis

Evgeniya Korovina

Institute of Linguistics, Russian Academy of Sciences
Moscow
jekorovina@gmail.com

Abstract

The study evaluates the hypothesis that the Rongorongo script of Easter Island functions as a syllabic substitution cipher where one symbol uniquely corresponds to one syllable. Using a genetic algorithm with a fitness function based on Rapa Nui n-gram statistics, we establish a performance baseline on controlled texts. Results show a strong correlation (0.75) between the algorithm's fitness score and decipherment accuracy, identifying a “noise threshold” at 2,500,000 points. We further demonstrate that cross-corpus genre variance significantly impacts recovery rates, with accuracy dropping by more than half when mismatched linguistic statistics are applied. Application to the CEIPP transliteration yields scores well below the noise threshold (1.0M–1.3M), suggesting a lack of a simple syllabic signal. However, testing the rongopy transliteration produces scores in the “gray zone” (up to 2.7M) and reveals stable mappings for high-frequency glyphs 200 and 006. The consistency of these results across independent inscriptions suggests that while a pure syllabic model is insufficient, specific structural simplifications of the script may capture latent linguistic patterns.

Keywords: Rongorongo, Rapa Nui, decipherment

1. Introduction

The decipherment of historical writing systems provides, on one hand, access to previously unknown aspects of the lives of their users (a subject of historical inquiry), and on the other, offers material for understanding how natural language can be represented in graphic form. One of the few such systems that remains undeciphered – at least in the sense of providing a verifiable and coherent reading of continuous text – is Rongorongo, the script of Easter Island, which fell out of use in the 1860s.

At the same time, there is little doubt among researchers that the language recorded in the Easter Island script is the ancestor of modern Rapa Nui or, at the very least, some Eastern Polynesian relative thereof (Horley 2021, Davletshin 2022). Furthermore, a relatively representative corpus of texts has come down to us – the total number of glyphs exceeds 12,000 (Horley 2021: 42), with nine inscriptions containing more than 500 glyphs each. These circumstances, it seems, create favorable starting conditions for the formal study of this writing system, particularly when compared to the Indus Valley (Harappan) script, the Epi-Olmec (Isthmian) script, or the Phaistos Disc.

A limiting factor for the direct comparison of Rongorongo material with Rapa Nui data is that modern Rapa Nui is a low-resource language with a relatively small corpus of published texts. Moreover, in the late 19th century, Rapa Nui underwent significant influence from Tahitian, which affected both its vocabulary and its grammar (Kieviet 2017: 15). Another obstacle consists of questions typical for undeciphered writing systems regarding the identification of glyphs: namely, determining which shapes should be considered variants of a single grapheme, and in which cases a shape represents a ligature of several glyphs.

Despite the enormous interest in attempts to decipher Rongorongo – a recent article by Wieczorek (2025) lists about 100 scholarly works published in the last 25 years alone, containing some 80 motivated suggestions for the reading of individual signs, varying widely in their degree of plausibility – the question of the script’s basic characteristics still seems to lack a satisfactory answer. For the most part, the signs that have been interpreted are described as logograms with nominal meanings, directly related to what the sign supposedly depicts. By contrast, signs that could potentially be interpreted as grammatical markers are almost entirely absent from such proposals. Moreover, since the late 1960s, the idea that Rongorongo exhibits a "telegraphic" style, largely devoid of grammar, has gained currency among researchers (cf. Кондратов 1968: 175; Pozdniakov & Pozdniakov 2007: 5). As a result, existing interpretations appear to be quite far removed from the structure of Polynesian languages, which makes their verification problematic and suggests the need for a different, more formal approach.

Unlike historical writing systems, whose successful decipherment has often relied on a combination of disparate factors (bilingual texts, iconographic context, and historical knowledge) and is therefore difficult to formalize, the principles underlying classical cryptography are far better formalized and algorithmically tractable. While various algorithms can efficiently solve simple substitution ciphers for alphabetic languages (see, e.g., Jakobsen 1995; Kang and Lee 2024), their application to syllabic systems presents a

Proceedings of LT4HALA 2026 @ LREC 2026, pages 177–183.
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

unique challenge. In such cases, the alphabet of potential referents is significantly larger than a standard phonemic alphabet yet remains more computationally tractable than a full logographic (word-based) system. In this paper, I treat Rongorongo as a large-alphabetic monoalphabetic substitution cipher, focusing on a one-to-one mapping between glyphs and syllables. It is important to clarify that this study does not account for homophonic substitution, where a single syllable could be represented by multiple distinct signs. While the script is widely considered to be logo-syllabic in nature, the 'purely syllabic' hypothesis frequently appears in literature. By modeling the script as a simple syllabic substitution system, I aim to provide a formal stress-test of this assumption. Testing the simble substitution hypothesis is a methodological necessity: if the script's structure is fundamentally more complex (e.g., homophonic or logo-syllabic), a simple substitution model should, by definition, fail to converge. Thus, the inability of the algorithm to find a stable solution serves as a formal indicator that the 'one-to-one' syllabic model is insufficient to describe the Rongorongo system. To do so, I will first analyze the feasibility of deciphering such a cipher using data from Rapa Nui – by artificially "encrypting" plain texts and testing whether the algorithm can recover the original – and then apply the same method to the transliterated Rongorongo data.

The paper proceeds as follows. Section 2 surveys prior formal and computational approaches to Rongorongo. Section 3 details the datasets, encompassing the Rongorongo corpus alongside texts in Rapa Nui and other Eastern Polynesian languages. Section 4 presents the methodology and core algorithm. Section 5 reports experimental results. Section 6 interprets their implications for Rongorongo decipherment and suggests avenues for future research. Section 7 concludes with a summary of key findings.

2. Previous Work

The first algorithmically oriented study of the Easter Island script was apparently undertaken in the 1960s by a research group led by Yuri Knorozov. The most comprehensive account of this work is presented in Kondratov (1968), where the author concludes, based on an analysis of the appearance of new glyphs, that Rongorongo falls formally between a logosyllabic and a purely logographic writing system. The study also identified glyphs that are evenly distributed throughout the texts, which were interpreted as potential grammatical markers. However, the technological limitations of the available computing technology at the time prevented further development of this line of research.

The next study of this kind was undertaken only in 2007, when Pozdniakov and Pozdniakov conducted an experiment based on the assumption that the number of distinct glyphs in the Easter Island script is substantially smaller than commonly thought, amounting to just over 50 signs. This hypothesis stemmed from the observation that published decipherment proposals – such as those by Fischer (1997) and Fedorova (2001), based on logosyllabic or logographic assumptions – remained far removed from the structure of Polynesian languages. Pozdniakov and Pozdniakov were able to demonstrate some statistical similarities, but their approach relied on heuristic comparisons of frequency distributions, which did not yield concrete readings or a reproducible mapping between signs and linguistic units, and the hypothesis that Rongorongo is predominantly syllabic in nature has not, overall, found broad acceptance among other researchers.

The next study of this kind was Harris (2010), who employed various computational methods – in particular, PCA and LSA – to uncover the structure of Rongorongo texts and the relationships between them. As a result, he obtained a formal clustering of the inscriptions that closely corresponded to the traditional classification.

Among the more successful attempts to identify formal similarities between Rapa Nui and the Easter Island script is the study by Horley, Davletshin, and Wieczorek (2018). The authors demonstrated linguistic parallels involving palindromes in Rongorongo, as well as various types of reduplication. However, this work was based on manual data analysis.

Finally, any brief overview must mention the work of Gregorio de Souza, who developed the rongopy repository and applied various algorithmic methods. In particular, he attempted to crack Rongorongo using a genetic algorithm scored by an LSTM language model. De Souza correctly identified that a direct frequency-based decipherment is hindered by the high variability of sign distributions across different subsets of the corpus. To mitigate this, he followed an approach like Aldarrab and May (2021), encoding sequences multiple times based on frequencies from random samples. Subsequently, in 2023, he sought to apply a Seq2Seq method for the same purpose. This work represents the most significant recent contribution to the computational analysis of the script.

As this overview demonstrates, while recent work has highlighted the challenges of frequency-based matching, there has been no systematic effort to formally measure the stability of these mappings under a controlled cryptanalytic framework. The present study addresses this gap.

3. Data

3.1 Rongorongo data

According to the counts of Horley (2021: 42), the total size of the Rongorongo corpus amounts to 12,289 glyphs (i.e., individually written elements). The majority of these – 10,879 glyphs – are found on twelve inscriptions, each exceeding 400 glyphs in length. For the purposes of this pilot study, only two of the longest and best-preserved inscriptions were selected: Tablet A (1,361 glyphs) and Tablet B (948 glyphs). These were chosen not only for their length and state of preservation, but also because, unlike, for example, I (Santiago Staff), they do not consist entirely of lists, which would preclude direct comparison with Polynesian language data. Facsimiles and drawings of these tablets are published in Barthel (1958), Fischer (1997), and Horley (2021).

However, for the purposes of the present analysis, it is not the glyph images themselves that are used, but rather transliterations of them. The primary catalogue for the study of the Easter Island script is the one developed by Barthel (1958), subsequently modified by CEIPP and made available online at http://kohaumotu.org/rongorongo_org/. While other catalogues exist (discussed below), their full transliterations have traditionally remained difficult to access. Recently, however, a machine-readable version based on Horley's structural analysis was made available via the rongopy repository (de Souza 2023).

In addition to the Barthel catalogue and the CEIPP version thereof – both containing around 500 glyphs – there exist catalogues that assume a significantly smaller number of signs. Notably, Pozdniakov & Pozdniakov (2007) identify 50 basic signs, while Horley (2021: 443) provides a list of 130 glyphs that "stand good chances of being independent". At the same time, Davletshin (2017) has shown that some graphically very similar signs are likely to be distinct based on collocation analysis, which may point to a larger number of signs in the writing system.

However, in addition to the usual debates about individual signs in the catalogue – typical for undeciphered writing systems – Rongorongo presents a specific problem: the presence of ligatured glyphs. The existence of such composite forms is evident from parallel passages, where two glyphs are written joined together in one text but separately in another. At the same time for many ligatures the nature of the script is such that such explicit demonstration of separateness is impossible in most cases. For example, see figure 1, glyph B206 could be interpreted as a combination of B200 and B006, as suggested by Horley and by Pozdniakov & Pozdniakov. However, in the case of B226, Horley following (Pozdniakov 2019) treats it not as a combination of B220 + B006, but as a distinct sign, whereas early he decomposes it even further, and in Pozdniakov & Pozdniakov (2007) he also breaks down B220 into B200 + B060.

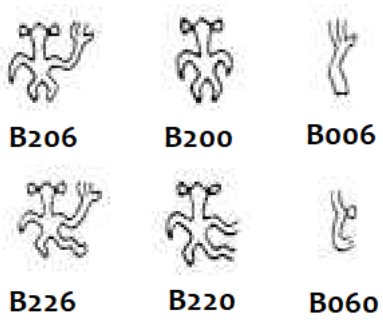


Figure 1: Appearance of the rongorongo glyphs mentioned above

Considering the issues outlined above, two types of transliteration were used for the purposes of this study: the original CEIPP encoding and an encoding from rongopy.

3.2 Rapa Nui data

The language of Easter Island is a low-resource language. The total size of the available corpus of texts in this language amounts to just over one million words (Kieviet 2017), approximately half of which consists of a recent translation of the New Testament (2018). Although the earliest transcriptions of Rapa Nui in the Latin alphabet date back to the late 18th century – when Rongorongo was still in use – the total volume of texts in the language remained around 130,000 words until the 1980s.

For the purposes of this study, five subcorpora of Rapa Nui texts were selected:

1. The text of Manuscript E (oral history, recorded by Rapa Nui speakers in the early 20th century).
2. Songs from the Campbell collection (1970).
3. The collection of texts published by Englert (1980), gathered in the 1930s (various genres: oral history, life stories, instructions).
4. Texts from Part 6 of the Rapa Nui textbook by Weber & Weber (1990) (Mai ki hāpī…).
5. The text of the Gospel of Matthew.

This selection, it is believed, provides a comprehensive representation of all the available Rapa Nui textual material. Each of the subcorpora exceeds 10,000 words in length, which is comparable to the volume of the Rongorongo corpus.

For the purposes of this experiment, all texts were converted to a unified format. All punctuation marks were removed, and orthography was standardized to a form that does not mark vowel

length or the glottal stop, as these were used inconsistently in early texts.

	Syllables
Manuscript E	29,422
Campbell	23,274
Englert	67,146
Mai ki hāpī… 6	54,289
Matthew	70,394

Table 1: Length of the texts selected for analysis in syllables

4. Methodology

The research is structured according to the following logic: before applying the algorithm to the actual Rongorongo tablets, it is necessary to understand its behavior and limitations on controlled material. To this end, a series of experiments is conducted on Rapa Nui language texts, where the true correspondence between signs and syllables is known. Only after establishing which syllables and with what degree of reliability the algorithm is able to recover under "laboratory conditions" can the results obtained on the undeciphered script be interpreted.

An important characteristic of the experiments is the length of the analyzed fragments – 1000 signs. This parameter was chosen deliberately: it roughly corresponds to the length of the surviving Rongorongo tablets (the largest of them contain 800-1200 signs). Thus, the simulated conditions are as close as possible to the real ones.

In this work, Rongorongo is considered as a simple substitution cipher, where each graphic sign corresponds to one syllable of the Rapa Nui language. Under this assumption, the decryption task reduces to finding a mapping from the set of unique signs (graphemes) to the set of syllables. Since the total number of distinct signs on the tablets exceeds 50, the search space (the number of possible keys) is astronomical, making direct brute-force enumeration impossible.

To solve this problem, a genetic algorithm is employed – an optimization method that mimics natural selection. The algorithm operates on a population of candidate keys, gradually improving their quality through selection and mutation operators.

A key feature of the developed method is a specialized fitness function that, instead of directly calculating probabilities, uses a system of bonuses and penalties. This allows the algorithm to more quickly identify linguistically plausible solutions.

When evaluating a candidate, the algorithm decrypts the fragment using the tested key and scans the resulting syllable sequence, identifying bigrams and trigrams that are present in the reference frequency dictionary. To enhance search efficiency, a special class is distinguished – the "golden fund" of linguistic chains, which includes sequences with a frequency of occurrence exceeding 0.5% in the reference corpus. The first occurrence of such a chain in the text receives an increased bonus, encouraging the algorithm to seek keys that generate the most characteristic sequences for the language.

An additional constraining factor is a penalty for violating the rank correspondence between the frequency of signs in the encrypted text and the frequency of syllables in the language. The algorithm proceeds from the assumption that the most frequent signs should correspond to the most frequent syllables. The penalty is proportional to the rank discrepancy; however, its strictness is dynamically reduced towards the end of each training phase. This allows the algorithm in its final iterations to consider non-standard correspondences that might turn out to be correct, even if they do not fit into the rigid rank model.

The process of searching for the optimal key is divided into two sequential stages:

Phase 1 (Core). In the first stage, the algorithm works only with the top 20 most frequent signs occurring in the encrypted fragment. The goal is to create a reliable frequency framework for the decryption, matching the core graphemes with the core syllable inventory.

Phase 2 (Expansion). After a stable solution for the top of the frequency list is fixed, the remaining signs (up to the top 50) are introduced. The algorithm builds the overall picture, preserving the already found correspondences and selecting options for the rarer graphemes.

Each stage uses a population of 80-110 individuals, and evolution continues for 110 generations. The genetic search operators include elitist selection (preserving the best solutions) and swap mutations (randomly exchanging values between two positions in the key).

The experiment is structured as follows:

- Formation of reference frequency profiles. Based on several corpora of Rapa Nui texts (Campbell, Manuscript E, Englert, etc.), frequency dictionaries of unigrams, bigrams, and trigrams are computed.
- Generation of test fragments. A continuous fragment of 1000 syllables is randomly extracted from the source text. Then, a random mapping "syllable → numeric code" (simulating a cipher) is created, and the fragment is encoded. Thus, for each test, the true key is known, allowing for result verification.

- Algorithm execution and metric recording. For each fragment, the decryption algorithm is run multiple times. Three key indicators are recorded:
    - Score – the internal evaluation of key quality, used to guide the evolution.
    - Accuracy – the percentage of unique signs for which the algorithm proposes the correct syllabic correspondence.
    - Coverage – the proportion of the text (in signs, accounting for repetitions) that is correctly deciphered. This indicator is the most important for the subsequent interpretation of results on the actual tablets, as it reflects what portion of the message could potentially be considered readable.

This procedure allows not only for testing the algorithm's effectiveness but also for accumulating empirical data on which specific syllables and with what reliability are recovered given a fixed text length (1000 signs). Subsequently, this data serves as the basis for interpreting the results obtained when working with the actual Rongorongo tablets.

5. Results

The algorithmic performance is summarized in Table 2. When the model is tested using N-gram statistics derived from its own internal corpus, it correctly identifies an average of 26 syllables, resulting in a coverage of over 80% of the text. These figures demonstrate the robustness and viability of the proposed pilot algorithm for deciphering materials of this nature.

However, the performance drops significantly in cross-corpus scenarios. When using statistics from a different corpus, the average number of correctly identified syllables falls to 10, covering only 37% of the text – a result clearly insufficient for a successful “decipherment.” This drop highlights the high degree of genre-specific variance within Rapa Nui texts. The lowest cross-genre performance was observed when analyzing fragments from Manuscript E using statistics from the Gospel of Matthew, yielding an average of only 4.6 correct syllables.

When utilizing an aggregate statistical model based on all five sub-corpora, the algorithm correctly identifies an average of 17 syllables, achieving a text coverage of 62%.

Data regarding the most frequently correctly identified syllables for select corpus pairs are presented in Table 3. When utilizing statistics from the same text as the source fragment, many syllables are recovered in over 90% of cases. However, even in the most challenging scenario – analyzing Manuscript E using statistics from the Gospel of Matthew – the syllables “a” and “i” are correctly identified in 67.2% and 52% of cases, respectively.

The fitness score serves as a relatively accurate predictor of the success of the syllabic decipherment. Its correlation with the percentage of correctly identified syllables is 0.75 (as shown in Figure 2). Specifically, when the score exceeds 4,000,000, a high accuracy rate in syllable identification can be expected. Conversely, if the score falls below 2,500,000, the resulting solution is likely nothing more than statistical noise. The range between these two values represents a “gray zone,” where both successful and unsuccessful decipherment attempts are possible.

When testing the Rongorongo transliterations, specifically the CEIPP version, the scores obtained for both analyzed tablets ranged between 1,000,000 and 1,300,000. These values are extremely low, falling well below the previously established noise threshold of 2,500,000. This indicates that any syllabic “solution” generated in this context is, likely, merely a byproduct of the algorithm’s inherent tendency to suggest a mapping, regardless of the data's linguistic validity. Consequently, even if a few signs were coincidentally identified correctly in each run, the lack of statistical significance makes it impossible to advance any verifiable claims.

Furthermore, it is noteworthy that each of the five sub-corpora yielded approximately the same low score ranges when applied to the Rongorongo inscriptions. This consistency across diverse linguistic datasets (ranging from the Gospel of Matthew to folklore manuscripts) suggests that the failure to reach the convergence threshold is not a result of genre mismatch or lexical bias in the reference data. Instead, it points to a fundamental structural discrepancy between the CEIPP transliteration and the simple substitution syllabic model.

When testing the rongopy transliteration, the results were significantly higher, ranging between 2,000,000 and 2,700,000, effectively placing them within the previously defined “gray zone.” Notably, the mappings assigned by the algorithm for certain glyphs showed considerably more stability.

For Tablet A, glyph 200 was consistently interpreted as either the syllable “a” or “i”, while glyph 006 was identified in most cases as “i” or “te”. Other signs continued to exhibit essentially random interpretations. A similar pattern emerged for Tablet B: glyph 200 was again mapped to “a” or “i” (with “a” being assigned in 80% of runs), while glyph 006 received an “i” interpretation in 53% of cases (with alternative hypotheses being a, e, te, u).

	Campbell	Manuscript E	Englert	Mai ki hāpī	Matthew	Joined
Campbell	25.5 (80.5%)	6.6 (29.8%)	6.9 (30.6%)	7.4 (34.4%)	6.3 (31.2%)	11.9 (49.9%)
Manuscript E	4.8 (22.7%)	22.6 (70.6%)	15.3 (53.7%)	8.9 (35.8%)	4.6 (23.3%)	14.7 (53.8%)
Englert	5.0 (21.0%)	14.5 (51.1%)	23.0 (71.9%)	10.1 (37.5%)	6.2 (27.3%)	15.0 (52.2%)
Mai ki hāpī	7.5 (34.6%)	11.7 (47.6%)	15.8 (59.4%)	27.6 (85.4%)	13.6 (50.5%)	20.1 (69.4%)
Matthew	7.5 (36.3%)	9.0 (41.4%)	11.1 (45.1%)	20.1 (69.7%)	31.7 (93.8%)	27.2 (86.4%)

Table 2: Average number of correctly identified syllables by text (rows) / corpus (columns) pair, and (in brackets) the percentage of the text read correctly (coverage).

	Campbell	Manuscript E	Englert	Mai ki hāpī	Matthew
a	90,4	87,2	83,2	94,4	99,2
u	91,2	79,2	75,2	92,8	96,8
i	91,2	78,4	82,4	94,4	99,2
te	92,8	76	86,4	96	99,2
o	94,4	76	80,8	94,4	99,2
ma	88,8	76	76	94,4	95,2
na	86,4	75,2	67,2	85,6	100
ka	71,2	75,2	76	85,6	99,2
he	36	74,4	80	81,6	100
e	92	73,6	74,4	96	100

Table 3: The most “cracked” syllables when a fragment is deciphered based on the statistics of its own text.

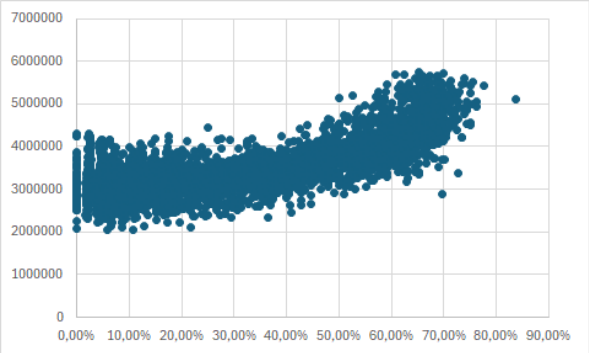


Figure 2: Correlation between the score value and the percentage of syllables deciphered

Such consistency across two independent texts suggests that these results might not be entirely accidental. However, given that the scores remain at the lower boundary of the “gray zone,” it stays possible that these patterns are merely a more structured form of statistical noise rather than a definitive decipherment.

6. Conclusion & Future Work

In this study, the syllabic hypothesis for the Rongorongo script was evaluated using a stochastic optimization algorithm. A fitness score metric was developed, which demonstrated a strong correlation (0.75) with actual decipherment accuracy in Rapa Nui control tests, establishing a clear threshold between linguistic signal and statistical noise.

While the results do not warrant a complete rejection of the syllabic hypothesis, the frequency characteristics of Rongorongo differ substantially from Latin-script Rapa Nui texts. This discrepancy suggests several possibilities: the texts may belong to an undocumented or lost genre; existing transliteration systems may be and likely is suboptimal; or Rongorongo may be a non-syllabic system incorporating numerous logograms or multi-glyph encoding. At the same time, the consistent mapping of glyphs 200 and 006 to the same values across two independent texts suggests that high-frequency signs might be identifiable through this method, though such interpretations remain speculative at this stage.

Future research will focus on testing the algorithm against logographic and logo-syllabic hypotheses, as well as models where a single Rapa Nui syllable corresponds to multiple signs. The research results should also be tested across different languages, for example, by comparing

the outcomes of decryption attempts on certain Rapa Nui texts with texts in Maori or Tahitian to further refine the cross-linguistic reliability of the framework. Additionally, this algorithm should be tested on fragments of varying lengths.

7. Acknowlegement

This research was supported by project FMNE-2025-0003 conducted at the Institute of Linguistics, Russian Academy of Sciences.

8. Bibliographical References

Aldarrab and May (2021) “Can Sequence-to-Sequence Models Crack Substitution Ciphers?” In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 7226–7235, Online. Association for Computational Linguistics.

Barthel, T. (1958). Grundlagen zur Entzifferung der Osterinselschrift. Walter de Gruyter.

Davletshin, A. (2017). “Allographs, graphic variants and iconic formulae in the Kohau Rongorongo script of Rapa Nui (Easter Island).” The Journal of the Polynesian Society 126 (1), 61-92.

Davletshin, A. (2022). "The script of Rapa Nui (Easter Island) is logosyllabic, the language is East Polynesian: Evidence from cross-readings." The Journal of the Polynesian Society, 131(2), 185-220.

de Souza (2023) *rongopy*: Ideas for the decipherment of rongorongo using machine learning in Python (https://jgregoriods.github.io/rongopy/)

Fischer, S. (1997). Rongorongo: the Easter Island script: history, traditions, texts. Oxford University Press.

Harris, M. (2010). Corpus Linguistics as a Method for the Decipherment of Rongorongo. Unpublished MA thesis in Applied Linguistics, Birkbeck University, London.

Horley P. (2021) Rongorongo: Inscribed Objects from Rapa Nui. Viña del Mar: Rapanui Press.

Jakobsen, T. (1995). "A fast method for cryptanalysis of substitution ciphers". Cryptologia, 19(3), 265-274.

Kang, H., and Lee, W. (2024). "A Robust Decryption Technique Using Letter Frequency Analysis for Short Monoalphabetic Substitution Ciphers" Journal of Computing Science and Engineering, Vol. 18, No. 3, pp. 144-151

Kieviet, P. (2017). A grammar of Rapa Nui. Language Science Press.

Pozdniakov, I. & Pozdniakov, K. (2007) “Rapanui Writing and the Rapanui Language: Preliminary Results of a Statistical Analysis”, Forum for Anthropology and Culture, 3: 3–36.

Pozdniakov K. (2019). Correlation of Graphical ’Distinctive Features’ in Rongorongo as an Additional Resource for the Construction of the Sign Catalog. “Easter Island and the Pacific”, 9th International Conference on Easter Island and the Pacific (EIPC 2015), Ethnological Museum Dahlem, Berlin, Jun 2015, Berlin, Germany. pp.393-406.

Wieczorek, R. (2025). "Aspects of rongorongo decipherment." Waka Kuaka: The Journal of the Polynesian Society, 134(3), 343-392.

Кондратов А.М. (1968) Статистические методы дешифровки некоторых письмен древнего востока и средиземноморья [Statistical Methods for Deciphering Certain Scripts of the Ancient East and the Mediterranean]. PhD thesis.

Федорова И.К. (2001) «Говорящие дощечки» с острова Пасхи. Дешифровка. Чтение. Перевод. ["Talking Tablets" from Easter Island. Decipherment. Reading. Translation.] Spb: MAE RAS.

Building a Corpus and Database for Rare and Undeciphered Scripts

Beáta Megyesi[1], **Rune Rattenborg**[2], **Benedek Láng**[3], **Michelle Waldispühl**[4], **and Mihály Héder**[5]
[1]Stockholm University, Sweden
[2]Lund University, Sweden
[3]Eötvös Loránd University (ELTE), Hungary
[4]University of Oslo, Norway
[5]Budapest University of Technology and Economics / HUN-REN SZTAKI, Hungary
{ beata.megyesi@ling.su.se, rune.rattenborg@klass.lu.se, lang.benedek@gtk.elte.hu, michelle.waldispuhl@ilos.uio.no, mihaly.heder@sztaki.hu }

Abstract

Historical sources written in rare or undeciphered scripts represent an immense but underexploited part of the world's cultural and linguistic heritage. Their study is often hindered by fragmentary preservation, non-standard symbol systems, and the absence of interoperable digital resources. While recent advances in imaging, transcription, and computational analysis have improved access to historical texts, most tools rely on large quantities of labeled data and standardized encodings, requirements that are rarely met for rare or unknown writing systems. This paper presents the design and methodology of a new corpus and database dedicated to rare and undeciphered scripts worldwide. The resource integrates high-quality images, transliterations, transcriptions, linguistic annotations, and metadata within a unified data model tailored for low-resource and non-standard scripts. By adhering to FAIR principles and existing standards for linguistic and cultural heritage data, the database enables reproducible, interdisciplinary research across philology, linguistics, cryptology, and computer science. The paper outlines the data collection and digitization workflow, describes the metadata and database architecture, and demonstrates applications in analysis and decipherment.

Keywords: decipherment, historical writing systems, rare scripts

1. Introduction

Historical sources written in rare or undeciphered scripts are among the most valuable yet least accessible parts of the world's cultural record. They preserve unique evidence of linguistic diversity, intellectual history, and cultural exchange, but remain systematically underused because they are difficult to access, standardize, and analyze. Examples of such writings include some of the most famous enigmas in the history of writing: the Indus Valley inscriptions from South Asia, the Rongorongo tablets of Easter Island, the Vinča signs of prehistoric southeastern Europe, the Linear A tablets from the Eastern Mediterranean and Proto-Elamite tablets from the ancient Near East, and the enigmatic Voynich manuscript from early modern Europe. Although each belongs to a distinct script tradition, material context, and historical setting, all present similar challenges: fragmentary evidence, uncertain linguistic affiliation, and a lack of standardized encoding or digital representation.

Traditional philological and epigraphic methods rely on expert knowledge, comparative materials, and well-established reference corpora. While these approaches have produced major breakthroughs in well-documented traditions, they encounter severe limitations when evidence is fragmentary, symbol inventories are idiosyncratic, or the underlying language and script remain unknown.

Digital technologies have substantially improved the documentation and legibility of historical materials. High-resolution imaging, multispectral enhancement, and handwritten text recognition (HTR) now form part of the standard toolkit for manuscript and epigraphic studies. Yet these tools depend on abundant labeled data and stable script encodings, conditions rarely met for rare or non-standard writing systems. In many cases, even a few hundred annotated pages are unattainable, and even basic Unicode (The Unicode Consortium, 2019) coverage may be lacking. Consequently, many historically and linguistically important sources remain outside the reach of computational methods and large-scale linguistic analysis.

Recent advances in historical cryptology and computational decipherment illustrate both the potential and the challenges of algorithmic assistance. Techniques such as language modeling, frequency analysis, and heuristic search can accelerate interpretation, but existing solutions remain specialized, fragmented, or dependent on expert intervention. There is still no generic, end-to-end workflow capable of supporting the full range of tasks required to process rare or undeciphered scripts—from image segmentation and symbol inventory creation to linguistic interpretation and contextualization.

A systematic, interoperable collection and database of historical sources with rare and un-

Proceedings of LT4HALA 2026 @ LREC 2026, pages 184–196.
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

known writing systems is therefore an essential step forward. Such a resource enables the preservation and comparative study of diverse scripts, supports the development of transferable AI and HTR models under low-data conditions, and provides a foundation for reproducible computational experiments. It also facilitates collaboration across disciplines—linking linguistics, archaeology, history, and computer science—and ensures that cultural heritage materials can be documented and analyzed within FAIR and open-data frameworks.

In this paper, we present the design and methodology of a new corpus and database dedicated to rare and undeciphered scripts. Section 2 reviews related work on the study and digitization of rare and unknown writing systems. Section 3 describes the data collection and digitization workflow, including source types, datasets, image acquisition, transcription, and metadata tailored to rare and undeciphered writings. Section 4 presents the database architecture and its alignment with established standards for linguistic and cultural heritage data. Section 5 discusses collaboration and long-term aims, and Section 6 concludes the paper.

2. Background

The study of historical writings provides an invaluable window into the past, offering insights into the cultures, languages, and intellectual traditions of earlier societies. Traditionally, research on ancient and rare scripts has relied on philological and epigraphic expertise: specialists meticulously examine texts within their linguistic and historical contexts, often focusing on sources from the same region, period, and language. While such methods have produced foundational breakthroughs—from the decipherment of Egyptian hieroglyphs by Champollion (1822) to the work on Linear B by Chadwick (1958) or the Maya script (Coe, 2011)—they remain limited by the fragmentary nature of the data, the scarcity of comparative material, and the time-intensive process of manual analysis.

Over the past few decades, the digital humanities have transformed access to and analysis of historical materials. Digitization initiatives, such as high-resolution imaging, 3D modeling, and optical character recognition (OCR), have made it possible to preserve, share, and annotate ancient texts at scale (Babeu, 2011; Santis and Rossi, 2019; Soriano and Espinosa, 2021). However, these technologies perform best on well-preserved and standardized scripts, and typically fail when confronted with irregular, damaged, or non-standard writings, features that characterize many rare and undeciphered sources. Moreover, variation in metadata standards, annotation practices, and encoding formats limit the interoperability of existing resources, making large-scale comparison and computational analysis difficult.

Current repositories dedicated to artifact corpora vary widely in scope: for instance, the cuneiform corpus contains over 550,000 inscriptions (Streck, 2010; Rattenborg et al., 2023), compared to around 500,000 Latin (Lloris, 2015), 200,000 Greek (Liddel, 2017), 15,000 Ancient South Arabian (Avanzini, 2009), 8,000 Runic (Williams et al., 2022), 6,000 Linear B (Aurora, 2015), and an estimated 20-100,000 texts combined for Egyptian hieroglyphic, hieratic, and demotic inscriptions (Trismegistos, 2025; Thesaurus Linguae Aegyptiae, 2025). Beyond these better-known cases, thousands of sources remain isolated: inscriptions on bone, wood, clay, parchment, and stone, or manuscripts written in unknown symbols or numeral sequences. Many of these represent linguistic traditions yet to be deciphered or even identified.

Efforts to analyze such material have increasingly turned to computational linguistics, cryptology, and AI. Historical cryptology—the systematic study of historical ciphers and secret writings often featuring non-standard symbol systems and low-resource language data (Kahn, 1967; Megyesi et al., 2024a)—has demonstrated how algorithmic methods can complement traditional expertise. The DECRYPT[1] project (Megyesi et al., 2020) pioneered this approach by creating large-scale digital resources for early modern European ciphertexts and cipher keys, including the DECODE cipher database (Megyesi et al., 2019; Héder and Megyesi, 2022) and Megyesi et al. (2024), the HistCorp multilingual historical corpora for European languages (Pettersson and Megyesi, 2018) and (Pettersson and Megyesi, 2024), and advanced cryptanalysis tools such as CrypTool 2 (Esslinger, 2024).

Building on prior research in automated text and cipher analysis, the work presented here broadens the scope from European sources to rare and undeciphered scripts worldwide, integrating AI, computer vision, and linguistic modeling to facilitate transcription, decipherment, and contextual interpretation.

Parallel research in automatic decipherment and transcription has shown the potential of computational approaches. Statistical and machine learning methods have been successfully applied to historical ciphers and ancient scripts (Dou and Knight, 2012; Ravi and Knight, 2011b,a; Knight et al., 2012; Reddy and Knight, 2011, 2012; Hauer and Kondrak, 2016; Lasry, 2018; Yin et al., 2019; Lasry et al., 2020; Leierzopf et al., 2021a,b; Kopal and Waldispühl, 2022; Souibgui et al., 2022; Lasry et al., 2023). Similar methods have been explored for undeciphered scripts such as Luwian hieroglyphs (de Lin and Knight, 2006), Ugaritic (Sny-

[1] de-crypt.org

der et al., 2010), and Iberian (Luo et al., 2021), though large-scale success remains elusive (Ferrara and Tamburini, 2022). Recent breakthroughs, such as the decipherment of Linear Elamite (Desset et al., 2022), Amorite vocabularies (George and Krebernik, 2022), and the Kushan script (Bonmann et al., 2023), highlight the enduring importance of systematic documentation, transcription, and interdisciplinary collaboration.

In parallel, research on script recognition and image enhancement (Hochberg et al., 1997; Easton and Knox, 2016; Fornés et al., 2017; Baró et al., 2019; Chen et al., 2020; Souibgui et al., 2022; Szigeti and Héder, 2022) and neural reconstruction of damaged manuscripts (Fetaya et al., 2020; Assael et al., 2022; Papavassileiou et al., 2023) has advanced rapidly. Platforms such as Transkribus (Tra, 2024) and eScriptorium (eSc, 2025) offer infrastructures for handwritten text recognition, but their reliance on large training datasets makes them unsuitable for rare or low-resource scripts. Even specialized tools such as Fabricius (Fab, 2024), developed for Egyptian hieroglyphs, remain restricted in scope. Consequently, generic and flexible transcription tools that can operate with minimal annotated data are still lacking.

These technological limitations underscore the continued need for human expertise and interdisciplinary cooperation. As Terras (2012, 2016) argues, digital palaeography and crowdsourced manuscript analysis must integrate historical knowledge with computational methods to achieve meaningful progress. The combination of domain expertise from philologists, linguists, archaeologists, computer scientists, and cryptologists offers a powerful model for advancing the study of rare and undeciphered scripts.

The present project builds directly on this interdisciplinary tradition. By creating a global corpus and database that integrates images, transliterations, linguistic annotation, and metadata, and eventually sign inventories, it aims to provide a unified and extensible infrastructure for studying the world's writing diversity. The resource is intended to support both humanistic and computational research, enabling the preservation, comparison, and analysis of rare scripts, and it aligns with the LREC mission to promote FAIR linguistic data. Designed according to the Linguistic Linked Open Data[2] (LLOD) principles and compatible with CLARIN[3] and ELRA[4] infrastructures, it contributes to the long-term goal of sustainable, reproducible, and inclusive language resources.

[2] https://linguistic-lod.org
[3] https://www.clarin.eu
[4] https://www.elra.info

3. The DESCRYPT Corpus

3.1. Scope and Source Types

The DESCRYPT[5] project collects and curates a wide range of materials featuring rare, non-standard, or unknown scripts, including but not limited to (i) ancient inscriptions (e.g., Ajami, Cuneiform, Linear A/B, Old South Arabian, Linear and Proto Elamite, runic corpora), (ii) manuscript traditions with specialized or hybrid notation (e.g., shorthand systems, also called stenographic writing, artificial language schemes, mixed symbol sets), and (iii) encrypted or otherwise encoded documents from later periods (diplomatic, scientific, or private correspondence; political or religious materials). The corpus includes both famous enigmas and lesser-known fragments found on bone, wood, clay, stone, parchment, or paper, whether preserved in archives, libraries, museums, or private collections. Figure 1 illustrates some well-known rare writings.

The first group also encompasses undeciphered and rare historical scripts in natural languages from later epochs (e.g., Rongorongo or the contentious so-called Hungarian runic script), while the second includes early modern artificial language schemes and "common writings" devised by philosophers and linguists such as John Wilkins, Athanasius Kircher, René Descartes, Isaac Newton, Gottfried Wilhelm Leibniz, and Marin Mersenne. These often aimed at constructing universal systems of representation and communication. Alongside such projects, shorthand systems, which virtually disappeared by the twelfth century and were reinvented around 1600, proliferated into numerous schools and national traditions; over four centuries, approximately 450 methods emerged in print, especially in English-speaking regions. Many of these shorthand systems later came to resemble ciphers when their explanatory texts were lost, blurring the distinction between stenography and cryptography. The DESCRYPT corpus therefore brings together a wide spectrum of enigmatic writings and encoded systems, including ciphertexts, cipher keys, and rare alphabets, whether solved, partially deciphered, or unsolved.

3.2. Repositories and datasets

Sources included in DESCRYPT are drawn from open access archaeological, philological, and linguistic research databases, museum and library catalogs, and similar authoritative repositories. As a result of the digital turn in the humanities, cataloging and curation of historical writings in recent decades have built a rich ecosystem of digital re-

[5] descrypt.org

Figure 1: Examples of rare scripts.

sources providing access to millions of records, representing historical and rare writing systems from all over the world (Hockey, 2008; Babeu, 2011). Included datasets are evaluated, documented, and mapped to internal metadata and transcription standards using transparent and fully retraceable procedures, and stored and reused in accordance with relevant licenses and permissions of the source repositories.

Datasets included in DESCRYPT already cover a diverse range of scripts illustrating the broader scope and potential of the resource. Current holdings include ancient inscriptions from Europe and the Middle East, among others cuneiform and related writing systems from Iraq, Syria, Iran and Turkey (Streck, 2010; Rattenborg et al., 2023), Linear B from the Eastern Mediterranean (Aurora, 2015), Rongorongo from the Easter Island, and Runic from Scandinavia (Institutionen för nordiska språk, 2020). Chronologically, records span from some of the earliest known writing systems to the early modern period. The database further contains a variety of shorthand systems and records employing artificial language schemes, predominantly European and dating to the Middle Ages and the early modern era, as well as extensive collections of historical ciphertexts assembled as part of the DECRYPT project within the DECODE database, now being further augmented by the DESCRYPT project.

The sources under study are exemplified in Table 1, which presents examples of ancient and modern historical writings and scripts, while Table 2 illustrates examples of encrypted texts, shorthand systems, and artificial language schemes.

3.3. Acquisition and Imaging

Where feasible, sources are linked to authoritative archival repositories; otherwise, the program undertakes high-resolution digitization following best practices in cultural heritage imaging. The pipeline emphasizes image quality and reproducibility to support downstream processing: dewarping, contrast enhancement, denoising, and illumination normalization are applied cautiously, with original images preserved. When available and relevant, 3D models and multispectral derivatives are associated with each record to aid symbol recovery on challenging surfaces.

Digitization relies on both high-resolution scanning and photography, ensuring that every record can serve as a reliable digital surrogate for scholarly analysis. Existing digital collections prepared by other research communities are not duplicated but are instead referenced and linked to the DESCRYPT infrastructure. Each source is annotated, catalogued, and integrated into the database according to the TEI standard, using a flexible schema capable of representing both ciphers and other rare or unknown writing systems.

The resulting, albeit still growing, corpus thus provides a searchable and standardized set of digital reproductions intended for systematic and comparative study.

3.4. Metadata and Contextualization

Given the global outlook of the DESCRYPT project, the collection and integration of source material requires extensive metadata to support linguistic, historical, and archaeological contextualization, as well as efforts at decipherment and translation. Ro-

Script	Region	Date	Status	Type	Language
Asia					
Cuneiform	Iraq, Syria, Turkey	3200 BCE–100 CE	Deciphered	Logo-syllabic	Sumerian, Akkadian
Proto-Elamite	Iran	3100–2900 BCE	Undeciphered	Logographic	Unknown
Linear Elamite	Iran	2300–1880 BCE	Part. dec.	Logo-syllabic	Elamite
Indus (Harappan)	Pakistan, India	2600–1900 BCE	Undeciphered	Logo-syllabic?	Unknown
Brahmi	India	200-300 BCE	Deciphered	Abugida	Prakrit, Sanskrit
Proto-Sinaitic	Sinai	1850–1550 BCE	Part. dec.	Abjad	Proto-Semitic
Ugaritic	Syria	1400–1200 BCE	Deciphered	Abjad	Ugaritic
Oracle Bone Script	China	1200-700 BCE	Deciphered	Logographic	Archaic Chinese
Khitan (Lg./Sm.)	N. China	920–1190 CE	Part. dec.	Logo-syllabic	Khitan
Tangut	W. China	1036–1227 CE	Largely dec.	Logographic	Tangut
Jurchen	Manchuria	1110–1400 CE	Part. dec.	Mixed	Jurchen
Bactrian/Sogdian var.	C. Asia	100–900 CE	Part. dec.	Alphabetic	Iranian
Africa					
Meroitic	Nubia	300 BCE–400 CE	Part. dec.	Abugida	Meroitic
Libyco-Berber (Tifinagh)	N. Africa	300 BCE–pres.	Part. dec.	Abjad	Berber
Nsibidi	Nigeria/Cam.	pre-1600 CE–pres.	Undeciphered	Proto-writing	Unknown
Europe					
Vinča symb.	Balkans	5300–4000 BCE	Undeciphered	Proto-writing	Unknown
Linear A	Crete	1800–1450 BCE	Undeciphered	Logo-syllabic	Unknown (Minoan)
Linear B	Greece/Crete	1450–1200 BCE	Deciphered	Logo-syllabic	Mycenaean Greek
Cypro-Minoan	Cyprus	1550–1050 BCE	Undeciphered	Syllabic	Unknown
Etruscan	Italy	700–100 BCE	Part. dec.	Alphabetic	Etruscan
Runic (Elder Futhark)	N. Europe	ca. 150–750 CE	Deciphered	Alphabetic	Proto-Norse
Americas					
Olmec (Cascajal)	Mexico	ca. 900 BCE	Undeciphered	Glyphic	Unknown
Zapotec	Oaxaca	500 BCE–500 CE	Part. dec.	Logo-syllabic	Zapotec
Epi-Olmec/Isthmian	S. Mexico	300 BCE–500 CE	Part. dec.	Logo-syllabic	Unknown (Mixe-Zoq.)
Maya (early)	Mesoam.	300 BCE–250 CE	Part. dec.	Logo-syllabic	Mayan
Rongorongo	Polynesia	1800–1900 CE	Undeciphered	Glyphic	Unknown

Table 1: Examples of deciphered, largely deciphered (Largely dec.) partially deciphered (Part. dec.), and undeciphered writing systems across major world regions.

bust metadata protocols are essential for integrating sources drawn from a wide range of writing traditions, documented according to highly diverse scholarly conventions and preserved with varying degrees of digital representation and accessibility.

In addition to automatically generated administrative fields, such as the record identifier, and creation and update timestamps, the metadata model comprises fields for file description, record identification, object description, origin and provenance, correspondence, content, and additional information. The *File description* records the person responsible for the entry, the collector, the record type (e.g., natural script, artificial script, or cipher), and the access mode. *Record identification* includes the name of the record, its current institution, city, country, and source, that is, the geographical and institutional location at which the inscribed object is currently held, for example a library or a museum, together with its source identifier, the URI in the source repository, and license information. *Object description* captures the type and material of the inscribed object, together with its measurements and, where relevant, hand and decoration. *Origin and provenance* fields record the place of origin, including city and region, provenance, provenance type, dating range, and historical or archaeological period. The *Content* field describes the inscription through script, language(s), genre, number of pages, and cipher-related features, while *correspondence* fields allow for the specification of author, sender, and recipient. Bibliography and notes along with acknowledgments are recorded separately as *additional information*.

Wherever possible, the schema normalizes description through controlled vocabularies aligned with established external authorities in linguistics, writing systems, and cultural heritage documentation, thereby strengthening data discovery, compatibility, reuse, and linked open data capability (Bizer et al., 2009; Berners-Lee, 2006; Palma and Megyesi, 2025). Record material, object type, and genre are mapped to the Getty Art and Architecture Thesaurus (Getty Research Institute, 2025); language and script identifiers follow ISO 15924 standards and RFC 5646-based categories and best-practice recommendations associated with the Internet Assigned Numbers Authority (IANA)[6]; and geographical and chronological references may be aligned with resources such as Geonames (Geonames), Pleiades (Pleiades), PeriodO, and Wikidata (Wikimedia Foundation, 2025). At the same time, the model retains free-text subtype and note fields in order to preserve legacy values and to accommodate cases in which controlled vocabularies do not

[6] https://www.iana.org

Category	Region	Date	Nature / Type
Historical ciphertexts			
Voynich manuscript	Central Europe?	15th c. CE	Undeciphered (cipher? language?)
Letters of Mary, Queen of Scots	Scotland/England/France	1586 CE	Homophonic substitution cipher
Papal diplomatic ciphers	Italy (Europe-wide use)	16th–17th c. CE	Homophonic and polyphonic substitution cipher
The Borg cipher	Northern Europe	17th c. CE	Simple substitution cipher
The Copiale cipher	Germany	ca. 1730 CE	Homophonic substitution cipher
Cipher keys / devices			
Alberti cipher disk	Italy	1467 CE	Polyalphabetic disk (rotating)
Cipher keys	Europe	14th–19th c. CE	Digit/symbol mappings for letters/words
Vigenère table	France	16th c. CE	Polyalphabetic key table
Jefferson disk	USA/France	1790s CE	Multi-alphabet cylinder
Ottoman cipher tables	Ottoman Empire	18th–19th c. CE	Bigram/word tables + numerals
Shorthand systems			
Tironian notes	Roman world	1st c. BCE	Latin shorthand symbols
Gabelsberger shorthand	Germany	1817 CE	Cursive phonetic shorthand
Pitman shorthand	United Kingdom	1837 CE	Phonetic, stroke thickness/position
Duployan shorthand	France/Canada	1860s CE	Curvilinear, widely adapted
Gregg shorthand	USA	1888 CE	Cursive, elliptical phonetic system
Japanese sokki (Waseda) shorthand	Japan	1917 CE	Mora-based, indigenous systems
Arabic shorthand (Haddad)	Egypt	1940s CE	Pitman-influenced, RTL
Artificial language schemes			
Marin Mersenne's La science universelle	France	1636 CE	Proposal for a universal science and signs
Francis Lodwick's Common Writing	England	1647 CE	Empirical–lexical language
John Wilkins' philosophical language	England	1668 CE	Taxonomic, a priori lexicon
Cave Beck' Pasigraphie	England	1657 CE	Universal writing, numerical codes for words
George Dalgarno's The Universal Character	Scotland, England	1661 CE	Symbolic writing for communication
G. W. Leibniz's Characteristica universalis	Germany	1666-1700 CE	Universal symbolic language
Solresol	France	1820s CE	Do-re-mi based, universalist

Table 2: Examples of historical ciphertexts, cipher keys/devices, shorthand systems, and artificial language schemes across regions and time periods.

provide sufficient granularity. The schema therefore combines standardization and interoperability with the flexibility required for heterogeneous documentary corpora.

Schemas for rendering records in TEI- and EpiDoc-compatible XML (Consortium, 2023; Elliotte et al., 2006–2022) are tailored to include linked metadata (Gaitanou et al., 2022) on object type, script, material, archaeological and historical context, chronological and geographical indicators of use, relations to other pertinent sources, and decipherment status. Each entry may also include separate components for the associated script and language, facilitating cross-referencing within and beyond the database. The metadata scheme is illustrated in Figure 2. This structured descriptive framework forms the basis for subsequent computational processing and collaborative annotation, while also facilitating structured export and comparative analysis.

4. The DECODE Database

To make the data openly accessible, we created a database that supports search, editing, and the upload of new records.

4.1. Design Principles

The DECODE database is a public, extensible repository that unifies *images*, *transcriptions*, *metadata*, *derived annotations*, and *processing artifacts* (segmentations, symbol clusters, model outputs) within a coherent, standards-based framework.

It evolves the DECRYPT infrastructure (Megyesi et al., 2024a) to accommodate a wider typology of writing systems and integrates language resources and tools for analysis.

The database is built from standard components, including a PHP application backed by a relational database, quasi-standard API interface with the *swagger* framework, mass input modules for UX, the API and from CSV format. The experience derived from previous versions is that it is best to provide several paths for data entry, in small and large quantities, as different users have different working methods and preferences. Four principles guide the design:

1. **Interoperability**: TEI/EpiDoc-compliant metadata structure; persistent identifiers; Geonames Open Data value sets and schema alignment to external catalogs; role-based authorization system built on standard components; and a RESTful API for programmatic access.

2. **Provenance and Traceability**: Records linking raw images, pre-processing steps, potential annotations and model inferences; full audit trails of human and programmatic edits and access; clear licensing and rights; metadata.

3. **Data Generalization**: Support for connecting with AI deployments that can train and evaluate models under low-resource constraints

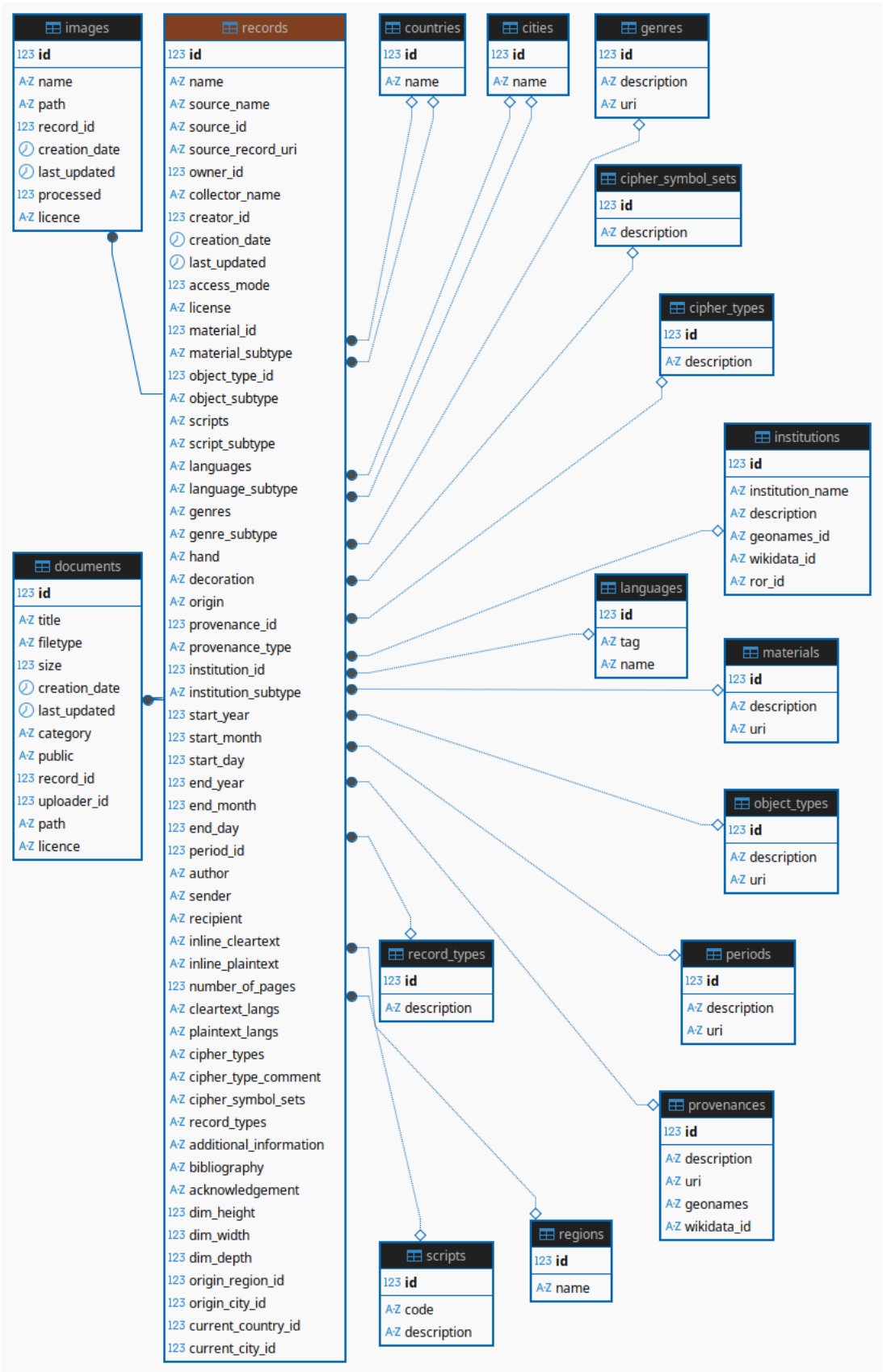


Figure 2: Metadata structure.

(few-shot settings), with benchmark splits for reproducible comparison and shared tasks.

4. **Scholarly Control**: Connectivity to expert-in-the-loop tooling to review, correct, and approve transcriptions and interpretations; mechanisms to represent uncertainty and competing hypotheses. Extensive user testing guide the development and integration of tools, ensuring their effectiveness in real-world applications.

4.2. Content and Schema

Each database record encapsulates:

- **Core Object**: High-resolution images (and, if available, 3D/multispectral derivatives), stable identifiers, source repository links, and rights statements.
- **Descriptive Metadata**: Object type, material, condition; script classification; chronology and provenance; associated persons/places in line with well-known Linked Open Data resources; and bibliographic references.
- **Textual Data**: Diplomatic transcription aligned to images and layout (blocks, lines, regions), with symbol-level segmentation and clustering, plus normalized representations if applicable.
- **Analytical Layers**: Language identification hypotheses; frequency profiles; co-occurrence statistics; cryptanalytic views for encoded sources; alignment to dictionaries or historical language models; and machine-inferred suggestions with confidence scores.
- **Relations and Collections**: Links across items by hand, inscription series, archive, or cipher family; curated sets for teaching, benchmarking, or shared tasks.

Figure 3 shows the current version of the database interface, including several records and selected metadata fields.

5. Collaboration Across Disciplines

The DESCRYPT project is inherently interdisciplinary, bringing together historians, philologists, linguists, cryptologists, computer scientists, and specialists in computer vision and manuscript studies in the joint development of the corpus and database. This combination of expertise ensures that the resource is historically grounded, methodologically robust, and relevant to multiple research communities working with rare writings.

An important objective is for the corpus to function as a *monitor corpus*: a living and continuously expanding resource that can be updated as new material is identified, additional images and descriptions become available, and existing analyses are revised. Because many rare writings survive only in fragmentary, dispersed, or otherwise unique forms, a static corpus would not adequately capture the changing state of documentation and interpretation. The corpus is therefore designed to support continuous enrichment and revision over time.

The database is intended to make this collaborative model operational by allowing domain experts to upload new records, revise existing entries, improve transcriptions and metadata, and contribute alternative readings or interpretations of rare writings and other unique materials. In this way, the platform serves not only as a repository, but also as a shared research environment that supports scholarly exchange and expert-led curation. Such collaboration is essential for possible decipherment of complex historical sources, where progress depends on combining domain-specific knowledge with computational, linguistic, and imaging methodologies (Láng and Megyesi, 2024).

6. Long Term Aims

The long-term goal of DESCRYPT is to develop the corpus and database into a mature research infrastructure for the study of rare, non-standard, and undeciphered writings. Beyond collecting and organizing material, the project aims to support the full research cycle from data acquisition and curation to analysis, interpretation, and reuse. Particular emphasis will be placed on extending coverage, refining data standards for poorly encoded writing systems, and enabling reproducible work across heterogeneous source types.

A central objective is to establish practical transcription and representation conventions for scripts that lack stable editorial or encoding standards. This includes Unicode-compliant solutions where possible, provisional encodings for otherwise unsupported symbols, and explicit links between image-level observations and abstract symbol inventories. Such conventions will make it easier to compare sources across traditions and to create reusable ground truth for computational analysis.

Another long-term aim is to connect the database more directly to analytical workflows. Future development will integrate services for layout analysis, symbol segmentation, transcription support, and exploratory decipherment, allowing researchers to move more efficiently from digitized objects to interpretable textual data. Language resources such as dictionaries, sign inventories, and historical corpora will be linked where relevant, while encoded materials will benefit from cryptanalytic functionality tailored to historical sources.

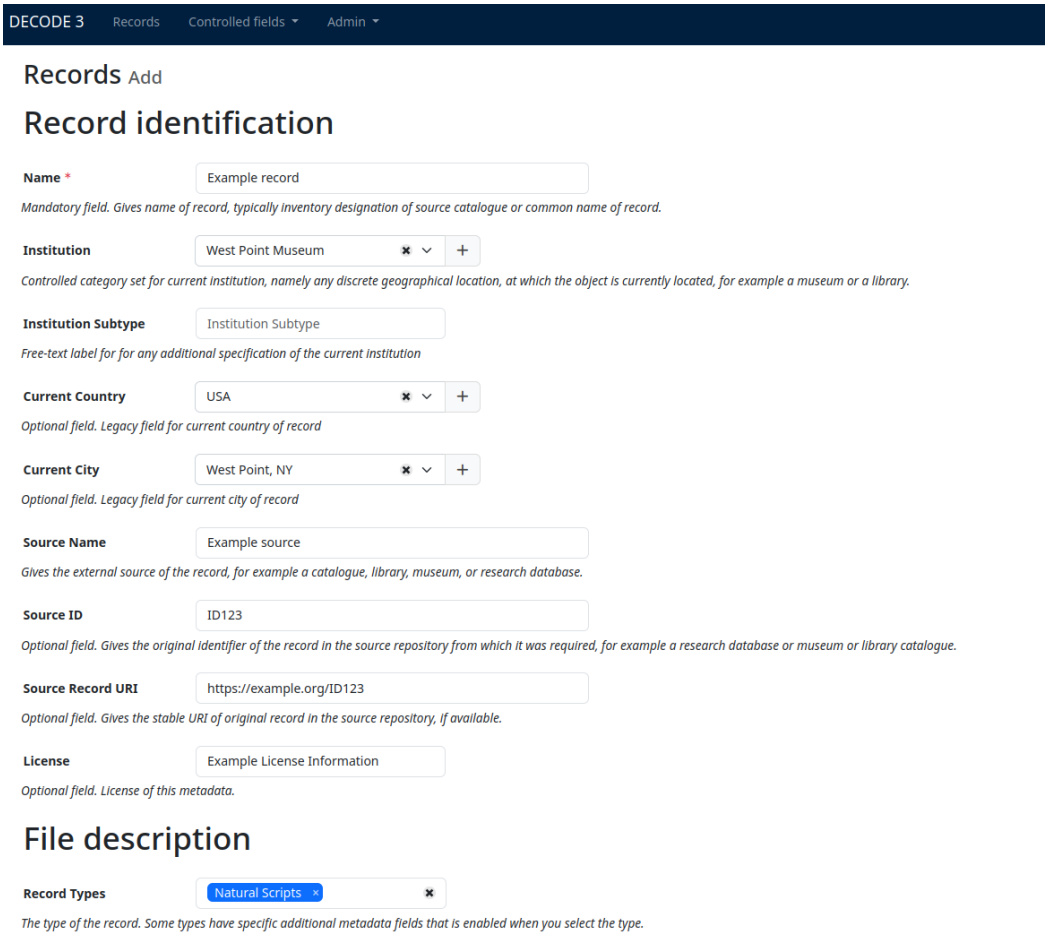


Figure 3: A screenshot of the DECODE database.

DESCRYPT is also intended to grow as a collaborative scholarly environment. The platform will support continued expert contribution through the addition of new records, revision of existing entries, and comparison of alternative readings and hypotheses. In parallel, the project aims to release stable data snapshots, benchmarks, and evaluation settings that can support transparent comparison of methods and encourage broader reuse within language technology and digital humanities.

In the longer term, DESCRYPT seeks to strengthen the visibility of rare writings within digital scholarship by making them more accessible as structured, analyzable, and reusable data.

7. Conclusion

This paper has introduced the design and implementation of a new corpus and database dedicated to rare and undeciphered scripts. The resource unifies heterogeneous materials—high-quality images, transcriptions, annotations, and metadata—within a standards-based infrastructure built for interoperability, reproducibility, and long-term preservation.

By combining established models from linguistic and cultural heritage data management with flexible solutions for low-resource and non-standard scripts, the database provides a foundation for systematic research across philology, linguistics, cryptology, and computer science. It supports both humanistic and computational workflows, enabling documentation, comparison, and analysis of writing systems that have long remained beyond the reach of large-scale digital methods.

Beyond corpus construction, the DESCRYPT project also aims to develop generic, data-lean pipelines for image processing, transcription, and symbol analysis, demonstrating how FAIR data principles and expert-in-the-loop methodologies can be applied to rare and complex scripts. Future work will expand coverage to underrepresented writing systems, enhance automated layout and symbol recognition, and deepen integration with infrastructures such as CLARIN and ELRA.

Ultimately, the initiative aims to transform rare scripts from isolated artifacts into analyzable linguistic data bridging cultural heritage preservation with computational approaches to the study of humanity's written past.

8. Ethical considerations

Building and sharing a corpus and database for rare and undeciphered scripts raises important ethical questions concerning cultural heritage, data stewardship, and responsible computational use. Many of the materials included in the resource originate from diverse cultural, religious, and political contexts, and some may carry particular significance for source communities or holding institutions. Digitization, description, and dissemination must therefore be undertaken with attention to provenance, attribution, rights, and, where relevant, appropriate restrictions on access.

A further ethical consideration concerns data ownership and reuse. Although the underlying sources are often historical, their digital reproductions, metadata, and annotations may involve new layers of intellectual contribution and institutional responsibility. DESCRYPT therefore seeks to balance openness and reuse with respect for licenses, archival agreements, and the interests of partner collections.

Responsible use of computational methods is equally essential. Automated analysis can support the study of rare writings, but it should not obscure uncertainty or replace expert judgement. For this reason, the project emphasizes transparency, traceability, and human oversight in transcription, annotation, and interpretation, so that computational outputs remain open to review, correction, and scholarly debate.

9. Limitations

The present resource remains a work in progress, and its coverage is necessarily uneven. Rare and undeciphered writings are often fragmentary, geographically dispersed, difficult to access, and documented according to highly variable scholarly conventions. As a result, some traditions are currently better represented than others, and the completeness and granularity of metadata, images, and transcriptions may vary across records.

A further limitation concerns standardization. Many of the scripts and symbol systems addressed in the project lack stable encoding conventions, established editorial practices, or agreed sign inventories. In such cases, provisional representations and working classifications are unavoidable, which may complicate comparison across sources and require later revision as scholarship advances.

The applicability of computational methods is also constrained by the nature of the material. Limited data availability, damaged or degraded surfaces, uncertain symbol boundaries, and unresolved linguistic affiliation all reduce the reliability of automatic approaches and make expert intervention indispensable. DESCRYPT therefore adopts an *AI-in-the-loop* approach, in which AI functions as an assistive tool within an expert-driven workflow: the scholar remains the principal agent of interpretation, while computational methods support analysis and hypothesis generation rather than driving the process and being corrected post hoc by the expert.

Finally, long-term sustainability depends on continued maintenance, technical development, and institutional support. Preserving the data, updating the platform, and adapting tools and workflows to changing research and infrastructure needs will require ongoing resources and active engagement from the wider scholarly community.

10. Acknowledgements

This work has been supported by Riksbankens Jubileumsfond, grant M24-0028: *Echoes of History: Analysis and Decipherment of Historical Writings* (DESCRYPT).

11. Bibliographical References

2024. Fabricius. Retrieved 2025-10-06.

2024. Transcribus. Retrieved 2025-10-07.

2025. eScriptorium. Retrieved 2025-10-06.

Nada Aldarrab. 2017. Decipherment of historical manuscripts. Master's thesis, University of Southern California.

Eugen Antal and Pavol Zajac. 2020. Hcportal overview. In *Proceedings of the 3rd International Conference on Historical Cryptology, HistoCrypt 2020*, pages 18–20. Linköping Electronic Press.

Yannis Assael, Thea Sommerschield, Brendan Shillingford, Mahyar Bordbar, John Pavlopoulos, Marita Chatzipanagiotou, Ion Androutsopoulos, Jonathan Prag, and Nando de Freitas. 2022. Restoring and attributing ancient texts using deep neural networks. *Nature*, 603:280–283.

Ali Assarpour and Kent D. Boklan. 2010. How we broke the union code (148 years too late). *Cryptologia*, 34(3):200–210.

Federico Aurora. 2015. Damos (database of mycenaean at oslo). annotating a fragmentarily attested language. In Pedro A. Fuertes-Olivera et al., editor, *Current Work in Corpus Linguistics: Working with Traditionally-conceived Corpora and Beyond*, volume 198 of *Procedia - Social and Behavioral Sciences*, pages 21–31.

Alessandra Avanzini. 2009. Origin and classification of the ancient south arabian languages. *Journal of Semitic Studies*, 54(1):205–220.

Alison Babeu. 2011. *Rome Wasn't Digitized in a Day: Building a Cyberinfrastructure for Digital Classicists*. Council of Library Information Resources.

Arnau Baró, Jialuo Chen, Alicia Fornés, and Beáta Megyesi. 2019. Towards a generic unsupervised method for transcription of encoded manuscripts. In *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage (DATeCH2019)*, Brussels, Belgium.

Tim Berners-Lee. 2006. Linked data. `https://www.w3.org/DesignIssues/LinkedData.html`. W3C Design Issues.

Christian Bizer, Tom Heath, and Tim Berners-Lee. 2009. Linked data: The story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5:1–22.

Svenja Bonmann, Jakob Halfmann, Natalie Korobzow, and Bobomullo Bobomulloev. 2023. A partial decipherment of the unknown kushan script. *Transactions of the Philological Society*, 121:293–329.

John Chadwick. 1958. *The Decipherment of Linear B*. Cambridge University Press.

Jean-Francois Champollion. 1822. *Lettre à M. Dacier relative à l'alphabet des hiéroglyphes phonétiques*.

Jialuo Chen, Mohamed Ali Souibgui, Alicia Fornés, and Beáta Megyesi. 2020. A web-based interactive transcription tool for encrypted manuscripts. In *Proceedings of the 3rd International Conference on Historical Cryptology, HistoCrypt 2020*, pages 52–59. Linköping Electronic Press.

Michael D. Coe. 2011. *Breaking the Maya Code*, 3rd edition. Thames & Hudson Ltd.

The TEI Consortium. 2023. *Guidelines for Electronic Text Encoding and Interchange P5 Version 4.7.0*. Last updated on 16th November 2023.

Shou de Lin and Kevin Knight. 2006. Discovering the linear writing order of a two-dimensional ancient hieroglyphic script. *Artificial Intelligence*, 170(4-5).

Francois Desset, Kambiz Tabibzadeh, Matthieu Kervran, Gian Pietro Basello, and Gianni Marchesi. 2022. The decipherment of Linear Elamite writing. *Zeitschrift für Assyriologie und vorderasiatische Archäologie*, 112(1):11–60.

Qing Dou and Kevin Knight. 2012. Large scale decipherment for out-of-domain machine translation. In *Empirical Methods in Natural Language Processing*.

Roger L. Easton and Keith Knox. 2016. Spectral imaging of manuscripts: Recovery of the past and preservation for the future. *Archiving Conference*, 2016(1):1–1.

Tom Elliotte, Gabriel Bodard, and Hugh Cayless. 2006–2022. Epidoc: Epigraphic documents in tei xml. Online material.

Bernhard Esslinger. 2024. *Learning and Experiencing Cryptography with CrypTool and SageMath*. Artech House, Norwood. Https://us.artechhouse.com/Learning-and-Experiencing-Cryptography-with-CrypTool-and-SageMath-P2378.aspx.

Silvia Ferrara and Fabio Tamburini. 2022. Advanced techniques for the decipherment of ancient scripts. *Lingue e linguaggio*, XXI(2):239–259.

Ethan Fetaya, Yonatan Lifshitz, Elad Aaron, and Shai Gordin. 2020. Restoration of fragmentary babylonian texts using recurrent neural networks. *Proceedings of the National Academy of Sciences*, 117(37):22743–22751.

Alica Fornés, Beáta Megyesi, and Joan Mas. 2017. Transcription of encoded manuscripts with image processing techniques. In *Proceedings of Digital Humanities*, Montreal, Canada.

Panorea Gaitanou, Ioanna Andreou, Miguel-Ángel Sicilia, and Emmanouel Garoufallou. 2022. Linked data for libraries: Creating a global knowledge space, a systematic literature review. *Journal of Information Science*, 50:204 – 244.

Geonames. GeoNames. Accessed: 2025-10-10.

Andrew George and Manfred Krebernik. 2022. Two remarkable vocabularies: Amorite-Akkadian bilinguals! *Revue d'assyriologie et d'archéologie orientale*, 116:113–166.

Getty Research Institute. 2025. Art & architecture thesaurus (AAT). Accessed: 2025-10-07.

Bradley Hauer and Grzegorz Kondrak. 2016. Decoding anagrammed texts written in an unknown language and script. In *Transactions of the Association for Computational Linguistics*, volume 4, pages 75–86. MIT Press.

Mihály Héder and Beáta Megyesi. 2022. The decode database of historical ciphers and keys: Version 2. In *Proceedings of the 5th International Conference on Historical Cryptology, HistoCrypt 2022*.

Judith Hochberg, Patrick Kelly, Timothy Thomas, and Lila Kerns. 1997. Automatic script identification from document images using cluster-based templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2):176–181.

Susan Hockey. 2008. The History of Humanities Computing. In Susan Schreibman, Ray Siemens, and John Unsworth, editors, *A Companion to Digital Humanities*, pages 3–19. Blackwell, Oxford & Malden, MA.

Uppsala universitet Institutionen för nordiska språk. 2020. Samnordisk runtextdatabas. http://www.nordiska.uu.se/forskn/samnord.htm. Accessed at 31 March 2026.

David Kahn. 1967. *The Codebreakers*, 2nd edition. New York.

Kevin Knight, Beáta Megyesi, and Christiane Schaefer. 2012. The copiale cipher. In *ACL Workshop on Building and Using Comparable Corpora (BUCC)*.

Nils Kopal. 2019. Cryptanalysis of homophonic substitution ciphers using simulated annealing with fixed temperature. In *Proceedings of the 2nd International Conference on Historical Cryptology, HistoCrypt 2019*, Mons, Belgium. NEALT Proceedings Series 37, Linköping Electronic Press.

Nils Kopal and Michelle Waldispühl. 2022. Deciphering three diplomatic letters sent by maximilian ii in 1575. *Cryptologia*, 46(2):103–127.

Benedek Láng and Beáta Megyesi. 2024. An STS analysis of a digital humanities collaboration: Trading zones, boundary objects, and interactional expertise in the DECRYPT project. *Humanities and Social Sciences Communications*, 11:618.

George Lasry. 2018. A methodology for the cryptanalysis of classical ciphers with search metaheuristics.

George Lasry, Norbert Biermann, and Satoshi Tomokiyo. 2023. Deciphering Mary Stuart's lost letters from 1578–1584. *Cryptologia*.

George Lasry, Beáta Megyesi, and Nils Kopal. 2020. Deciphering papal ciphers from the 16th to the 18th century. *Cryptologia*, pages 479–540.

Ernst Leierzopf, Nils Kopal, Bernhard Esslinger, Harald Lampesberger, and Eckehard Hermann. 2021a. A massive machine-learning approach for classical cipher type detection using feature engineering.

Ernst Leierzopf, Vasily Mikhalev, Nils Kopal, Bernhard Esslinger, Harald Lampesberger, and Eckehard Hermann. 2021b. Detection of classical cipher types with feature-learning approaches.

Peter Liddel. 2017. Greek inscriptions: insights and resources in the classroom and beyond. *Journal of Classics Teaching*.

Francisco Beltrán Lloris. 2015. The epigraphic habit in the roman world. In Ch. Bruun and J. Edmonson, editors, *The Oxford Handbook of Roman Epigraphy*, pages 131–148. Oxford University Press, Oxford—New York.

Jiaming Luo, Frederik Hartmann, Enrico Santus, Regina Barzilay, and Yuan Cao. 2021. Deciphering undersegmented ancient scripts using phonetic prior. volume 9, pages 69–81.

Beáta Megyesi, Nils Blomqvist, and Eva Pettersson. 2019. The DECODE database: Collection of historical ciphers and keys.

Beáta Megyesi, Bernhard Esslinger, Alicia Fornés, Nils Kopal, Benedek Láng, Georg Lasry, Karl de Leeuw, Eva Pettersson, Arno Wacker, and Michelle Waldispühl. 2020. Decryption of historical manuscripts: the decrypt project. *Cryptologia*.

Beáta Megyesi, Alicia Fornés, Nils Kopal, Benedek Láng, Michelle Waldispühl, Vasily Mikhalev, and Bernhard Esslinger. 2024a. Historical cryptology.

Beáta Megyesi, Crina Tudor, Benedek Láng, Anna Lehofer, Nils Kopal, Karl de Leeuw, and Michelle Waldispühl. 2024b. Keys with nomenclatures in the early modern europe. *Cryptologia*, 48(2):97–139.

Cosimo Palma and Beáta Megyesi. 2025. DECODE2LOD: Connecting the DECODE Database with the Linked Open Data Cloud. In *HistoCrypt 2025 - The International Conference on Historical Cryptology*, Poznań.

Katerina Papavassileiou, Dimitrios Kosmopoulos, and Gareth Owens. 2023. A generative model for the mycenaean linear b script and its application in infilling text from ancient tablets. *Journal on Computing and Cultural Heritage*, 16.

Eva Pettersson and Beáta Megyesi. 2018. The histcorp collection of historical corpora and resources. In *Proceedings of the Third Conference on Digital Humanities in the Nordic Countries*.

Pleiades. Pleiades: A Gazetteer of Past Places. Accessed: 2025-10-10.

Rune Rattenborg, Gustav Ryberg Smidt, Carolin Johansson, Nils Melin-Kronsell, and Seraina Nett. 2023. The archaeological distribution of the cuneiform corpus. *Altorientalische Forschungen*, 50(2):178–205.

Sujith Ravi and Kevin Knight. 2011a. Bayesian inference for Zodiac and other homophonic ciphers. In *Proceedings of the 49th Annual Meeting of the ACL: Human Language Technologies (ACL-HLT 2011)*, pages 352–360, Portland, Oregon, USA. ACL.

Sujith Ravi and Kevin Knight. 2011b. Deciphering foreign language. In *Proceedings of the 49th Annual Meeting of ACL: Human Language Technologies (ACL-HLT 2011)*, pages 12–21, Portland, Oregon, USA. ACL.

READ-COOP SCE. 2024. Transkribus. Accessed: 2025-10-06.

Sravana Reddy and Kevin Knight. 2011. What we know about the voynich manuscript. In *ACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*.

Sravana Reddy and Kevin Knight. 2012. Decoding running key ciphers. *ACL*.

Annamaria De Santis and Irene Rossi. 2019. Crossing experiences in digital epigraphy: From practice to discipline. *De Gruyter*.

Benjamin Snyder, Regina Barzilay, and Kevin Knight. 2010. A statistical model for lost language decipherment. *ACL*.

Isabel Velázquez Soriano and David Espinosa Espinosa. 2021. *Epigraphy in the Digital Age. Opportunities and Challenges in the Recording, Analysis and Dissemination of Inscriptions*. Archaeopress Archaeology, Oxford.

Mohamed Ali Souibgui, Asma Bensalah, Jialuo Chen, Alicia Fornés, and Michelle Waldispühl. 2022. A user perspective on HTR methods for the automatic transcription of rare scripts. the case of Codex Runicus. *ACM Journal on Computing and Cultural Heritage*.

Michael P. Streck. 2010. Großes Fach Altorientalistik: Der Umfang des keilschriftlichen Textkorpus. *Mitteilungen der Deutschen Orient-Gesellschaft*, 142:35–58.

Ferenc Szigeti and Mihály Héder. 2022. The transcript tool for historical ciphers by the decrypt project. In *Proceedings of the 5th International Conference on Historical Cryptology*, pages 208–211.

Melissa Terras. 2012. Image processing and digital humanities. In N. M. Terras, J. Nyhan, and C. Warwich, editors, *Digital Humanities in Practice*. Facet Publishing.

Melissa Terras. 2016. *Crowdsourcing in the Digital Humanities*. Wiley-Blackwell.

The Unicode Consortium. 2019. *The Unicode® Standard, Version 12.0 – Core Specification*. The Unicode Consortium, Mountain View, CA.

Thesaurus Linguae Aegyptiae. 2025. Thesaurus Linguae Aegyptiae. Accessed: 2025-10-08.

Trismegistos. 2025. Trismegistos. https://www-trismegistos-org. Accessed: 2025-10-08.

Wikimedia Foundation. 2025. Wikidata: A free and open knowledge base. Accessed: 2025-10-07.

Henrik Williams, Marco Bianchi, and Christiane Zimmermann. 2022. Corpus editions of runic inscriptions in supranational databases. *Futhark: International Journal of Runic Studies*, 12:117–135.

Xusen Yin, Nada Aldarrab, Beáta Megyesi, and Kevin Knight. 2019. Decipherment of historical manuscript images. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 78–85, Sydney, NSW, Australia.

12. Language Resource References

Federico Aurora. 2015. DAMOS (Database of Mycenaean at Oslo). Annotating a fragmentarily attested language. In Pedro A. Fuertes-Olivera et al., editor, *Current Work in Corpus Linguistics: Working with Traditionally-conceived Corpora and Beyond*, volume 198 of *Procedia - Social and Behavioral Sciences*, pages 21–31.

Uppsala universitet Institutionen för nordiska språk. 2020. Samnordisk runtextdatabas. `http://www.nordiska.uu.se/forskn/samnord.htm`. Accessed at 31 March 2026.

Beáta Megyesi and Mihály Héder and Benedek Láng. 2024. *The DECODE database: A corpus of historical ciphertexts and cipher keys*. Department of Linguistics, Stockholm University. PID de-crypt.org. Retrieved 2025-10-06.

Eva Pettersson and Beáta Megyesi. 2024. *HistCorp: Historical Corpora*. Department of Linguistics and Philology. PID https://www2.lingfil.uu.se/person/pettersson/histcorp/. Retrieved 2025-10-06.

A Layered Annotation Workflow for Semitic Epigraphy

Tal Bernstein[1], Shai Gordin[2,3], Letizia Cerqueglini[1]

[1] Tel Aviv University, [2] Ariel University

[1] The Department of Hebrew Language and Semitic Linguistics, Tel Aviv University, Tel Aviv 69978, Israel

[2] Department of Land of Israel and Archaeology, Ariel University, Ariel 40700, Israel

[3] Digital Humanities and Social Sciences Hub, Open University of Israel, Ra'anana, Israel

talbernstein@mail.tau.ac.il, shaigo@ariel.ac.il, cerqueglini@tauex.tau.ac.il

Abstract

This paper presents a layered annotation workflow for the historical linguistic study of Semitic epigraphic texts. Using a curated Phoenician corpus primarily based on *Kanaanäische und Aramäische Inschriften* (KAI), the system models inscriptions as multi-layered objects that encode graphemic, morphosyntactic, phonological, semantic, and contextual information as independently queryable layers. Annotation is embedded in a structured editorial workflow supporting peer review, expert validation, version tracking, and the representation of variant readings and uncertainty. A case study demonstrates how recurring formulaic constructions can be modeled as morphosyntactic configurations retrievable across inscriptions. Although the current corpus is limited in scope, the data model is language-agnostic, designed for extension to other Semitic epigraphic traditions.

Keywords: Layered annotation, Semitic epigraphy, historical linguistics, collaborative annotation, linguistic corpora, Phoenician inscriptions

1. Introduction

1.1 Background and Motivation

Over the past two decades, research on ancient languages has increasingly relied on digital corpora and language technologies. In Semitic epigraphy, this has led to the digitization of inscriptions and the development of online repositories, such as DASI (Digital Archive for the Study of Pre-Islamic Arabian Inscriptions) and OCIANA (Online Corpus of the Inscriptions of Ancient North Arabia), which provide large-scale access to epigraphic data with rich metadata (Avanzini, Prioletta & Rossi 2014; Al-Jallad 2015).

Epigraphic research nevertheless faces challenges such as fragmentary preservation, orthographic variation, and diachronic and diatopic diversity. Whereas many digital resources offer high-quality editions and metadata, fewer support fine-grained linguistic annotation within collaborative editorial workflows. These limitations are particularly evident in the Phoenician and Punic corpus, whose wide geographical and chronological spread calls for tools capable of representing both structural regularities and local variation in a systematic and reusable way.

1.2 Research Problem

How should the linguistic complexity of Semitic epigraphic texts be represented in a digital environment, to support historical linguistics research while remaining philologically responsible? Classical scholarship has long identified recurring formulaic structures and morphosyntactic patterns in Phoenician and Punic inscriptions, as well as in related Northwest Semitic languages such as Moabite, Ammonite, and Ekronite, often interpreted as a continuum of closely related Canaanite varieties (Koller 2013; Sanders 2008). However, such analyses are typically expressed in descriptive form and remain difficult to reproduce or compare systematically across corpora.

From a computational perspective, the challenge goes beyond digitizing inscriptions to designing an annotation model and workflow that encodes multiple layers of linguistic interpretation without flattening variation or obscuring editorial uncertainty. This includes supporting and preserving alternative readings, and enabling structured comparison across inscriptions distributed through time and space. It also calls for distinguishing between a *diplomatic representation*, which reflects the inscription as written, with line breaks, orthographic variation, damage, and restorations, and the *normalized representation*, which reflects linguistic interpretation, a standardized orthography, morphological segmentation, and lexical identification. Addressing such a problem requires an infrastructure in which linguistic annotation is tightly integrated with collaborative editorial processes.

1.3 Contribution of This Work

This paper presents a layered linguistic

annotation (LLA) workflow and digital infrastructure for Semitic epigraphy. It introduces a multi-layer annotation model grounded in Semitic linguistic principles, embeds it in a structured editorial workflow with peer review and expert approval, and demonstrates its analytical potential through a case study of Phoenician formulaic morphosyntax. Rather than aiming at full automation, such a workflow supports expert-driven scholarship through structured encoding and comparison of linguistic interpretations, and is extensible to other Semitic epigraphic traditions with comparable characteristics.

1.4 Structure of the Paper

The paper is structured as follows: Section 2 reviews related work. Section 3 describes the corpus. Section 4 presents the annotation model and workflow. Section 5 outlines the implementation. Section 6 presents a linguistic case study. Section 7 discusses limitations and future directions, followed by a brief conclusion.

2. Related Work

2.1 Annotation Frameworks for Historical and Ancient Languages

Digital corpora of historical languages increasingly rely on structured annotation frameworks to support linguistic analysis and computational reuse. TEI-based models and domain-specific standards such as EpiDoc alongside graph-oriented frameworks like LAF/GrAF, enable multi-level encoding of textual, linguistic, and contextual information (Bodard & Stoyanov 2016; Ide & Suderman 2014). Yet annotation practices remain largely project-specific, shaped by individual research goals and source materials rather than shared conventions (De Santis & Rossi 2018).

In parallel to XML and graph-based annotation frameworks, recent work has explored RDF and Linked Data approaches for representing LLA and interlinked textual resources. The Open Annotation Data Model provides a general framework for representing annotations as web resources, while models such as POWLA demonstrate how multi-layer linguistic corpora can be represented in RDF and Linked Data approaches for linguistic annotation (Sanderson et al. 2017; Chiarcos 2012; Chiarcos et al. 2012). These approaches emphasize interoperability, dataset linking, and flexible representation of annotation layers.

This tension is particularly evident in epigraphic corpora. In inscriptions such as the Phoenician Ahiram sarcophagus or Eshmunazar texts, fragmentary preservations, restored letters and competing interpretations[1] require concurrent representation of diplomatic readings, normalized forms and morphosyntactic analysis. Although existing standards allow such layered encoding, they do not inherently provide an integrated editorial workflow in which variant readings, certainty levels and linguistic interpretation remain dynamically interlinked and collaboratively validated. The challenge, therefore, is not only structural representation, but the coordination of LLA with transparent scholarly decision-making.

2.2 Digital Epigraphic Corpora and Tools

Digital epigraphic corpora have substantially improved the preservation, accessibility, and dissemination of inscribed materials by combining textual editions with structured metadata and visual documentation. Such platforms enable browsing, cataloguing, and basic textual search, and play an important role in safeguarding endangered cultural heritage (Vitale et al. 2021). However, their development is often constrained by funding cycles, software obsolescence, and long-term maintenance challenges, motivating infrastructure initiatives focused on persistent identifiers and archival solutions (Derntl et al. 2023). These efforts address sustainability and access, but many epigraphic corpora remain oriented toward documentation rather than toward fine-grained linguistic annotation or iterative scholarly interpretation.

2.3 Semitic and Northwest Semitic Digital Resources

Numerous Semitic corpora, both epigraphic and manuscript-based, have been digitized through libraries- and academic institutions-led projects. Codex and literary traditions in Arabic, Aramaic, Ge'ez and Hebrew are represented in major initiatives such as e.g., Arabic Collections Online (ACO; NYU Libraries), Corpus Coranicum

[1] The final two words of the Ahiram sarcophagus inscription (KAI 1) are read by Donner & Röllig (2002) as לפן גבל (LPN GBL, "before Byblos"), whereas Avishur (1979), unresolveably reads לפפ שבל (LPP ŠBL), noting that "this phrase is difficult and there are no comments worth mentioning". Others have suggested an alternative לעד שרל (L ʿD ŠRL), reflecting ongoing uncertainty at both graphemic, lexical and semantic levels.

(Berlin-Brandenburg Academy of Sciences and Humanities), the Ktiv project of the National Library of Israel, and syri.ac: An Annotated Bibliography of Syriac Resources Online. Cuneiform texts have also been made available through large digital platforms such as the Cuneiform Digital Library Initiative (CDLI), the Open Richly Annotated Cuneiform Corpus (ORACC), the Ebla Digital Archives (EbDA), and the Amarna Letters project in Akkadian in the Eastern Mediterranean World (AEMW/Amarna, Lauinger & Yoder 2014). These resources are typically organized around specific languages, regions, scripts, or textual traditions, reflecting domain-specific priorities. Even though coverage varies, most corpora emphasize cataloguing and access rather than fine-grained linguistic annotation.

2.4 Positioning the Present Workflow

The proposed workflow builds on existing annotation and epigraphic frameworks while addressing their limitations for linguistic analysis. It adopts a modular, multi-layered model tailored to Semitic inscriptions, enabling the representation of graphemic, morphosyntactic, semantic, and contextual information as interrelated yet independently queryable layers. Unlike many corpora that prioritize static editions, this workflow integrates linguistic annotation with a structured editorial process, including peer review, expert validation, and transparent versioning. Annotations retain variant readings and degrees of certainty, preserving interpretive flexibility and philological rigor. The resulting data support structured querying and export to standard formats (e.g., EpiDoc, LAF), enabling interoperability and comparative synchronic and diachronic research.

3. Corpus Overview

3.1 Linguistic Scope

The initial phase of the platform's corpus centers on Phoenician inscriptions from the Levantine "Motherland" (Byblos, Sidon, Tyre, and surrounding regions), drawing from the canonical selection in *Kanaanäische und Aramäische Inschriften* (KAI, Donner & Röllig 2002). Though selective, KAI serves as a backbone for Northwest Semitic epigraphy, offering historically and linguistically significant texts. The platform demonstrates the annotation and editorial workflow through diverse literary types, including royal inscriptions, temple dedications, funerary stelae, tariffs, and bilingual texts[2] on stone, metal, ostraca, and pottery.

3.2 Epigraphic Scope and Sources

Each inscription in the platform is encoded as an ordered array of lines, segmented into tokens and subdivided into graphemic segments, representing clitics, stems, suffixes, and inflectional morphemes. Linguistic annotations are stored in separate queryable layers for lemma, root, part of speech, morphology, phonology, and context-sensitive glosses. Transcriptions follow published editions and support both diplomatic and normalized readings (see Section 1.2), in accordance with established editorial practices[3].

3.3 Digital Preparation and Metadata Model

Metadata fields include inscription name and variants, language, dialect, script, provenance and current locations (with GIS coordinates), archaeological dating, discovery date and legacy sigla (e.g., KAI, CIS). Bibliographic references are linked to online copies, when available, and visual media is normally derived from Wikimedia Commons files. The easily customizable model supports object classification details and physical dimensions, following best practices from Trismegistos (Depauw & Gheldof 2014) and DASI projects.

The platform uses a standard Latin-based transliteration system alongside full Unicode display of the original script, as well as auto-transliteration into square Hebrew script to support orthographic comparison. In line with KAI, the system applies a consistent set of epigraphic sigla to mark restorations, damage, and other textual features. These can be presented in an optional layer of graphemic annotations to preserve the clarity of the main text while supporting detailed scholarly analysis.

3.4 Corpus Growth and Future Extensions

The current corpus, built from an initial set of about twenty inscriptions, functions as a testing ground for the full editorial workflow: editors annotate texts from established scholarly

[2] See Gzella (2011) for linguistic and historical overview.

[3] Cf. Sahle (2016) that defines and characterizes Scholarly Digital Editions (SDEs) as the critical representations of historical documents, distinct from digitized text or digital libraries.

editions, submit their work for peer review, for the inscription data to receive a final approval from a subject specialist before being published and added to the platform. The system is designed for modular growth and for expansion across languages, since its annotation model is grounded in core Semitic morphological concepts such as root, stem, and state, among others. This foundation enables cross-linguistic querying across Canaanite and the wider Semitic family as the project grows. The corpus and its editorial tools were conceived not as a static repository, but as a collaborative research environment built for comparative linguistic work.

4. Layered Annotation Workflow

4.1 Conceptual Architecture

The annotation framework is based on the principle that Semitic inscriptions require multiple interdependent interpretive layers. Unlike monolithic transcription models, the platform adopts a modular architecture in which distinct annotation layers interact. This architecture reflects both the non-linear nature of epigraphic interpretation and the scholarly need to document variation, uncertainty, and contextual dependency at multiple levels.

The system draws conceptual inspiration from multi-layer annotation standards in computational linguistics (e.g., LAF/GrAF; Ide & Romary 2004), but tuned to the specificities of Semitic historical texts.

4.2 Discrete Annotation Layer Tiers

- **Paleographic Layer:** Optional, showing relevant graphemic variation, uncertainty and visual distinctions (e.g., regional ductus).
- **Graphemic Layer:** Encodes the literal signs as segmented from the inscription, aligning each grapheme to its linear position and Unicode representation.
- **Morphosyntactic Layer:** Segments each token into its internal morphemes (e.g., prefixes, stems, suffixes), with annotations for lemma, root, and setting the main part of speech of that token, which then contextually offers additional features to set, such as gender, number, stem, and state.
- **Semantic and Pragmatic Layers:** Include translation equivalents, glosses, context-sensitive disambiguation, semantic field and entity type.
- **Contextual Layer:** Stores editorial notes, variant readings, references to published debates, and degrees of certainty.

Editorial interventions are modeled as span-based annotations over tokens. Tokens carry both diplomatic and normalized forms linked through token identifiers. In other words, lexical annotation, associated with normalized word entities, is linked to one or more tokens. Multiple tokens across lines may comprise a single lexical word, thus preserving the distinction between inscriptional form and linguistic interpretation. Such separation preserves epigraphic fidelity while enabling structured linguistic analysis and reproducibility. Because layers are modular and linked through unique identifiers, scholars can annotate and publish different aspects of an inscription without collapsing them into a single representation.

4.3 Data Flow and Inter-layer Relationships

The system follows a bottom-up workflow: editors enter graphemes, segment tokens, and then may add morphological, phonological and syntactic analysis. Layers remain interlinked but revisable; semantic reinterpretation may trigger morphological updates, all tracked through internal logs to ensure transparency. Prior to the expert-led annotation, the platform already provides limited semi-automated input assistance, with tokens split from transliterated or original script line-level entry. Editors can also copy annotations of previously set tokens through indexed lookups, but keeping their editorial control.

4.4 Collaborative Workflow Design

The annotation process is structured as a staged editorial workflow. Each inscription is assigned to an editor responsible for initial graphemic and linguistic encoding. Submitted annotations enter a peer review, where collaborators may comment, suggest revisions, and flag uncertainties, for the editor's consideration. Final approval is granted by a chief editor, after which the inscription becomes part of the published corpus. All stages are versioned and logged, ensuring traceability of editorial decisions.

4.5 Workflow Overview

Figure 1 shows the staged progression from assignment to publication. The dashed path shows an iterative peer feedback before delivery for approval, reflecting the non-linear nature of philological revision. The model emphasizes controlled state transitions rather than open editing, ensuring accountability while preserving collaborative input.

A typical workflow for an inscription includes the following stages:

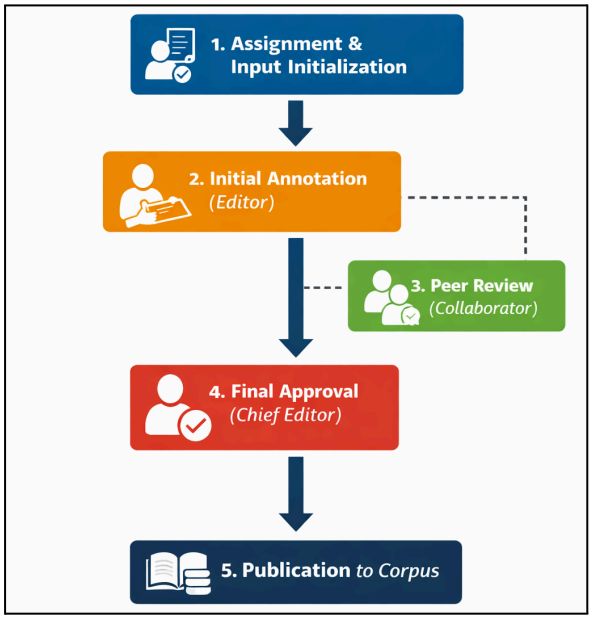


Figure 1. Collaborative Editorial Workflow for Layered Annotation of Semitic Inscriptions.

1. **Assignment & Input Initialization**: An editor is assigned an inscription (e.g., from KAI). The base text is initialized with line and token segmentation in both transliteration and original script Unicode, alongside a sourced transcription.
2. **Initial Annotation (Editor)**: The assigned editor verifies token segmentation and adds token and segment annotation, populating morphological, semantic, and contextual layers. Published editions, lexica, and relevant literature are consulted throughout. Optional commentary is attached at the grapheme, segment or token level, giving room to supplemental or even sometimes contradictionary interpretations as either token-linked linguistic notes or using the built-in Apparatus mechanism.
3. **Peer Review (Collaborator)**: A second editor reviews the annotations. Feedback includes disagreements about segmentation, alternate morphological readings, or suggested corrections in metadata or glossing. Comments and change suggestions are tracked in the platform and reviewed by the editor before final approval. Said disagreements are not automatically resolved but recorded as alternative editorial interpretations linked to relevant tokens or spans. The workflow offers multiple proposed readings or analyses to be reviewed by a chief editor, who makes the final decision, either preserving these proposals in the inscription's notes or version history or discarding them. This ensures transparency of editorial decisions and allows future revisions if new interpretations emerge.
4. **Final Approval (Chief Editor)**: A designated field expert performs a final pass, validating the quality and completeness of the annotation. Submissions, in whole or partially, can be approved as-is or sent back for revision. All editorial decisions are recorded in a transparent log.
5. **Publication**: Approved inscriptions may then be published into the corpus. All layers are stored as discrete, queryable structures. Users may browse, search, and later export annotations per layer or as full bundled records, in addition to the inscription's text and metadata.

Each step in the workflow is timestamped and version-controlled, supporting rollback. A draft lives alongside a public canonical version (should one already exist), until its editorial work is finalized, reviewed and approved, prior to being published for the first time or as a replacement to the existing copy. Researchers can reconstruct editorial decisions and preserve the academically-expected transparency in epigraphy.

The interface supports sequential multi-user annotation with locked editing states, visual differences tracking for chief editors, and source citation fields to anchor interpretations to specific scholarly sources.

The platform's export capabilities will include structured formats (e.g., TEI, LAF), ensuring compatibility with infrastructures like CLARIN or ELRC.

5. Implementation and Infrastructure

5.1 System Architecture

The platform is a browser-based annotation environment with a React frontend and Firebase backend, centered on Firestore, Google's cloud-hosted NoSQL database. Designed for collaborative editing, Firestore provides real-time synchronization, offline access, and fine-grained access control, enabling a responsive, multi-user workflow without custom server infrastructure. It maps naturally onto the hierarchical structure of epigraphic data, where inscriptions contain

metadata, segmented text, and apparatus entries within nested Firestore documents. This supports efficient editing and UI rendering without relying on complex relational joins. Its declarative security rules enable role-specific editorial permissions, essential for managing contributions across contributor, reviewer, and approver roles. Priority is thus given to operational simplicity and scholarly usability: versioned real-time updates are enforced via TypeScript schemas, reducing backend complexity. While Firestore has limits, e.g., no native joins or complex aggregations, these are mitigated through export pipelines planned for integration with SQL backends or TEI/XML-compatible formats for long-term interoperability and corpus-wide analysis (cf. Sahle 2016; Elliott et al. 2017-2026).

5.2 Data Model and File Formats

The platform uses a structured JSON data model reflecting the distinction between physical inscriptional units and linguistic analysis. An inscription comprises metadata, an ordered sequence of lines, and a lexical layer linking tokens to abstract word-level objects[4]. Lines consist of tokens representing physical units of the inscription while tokens store graphemic content, diplomatic and normalized forms, and optional morphological segments. Linguistic analysis is modeled separately: word objects linked via token reference IDs store lemma, root, part of speech, morphology, phonology gloss, and related features. Editorial interventions and variant readings are represented through span-based annotations at either the line or inscription (apparatus) level .

The schema is enforced through TypeScript interfaces to ensure structural consistency and machine readability. While conceptually aligned with standards such as EpiDoc, the model is implemented natively in JSON, with planned export to TEI-compliant formats. A representative JSON example of the inscription data model is provided in Appendix A. Detailed sample annotated data will be made available upon publication.

5.3 Storage, Versioning, and Export

Firestore functions as the platform's primary storage layer, supporting structured queries and cloud-based scalability. Each modification generates a timestamped audit trail at both token and document levels, enabling editors to track annotation history and restore prior states.

The platform is planned to support data export in standard scholarly formats, including TEI-XML for EpiDoc compatibility, TSV or CSV for linguistic analysis, and JSON-LD for semantic web integration. While current exports are user-initiated, the architecture is designed for automated format generation through cloud functions as the corpus scales.

5.4 Interoperability and Integration Potential

Although designed for human-led annotation, the modular architecture facilitates integration with computational tools and external research infrastructures. Planned extensions include interoperability with NLP pipelines for lemmatization, pattern detection, and grapheme-level classification.

The platform is being developed to integrate with research infrastructures such as CLARIN (Váradi et al. 2008) and other archival standards. In addition to TEI and LAF/GrAF export, future development will include RDF serialization compatible with Linked Data frameworks such as the Open Annotation Model and POWLA, allowing annotated inscriptions and linguistic data to be linked with external repositories and research datasets. The system is also designed with compatibility in mind for cultural heritage and museum data standards such as CIDOC-CRM (Crofts, Doerr, & Gill 2003) and Linked Open Data frameworks (Bizer, Heath, & Berners-Lee 2009), enabling connections between inscriptions, places, persons and related historical datasets.

6. A Linguistic Case Study Enabled by Layered Annotation

6.1 Research Question and Linguistic Background

Formulaic language is a hallmark of Phoenician and other Northwest Semitic inscriptions, especially in royal and votive contexts. Across the corpus, recurring motifs such as dedication formulae, genealogical openings, and blessing or curse sequences tend to follow stable morphosyntactic patterns while exhibiting significant variation in orthography, morphology, and vocabulary. Traditionally, these patterns have been discussed qualitatively and compared through philological judgment.

[4] This distinction is required as inscriptional word boundaries may not align with physical line boundaries (KAI 10:L.10↔11, L.11↔12)

This case study examines whether recurring dedication formulae can be operationalized as structured morphosyntactic configurations that are computationally retrievable across inscriptions. Rather than simply identifying new linguistic patterns, the platform shows how familiar constructions can be modeled as structured units through LLA, clarifying their internal architecture, their points of variation, and their reproducibility for further research.

6.2 Data Selection and Annotation Basis

Drawing on a subset of the Phoenician corpus encoded in the platform, the focus is given to five royal and two dedicatory inscriptions (e.g. the Ahiram Sarcophagus, Elibaʿl Osorkon Bust), dated between the 10th and the 3rd centuries BCE, featuring particularly prominent formulaic language.

All selected inscriptions are tokenized and segmented into graphemic and morphological units. Linguistic annotations include lemma, root, part of speech, and relevant morphological features, alongside both diplomatic and normalized representations. This annotation provides a consistent basis for identifying recurring constructions while maintaining alignment with published editions.

6.3 Representing Formulaic Morphosyntax

As a concrete example, the case study focuses on dedication and self-presentation clauses built around verbs of making, dedicating, or establishing, typically followed by an agent (often a royal or donor proper name), an object, and a recipient deity introduced by a prepositional marker. While surface realization varies across inscriptions, the underlying morphosyntactic structure is stable.

Within the LLA model, each component of such a clause is encoded explicitly: the verbal predicate is identified by lemma and morphological features; proper names are tagged as nominal tokens with onomastic and grammatical properties; prepositions and pronominal elements are segmented at the morpheme level; and divine names are annotated as distinct lexical entities. Because token order and segmentation are preserved, equivalent constructions can be compared across inscriptions despite orthographic or morphological variation.

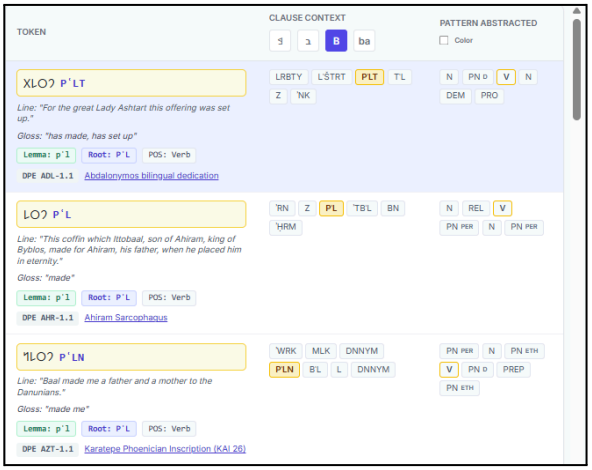


Figure 2: Retrieval of verbal forms

The query in Figure 2 filters tokens by morphological features (POS=**verb**, root=√**PʿL**) and retrieves their linked clause-level annotations. Such structured retrievals expose recurrent predicate-agent-recipient sets across inscriptions. Since these relations are encoded explicitly in the data model, equivalent morphosyntactic templates can be compared independently of surface variation. Formulaic structures can therefore be modeled as configurations of linguistic roles rather than linear text sequences. This representation makes formulaic morphosyntax comparable across inscriptions while maintaining philological interpretation.

6.4 Implications for Comparative Semitic Studies

Although initialized with a limited dataset, the model of structured fields is grounded in shared Semitic morphological principles: root-based lexemes, affixal morphology, state distinctions and syntactic roles. This encoding allows comparison of equivalent constructions across the entire Northwest Semitic corpora and beyond, transcending both time, space, language and dialect boundaries, as shown in Figure 3.

In this respect, not only qualitative comparison is supported, but also systematic investigation of formulaic morphosyntax, lexical distributions and structural variation across related varieties. Bilingual inscriptions further provide controlled environments for aligning normalized representations across linguistic systems. LLA functions as an analytical substrate for comparative Semitic linguistics, enabling reproducible cross-corpus research.

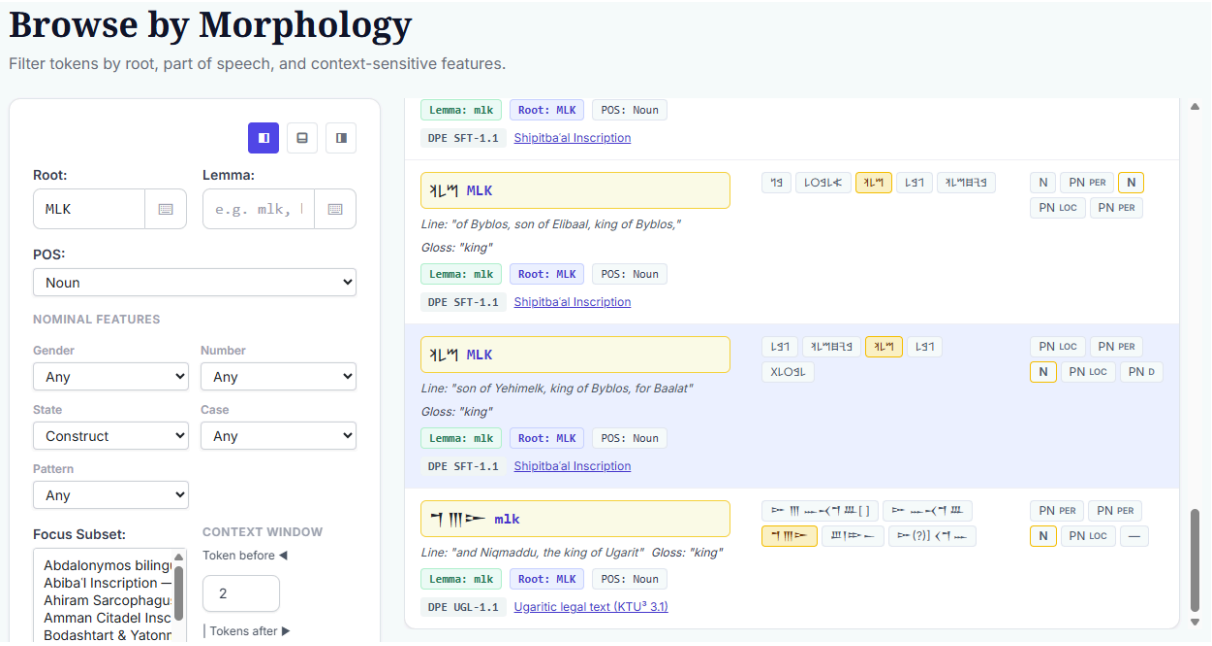


Figure 3. Cross-inscription morphological lookup

7. Limitations and Future Work

7.1 Current Limitations

The platform and corpus are currently in an early stage of development, and several limitations must be acknowledged. The corpus is intentionally restricted to a curated subset of Phoenician inscriptions, with only limited coverage of other Canaanite and Northwest Semitic languages. While sufficient for demonstrating the annotation architecture and supporting focused case studies, this scope does not yet allow large-scale quantitative analysis or broad typological generalizations.

Although inscriptions are tokenized and morphologically segmented, annotation depth and completeness vary across texts, reflecting the iterative and collaborative nature of the editorial workflow. More advanced analytical operations, such as pattern-based multi-token queries or large-scale aggregation, are not yet available at the interface level, despite being supported conceptually by the data model.

Finally, the system currently prioritizes human-led annotation and editorial control. Automated linguistic support, such as morphological pre-annotation or pattern suggestion based on recurring structures, remains under development and is not yet fully integrated into the workflow.

7.2 Planned Development

Corpus coverage and usability are expanded as a main focus. This includes adding further Phoenician, Punic, and other Northwest Semitic inscriptions while maintaining editorial consistency using the review and approval workflow. Interface improvements will support more effective browsing and comparison of annotated structures, including expressive queries over linguistic features and token sequences.

From a technical perspective, future work includes export to standardized formats such as TEI/EpiDoc, ensuring interoperability and long-term preservation. Semi-automated tools, including AI-assisted draft annotation or suggestion mechanisms, are envisaged as optional preliminary editorial aids, to support routine tasks while adhering to expert oversight and final scholarly authority.

While the annotated corpus will be released in stages, with editorial considerations in mind, the platform code is planned to be released as an open-source project, following rigorous stability and multiple user roles operability tests. Separating the annotation infrastructure from the corpus data will allow other projects to reuse the platform for different epigraphic or historical linguistic corpora and contribute to further development of the annotation workflow.

7.3 Long-term Research Roadmap

In the longer term, the project aims to establish a generalizable annotation workflow applicable to a wider range of Semitic language corpora, extending beyond Phoenician to other Canaanite

and Northwest Semitic traditions[5], as well as closely related epigraphic corpora. By maintaining a layered, language-agnostic representation grounded in shared morphological principles, the infrastructure is designed to support comparative research across time, region, and genre.

Future directions include closer integration with analytical environments for corpus and historical linguistics, and the exploration of computational approaches to modeling formulaic language at scale. Throughout these developments, computational methods are intended to support, rather than replace, philological expertise, ensuring that scholarly responsibility and editorial control remain central to the workflow.

8. Conclusion

This paper has presented a layered annotation workflow and digital infrastructure for the computational representation of Semitic epigraphic texts. Fine-grained LLA combined with a structured editorial workflow aligns philological practice with reproducible computational research. Inscriptions are structured, versioned research objects whose analytical layers can be encoded, reviewed, and compared systematically.

Using the focused corpus and a concrete linguistic case study, the paper has shown how well-established epigraphic and morphosyntactic phenomena are explicit, queryable, and comparable across texts. The layered data model enables the representation of graphemic, morphological, semantic, and contextual information. In other words, synchronic and diachronic analysis isn't flattening the interpretive complexity inherent to epigraphic sources.

Although the system is currently limited in scope and remains under active development, its modular design and language-agnostic principles make it extensible to other Semitic epigraphic traditions and related corpora. The workflow demonstrates how computational tools can augment, rather than replace, expert-driven scholarship. By foregrounding annotation as an interpretive and collaborative process, this work contributes a sustainable model for future digital research on historical and ancient languages.

[5] I.e., the epigraphic and written traditions that are linguistically, structurally, and methodologically adjacent to Northwest Semitic corpora, and therefore compatible with the same LLA principles.

9. Required Statements

9.1 Ethics Statement

This research is based on the study of historical and ancient epigraphic texts that are part of the public cultural heritage. The corpus consists exclusively of inscriptions that have been previously published in scholarly editions and does not involve human subjects, personal data, or sensitive information. The annotation workflow is designed to support transparent scholarly interpretation, including the explicit representation of uncertainty, alternative readings, and editorial responsibility. No ethical risks associated with data collection, annotation, or dissemination have been identified.

9.2 Data and Code Availability

The corpus and annotation workflow described in this paper are under active development. The annotated data are derived from published epigraphic sources and are encoded within a custom digital infrastructure designed for collaborative scholarly annotation. At the time of submission, the full dataset and source code are not yet publicly released due to ongoing editorial work and review processes. The authors intend to make selected components of the corpus, along with documentation of the annotation model and workflow, available in future releases, in accordance with licensing constraints and best practices for the sharing of language resources. The platform infrastructure is designed to be released independently of the corpus data, allowing reuse of the annotation environment by other projects.

9.3 Acknowledgements

The authors would like to thank colleagues and mentors for scholarly guidance and discussion that contributed to the development of this project. Particular thanks are due to academic collaborators for their feedback on the conceptual design, linguistic scope, and methodological framing of the workflow. Any remaining errors or limitations are the sole responsibility of the authors.

10. Bibliographical References

Al-Jallad, A. (2015). An Outline of the Grammar of the Safaitic Inscriptions. Brill.

Avanzini, A., Prioletta, A., & Rossi, I. (2014). The Digital Archive for the Study of Pre-Islamic Arabian Inscriptions: an ERC project. https://doi.org/10.13131/UNIPI.ARABIANTICA.0000000002

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data: The story so far. International Journal on Semantic Web and Information Systems, 5(3), 1–22.

Bodard, G., & Stoyanova, S. (2016). Epigraphers and encoders: Strategies for teaching and learning digital epigraphy. In G. Bodard & M. Romanello (Eds.), Digital classics outside the echo-chamber: Teaching, knowledge exchange & public engagement (pp. 51–68). Ubiquity Press. https://doi.org/10.5334/bat.d

Chiarcos, C. (2012). POWLA: Modeling linguistic corpora in OWL/DL. In C. Chiarcos, S. Nordhoff, & S. Hellmann (Eds.), Linked data in linguistics: Representing and connecting language data and language metadata (pp. 225–246). Springer. https://doi.org/10.1007/978-3-642-30284-8_22

Chiarcos, C., Nordhoff, S., & Hellmann, S. (Eds.). (2012). Linked data in linguistics: Representing and connecting language data and language metadata. Springer. https://doi.org/10.1007/978-3-642-28249-2

Crofts, N., Doerr, M., & Gill, T. (2003). The CIDOC Conceptual Reference Model: A standard for communicating cultural contents. Cultivate Interactive, 9.

De Santis, A., Rossi, I. 2018. Crossing Experiences in Digital Epigraphy. From Practice to Discipline. Warsaw-Berlin: De Gruyter.

Depauw, M., & Gheldof, T. (2014). Trismegistos: An interdisciplinary portal of the ancient world. In S. Berti & M. Costa (Eds.), Ancient Worlds in Digital Culture (pp. 293–303). Brill. https://www.trismegistos.org/

Derntl, M., Gietz, P., Helling, P. 2023. FORGE 2023 - Anything Goes?! Forschungsdaten in den Geisteswissenschaften - kritisch betrachtet. Tübingen, 4.-6. Oktober 2023. Zenodo. https://10.5281/zenodo.8341605.

Di Filippo, F. 2018. “Sinleqiunnini: Designing an Annotated Text Collection for Logo-Syllabic Writing Systems”. In De Santis, Annamaria and Irene Rossi (eds.) Crossing Experiences in Digital Epigraphy (pp. 49–64). Warsaw-Berlin: De Gruyter.

Donner, H. and Röllig, W. Kanaanäische und Aramäische Inschriften, 5th ed. (Wiesbaden: Harrassowitz, 2002).

Elliott, T., Bodard, G., Mylonas, E., Stoyanova, S., Tupman, C., Vanderbilt, S., et al. (2017–2026). EpiDoc guidelines: Ancient documents in TEI XML (Version 9). https://epidoc.stoa.org/gl/latest/

Gzella, H. (2011). Phoenician. In H. Gzella (Ed.), Languages from the World of the Bible (pp. 55-75). Berlin: De Gruyter.

Ide, N., & Romary, L. (2004). A registry of standard data categories for linguistic annotation. In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC’04) (pp. 135-138). European Language Resources Association.

Ide, N., & Suderman, K. (2014). The linguistic annotation framework: A standard for annotation interchange and merging. Language Resources and Evaluation, 48(3), 395–418. https://doi.org/10.1007/s10579-014-9268-1

Koller, A. 2013. “Ancient Hebrew מעצד and עצד in the Gezer Calendar”. Journal of Near Eastern Studies 72: 179-193.

Pierazzo, E. (2016). Textual scholarship and text encoding. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.), A new companion to digital humanities (pp. 307-321). Wiley-Blackwell.

Renan, E. (1881-1962). Corpus Inscriptionum Semiticarum (CIS). Historical multi-volume collection of Semitic inscriptions.

Sahle, P. (2016). What is a scholarly digital edition? In M. J. Driscoll & E. Pierazzo (Eds.), Digital scholarly editing: Theories and practices (pp. 19-39). Open Book Publishers. https://doi.org/10.11647/OBP.0095

Sanders, S. 2008. “Writing and Early Iron Age Israel: Before National Scripts, Beyond Nations and States”. In Tappy, Ron and P. Kyle McCarter (eds.) Literate Culture and Tenth-Century Canaan: The Tel Zayit Abecedary in Context. Winona Lake, IN (pp. 100-102).

Sanderson, R., Ciccarese, P., & Van de Sompel, H. (2017). Web annotation data model. World Wide Web Consortium (W3C). https://www.w3.org/TR/annotation-model/

TEI Consortium. (2022). Guidelines for Electronic Text Encoding and Interchange (Version P5, 4.5.0). https://tei-c.org/release/doc/tei-p5-doc/en/html/

Váradi, T., Wittenburg, P., Krauwer, S., Wynne,

M., Koskenniemi, K., & others. (2008). CLARIN: Common language resources and technology infrastructure. In Proceedings of LREC 2008. ACL Anthology. https://aclanthology.org/L08-1008/

Vitale, V., de Soto, P., Simon, R., Barker, E., Isaksen, L. and Kahn, R. 2021. Pelagios - Connecting Histories of Place. Part I: Methods and Tools. International Journal of Humanities and Arts Computing, 15/1-2: 5–32.

10.1. Digital Corpora and Online Language Resources

ACO Arabic Collections Online. (n.d.). *Arabic Collections Online*. Digital Library Technology Services. https://dlib.nyu.edu/aco/

AICC AI Cuneiform Corpus. https://aicuneiform.com/

Beta maṣāḥəft: Manuscripts of Ethiopia and Eritrea. www.betamasaheft.uni-hamburg.de

CASPO Corpus of Akkadian Shuila Prayers Online. http://www.shuilas.org/

CDL Cambridge Digital Library. Ethiopian Manuscripts. https://cudl.lib.cam.ac.uk/collections/ethiopianmanuscripts/1

CDLI Cuneiform Digital Library Initiative. (n.d.). *Home*. https://cdli.earth/

Corpus Coranicum https://www.arabic-latin-corpus.philosophie.uni-wuerzburg.de/

CSAI Corpus of South Arabian Inscriptions. https://dasi.cnr.it/index.php?id=42&prjId=1&corId=0&colId=0

DANES Online Resources for Digital Ancient Near Eastern Studies. https://opendanes.org/nav/DANES-resources.html

Digital Archive for the Study of Pre-Islamic Arabian Inscriptions (DASI). (n.d.). *DASI*. https://dasi.cnr.it/

DCGAS Digital Corpus for Graeco-Arabic Studies. https://www.graeco-arabic-studies.org/home.html

Ebla Digital Archives (EbDA). (n.d.). *Ebla Digital Archives*. http://ebda.cnr.it/

eHammurabi https://ehammurabi.org/

EPIDAT The Database of Jewish Epigraphy. https://www.re3data.org/repository/r3d100012712

National Library of Israel. (2024). Ktiv: The International Collection of Digitized Hebrew Manuscripts. https://www.nli.org.il/en/discover/manuscripts/hebrew-manuscripts

Lauinger, J. and Yoder, T. 2014. AEMW/ Amarna. https://oracc.museum.upenn.edu/aemw/amarna/index.html

OCIANA. (n.d.). *Online Corpus of the Inscriptions of Ancient North Arabia*. https://dasi.cnr.it/index.php?id=42&prjId=4

Open Richly Annotated Cuneiform Corpus (ORACC). (n.d.). *ORACC*. https://oracc.museum.upenn.edu/

PDLIM Princeton Digital Library of Islamic Manuscripts. https://dpul.princeton.edu/islamicmss

PFDP Polonsky Foundation Digitization Project. http://bav.bodleian.ox.ac.uk/hebrew-manuscripts

QAC The Quranic Arabic Corpus. https://corpus.quran.com/

syri.ac. (n.d.). *An Annotated Bibliography of Syriac Resources Online*. https://syri.ac/digimss

Appendix A. Inscription Data Model

A.1 Overview

This appendix presents a representative JSON structure for an inscription document in the annotation platform. The data model separates inscription-level metadata, line-based diplomatic transcription, token- and grapheme-level epigraphic annotation, and word-level linguistic annotation. In the implementation, each inscription is stored as a single document containing nested arrays for lines, translations, editorial spans, and tokens, together with a separate `wordsById` object for lexical entries that may link to one or more tokens. This design preserves the diplomatic transcription while allowing linguistic, phonological, and editorial annotations to be queried independently.

A.2 Top-level Inscription Document Structure

A representative inscription document contains metadata, line objects, word-level records, and workflow state:

```
{
  "metadata": {
    "iid": "DPE-AHR-1.1",
    "ctg_id": "DPE-AHR-1",
    "legacy_file_id": "KAI_1",
    "language": "Phoenician",
    "script": "Phoenician Alphabet",
    "date": "10th Century BCE",
    "date_year_sort": -950,
    "period": "Iron Age IIA",
    "provenance": {
      "place": "Byblos, Lebanon",
           "coordinates": { "lat": 34.1236,
"long": 35.6513 }
    },
    "location": {
      "place": "National Museum of Beirut"
    },
    "legacy_refs": [
       { "type": "KAI", "number": "1", "ref":
"KAI 1" }
    ],
    "sigla": [
      { "label": "KAI 1" }
    ],
    "bibliography": [
       { "citation": "Donner, H. & Röllig, W.
(1962). KAI, vol. 1, no. 1." }
    ],
    "workflowStatus": "published",
    "visibility": "public"
  },
  "lines": [...],
  "wordsById": {...},
  "peerReviewState": "completed"
}
```

The `metadata` object stores bibliographic, chronological, geographic, and administrative information. The inscription text itself is represented in `lines`, while `wordsById` stores the separate lexical layer.

A.3 Line and Token Structure

The diplomatic transcription is stored as an ordered array of `lines`. Each line contains `tokens` and may also contain `translations` and editorial spans.

```
{
  "id": "L1",
  "line_number": 1,
  "translations": [
    {
      "id": "tr-L1-en",
      "language": "en",
       "text": "The coffin which Ittobaal, son
of Ahiram, king of Byblos, made",
      "source": "Gibson 1982"
    }
  ],
  "editorialSpans": [
    {
      "id": "es-1",
      "kind": "supplied",
      "cert": "high",
       "start": { "tokenId": "T1.3", "pos": 0
},
      "end": { "tokenId": "T1.3", "pos": 3 },
      "sourceEdition": "KAI"
    }
  ],
  "tokens": [
    {
      "id": "T1.1",
      "form": "ʾrn",
      "script_form": "𐤀𐤓𐤍",
      "normalized_form": "ʾarōn",
      "wordId": "W1.1",
      "lemma": "ʾrn",
      "root": "ʾry",
      "pos": "Noun",
      "gloss": "coffin",
      "morphology": {
        "gender": "Masculine",
        "number": "Singular",
        "state": "Absolute"
      }
    }
  ]
}
```

This structure keeps the physical sequence of the inscription at line and token level, while allowing each token to carry normalized, lexical, and morphological information. `editorialSpans` encode line-level editorial phenomena such as supplied text, gaps, or corrections.

A.4 Grapheme and Segment Annotation

Token-internal annotation can be represented at grapheme and segment level. Grapheme objects capture character-level epigraphic information such as attestation, restoration, damage, and certainty, while segment objects encode morpheme-level analysis aligned to graphemes.

```
{
  "id": "T1.1",
  "form": "ʾrn",
  "graphemes": [
    {
      "id": "T1.1:g:1",
      "n": 1,
      "base": "ʾ",
      "presence": "attested",
      "certainty": "secure",
      "damage": "none",
           "anchor": { "kind": "tokenSpan",
"start": 0, "end": 1 }
    },
    {
      "id": "T1.1:g:3",
      "n": 3,
      "base": "n",
      "presence": "restored",
      "certainty": "uncertain",
      "damage": "none",
           "anchor": { "kind": "tokenSpan",
"start": 2, "end": 3 }
    }
  ],
  "segments": [
    {
      "id": "T1.1:m:1",
      "n": 1,
      "kind": "stem",
      "form": "ʾrn",
      "lemma": "ʾrn",
      "gloss": "coffin",
      "morph": "N.ms.abs",
      "anchor": {
        "kind": "graphemes",
                "graphemeIds": ["T1.1:g:1",
"T1.1:g:2", "T1.1:g:3"]
      }
    }
  ]
}
```

This layer allows epigraphic and morphological annotation to remain linked without collapsing them into a single representation.

A.5 Token-Word Linking

The model distinguishes between inscriptional tokens and word-level lexical records. Tokens preserve the diplomatic sequence of the inscription and may contain local annotation, while `wordsById` stores a separate lexical object that can link to one or more tokens through `tokenRefs`. This is particularly useful when an interpreted lexical word spans multiple inscriptional units, or when clitics and compound constructions must be represented without losing the original inscriptional segmentation.

A representative word-level object is shown below:

```
{
  "W1.1": {
    "id": "W1.1",
    "tokenRefs": [
        { "lineId": "L1", "tokenId": "T1.1",
"order": 0 }
    ],
    "leaderTokenId": "T1.1",
    "leaderLineId": "L1",
    "displayPolicy": "leaderOnly",
    "lemma": "ʾrn",
    "root": "ʾry",
    "pos": "Noun",
    "gloss": "coffin",
    "morphology": {
      "gender": "Masculine",
      "number": "Singular",
      "state": "Absolute"
    }
  }
}
```

In this structure, `tokenRefs` records the inscriptional token or tokens associated with a lexical entry, while `leaderTokenId` and `leaderLineId` identify the default display anchor. This separation makes it possible to preserve inscriptional form and editorial segmentation at token level while storing normalized lexical analysis in a reusable word-level layer.

A.6 Translation and Metadata Structure

Translations are stored per line rather than only at inscription level, which allows multiple aligned renderings and source attributions.

```
{
  "translations": [
    {
      "id": "tr-L1-en",
      "language": "en",
       "text": "The coffin which Ittobaal, son
of Ahiram, king of Byblos, made",
      "source": "Gibson 1982"
    }
  ]
}
```

Metadata fields are stored in a separate inscription-level object and include identifiers, chronology, provenance, present location, bibliography, sigla, and legacy references. A compact example is shown below:

```
{
  "metadata": {
    "iid": "DPE-AHR-1.1",
    "ctg_id": "DPE-AHR-1",
    "legacy_file_id": "KAI_1",
    "language": "Phoenician",
    "script": "Phoenician Alphabet",
    "date": "10th Century BCE",
    "period": "Iron Age IIA",
    "provenance": {
      "place": "Byblos, Lebanon",
           "coordinates": { "lat": 34.1236,
"long": 35.6513 }
    },
    "location": {
      "place": "National Museum of Beirut"
    },
    "legacy_refs": [
       { "type": "KAI", "number": "1", "ref":
"KAI 1" }
    ],
    "sigla": [
      { "label": "KAI 1" }
    ]
  }
}
```

In the implementation, the metadata layer may additionally include multilingual names, images, bibliographic records, visibility settings, and workflow state.

A.7 Main Layers in the Inscription Document

The main components of the data model are as follows:

- `metadata`: per document (root), bibliographic, geographic, chronological, and administrative information
- `lines[]`: per document, the physical line divisions of the inscription:
  - `translations[]`: per-line translations, optionally with source attribution
  - `editorialSpans[]`: per-line, editorial ranges such as supplied, unclear, or corrected text
  - `tokens[]`: per-line, inscriptional token units with local annotation:
    - `graphemes[]`: per token, grapheme-level segmentation and physical condition flags
    - `segments[]`: per token, morpheme-level segmentation aligned to graphemes
    - `phonology`: per token, phonological transcription and feature bundles
- `wordsById`: per document (root), word-level lexical objects linked to (spanning) one or more tokens, for clitics, compound constructions, etc.

A representative set of key enumerated values includes:

- `presence`: `attested`, `restored`, `expanded`, `corrected`, `missing`
- `certainty`: `secure`, `uncertain`, `highly_uncertain`
- `damage`: `none`, `damaged`, `illegible`
- `pos`: `Noun`, `Verb`, `ProperNoun`, `Adjective`, `Adverb`, `Preposition`, `Conjunction`, `Demonstrative`, `DefiniteArticle`, `Numeral`, `Suffix`, `Particle`, `Pronoun`
- `state`: `Absolute`, `Construct`, `Emphatic`
- `editorialSpan.kind`: `supplied`, `gap`, `unclear`, `add`, `del`, `corr`, `choice`, `abbrExpan`

Overview of the Dependency Parsing Task at EvaLatin 2026

Federica Iurescia, Marco Passarotti, Rachele Sprugnoli
Università Cattolica del Sacro Cuore
Largo Gemelli 1, 20123 Milan, Italy
federica.iurescia@unicatt.it; marco.passarotti@unicatt.it; rachele.sprugnoli@unicatt.it

Abstract

This paper presents the organization, methodology, and outcomes of the Dependency Parsing shared task held within the fourth edition of EvaLatin, a campaign dedicated to the evaluation of Natural Language Processing tools for Latin. EvaLatin aims to promote and advance research in language technologies for Latin, fostering the development of robust and linguistically informed computational approaches. The paper provides a detailed description of the data released for the shared task. It also outlines the evaluation framework and metrics adopted for assessing system performance. The results achieved by participating teams are reported and comparatively analyzed, highlighting strengths, limitations, and emerging trends in current approaches to Latin dependency parsing. Finally, the paper discusses the main challenges posed by the task and suggests directions for future research in the field.

Keywords: Latin, evaluation, dependency parsing

1. Introduction

EvaLatin 2026 is the fourth edition of the campaign devoted to the evaluation of Natural Language Processing (NLP) tools for the Latin language. As in 2020 (Sprugnoli et al., 2020), 2022 (Sprugnoli et al., 2022), and 2024 (Sprugnoli et al., 2024), EvaLatin is proposed as part of the *Workshop on Language Technologies for Historical and Ancient Languages* (LT4HALA), co-located with LREC 2026.[1] Similar to what happens in other international evaluation campaigns, participants were provided with shared test data that are made freely available for research purposes to encourage further improvement of language technologies for Latin. A shared evaluation script was also provided. Data, scorer and detailed guidelines are available in a dedicated GitHub repository.[2]

EvaLatin 2026[3] is organized around 2 tasks: Dependency Parsing and Named Entity Recognition (Boano et al., 2026). In what follows, the organization and results of the Dependency Parsing tasks are detailed.

2. Dependency Parsing

The aim of the task is to provide syntactic analysis of Latin texts following the Universal Dependencies (UD) framework (de Marneffe et al., 2021). The output submitted by the participants is a CoNLL-U file,[4] with indications of the syntactic head and of the dependency relations in the fields 7 (HEAD) and 8 (DEPREL), respectively. Other annotation levels (tokenization, sentence splitting, part-of-speech tagging and morphological features) are given as gold labels, since they are not the focus of the evaluation.

3. Data

No specific training data are released but participants are free to make use of any resource they consider useful for the task, including the Latin treebanks already available in the UD collection. In this regard, one of the challenges of this task is to understand which treebank (or combination of treebanks) is the most suitable to deal with new test data. The task aims at improving a state of the art that is not optimal. UD treebanks currently show different degrees of harmonization, and Latin is not an exception in this respect (Gamba and Zeman, 2023a); (Gamba and Zeman, 2023b).

Reflecting this situation, some texts in the dataset include punctuation, some do not, as this is the actual state of the art for Latin treebanks and corpora. Texts provided as test data are by one Classical author (Seneca the Younger) and one Medieval author (Thomas Aquinas) for a total of more than 8,000 tokens. Each author is taken as a representative of a specific text genre: Seneca the Younger for poetry, more specifically tragedy, with the *Phoenissae* (4,155 tokens), composed in the 1st century CE; Thomas Aquinas for prose, more specifically philosophical-theological treatise, with a portion of the *Summa Theologiae* (*ST*) (4,151 tokens), written

[1] https://lrec2026.info/

[2] https://github.com/CIRCSE/LT4HALA/tree/master/2026/data_and_doc

[3] EvaLatin is an initiative organized by the CIRCSE research centre at the Università Cattolica del Sacro Cuore in Milan, Italy. https://centridiricerca.unicatt.it/circse/en.html

[4] https://universaldependencies.org/format.html

Proceedings of LT4HALA 2026 @ LREC 2026, pages 211–218.
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

```
# sent_id = CaesBG4-A-01-607
# text = neque multum frumento sed maximam partem lacte atque pecore uiuunt multumque sunt in uenationibus
1   neque   neque   CCONJ   S   Polarity=Neg   _   _   _   LiLaflcat=i
2   multum  multum  ADV M   Degree=Pos  _   _   _   LASLAVariant=2|LiLaflcat=i
3   frumento    frumentum   NOUN    A2  Case=Abl|Gender=Neut|InflClass=IndEur0|Number=Sing  _   _   _   LiLaflcat=n2
4   sed sed CCONJ   S   _   _   _   _   LiLaflcat=i
5   maximam magnus  ADJ C1  Case=Acc|Degree=Abs|Gender=Fem|InflClass=IndEurA|Number=Sing   _   _   _   LiLaflcat=n6
6   partem  pars    NOUN    A3  Case=Acc|Gender=Fem|InflClass=IndEurI|Number=Sing   _   _   _   LiLaflcat=n3
7   lacte   lac NOUN    A3  Case=Abl|Gender=Masc|InflClass=IndEurI|Number=Sing  _   _   _   LiLaflcat=n3
8   atque   atque   CCONJ   S   _   _   _   _   LASLAVariant=1|LiLaflcat=i
9   pecore  pecus   NOUN    A3  Case=Abl|Gender=Neut|InflClass=IndEurX|Number=Sing  _   _   _   LASLAVariant=1|LiLaflcat=n3
10  uiuunt  uiuo    VERB    B3  Aspect=Imp|InflClass=LatX|Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act   _   _   _   LiLaflcat=v3
11-12   multumque   _   _   _   _   _   _   _   _
11  multum  multum  ADV M   Degree=Pos  _   _   _   LASLAVariant=2|LiLaflcat=i
12  que que CCONJ   S   _   _   _   _   LiLaflcat=i
13  sunt    sum AUX B6  Aspect=Imp|InflClass=LatAnom|Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin  _   _   _   LASLAVariant=1|LiLaflcat=v6
14  in  in  ADP R   AdpType=Prep    _   _   _   LiLaflcat=i
15  uenationibus    uenatio NOUN    A3  Case=Abl|Gender=Fem|InflClass=IndEurX|Number=Plur   _   _   _   LiLaflcat=n3
```

Figure 1: Example of the test data format.

in the 13th century.[5]

Data for Seneca the Younger are taken from the *Opera Latina* corpus,[6] (Denooz, 2004) a linguistic resource manually annotated since 1961 by the Laboratoire d'Analyse Statistique des Langues Anciennes (LASLA) at the University of Liège, Belgium. Original data were converted into the annotation formalism of UD, automatically parsed with a Stanza model (Qi et al., 2020) retrained on the UD_Latin-CIRCSE treebank,[7] and manually revised by an expert of Latin language and literature.

Data for Thomas Aquinas are taken from the *Index Thomisticum* (Passarotti, 2019).[8] They were manually annotated with tokenization, sentence splitting, lemmatization, part-of-speech tagging, and syntactic annotation originally according to Prague Dependency Treebank formalism and then converted into UD format.

Data are distributed in the CoNLL-U format. Accordingly, the annotations are plain text files having the `.conllu` extension and encoded in UTF-8. An example of the test data is provided by Figure 1.

4. Evaluation

The scorer employed for the evaluation of the Dependency Parsing task is a revised version of the one developed for the *CoNLL18 Shared Task on Multilingual Parsing from Raw Text to Universal Dependencies*.[9] Our version of the script allows for the evaluation of parsing either with or without taking dependency relations subtypes into account, and is available on the EvaLatin 2026 web page.[10]

The evaluation starts by aligning the system-produced tokens to the gold standard one; given that we provide test data already tokenized, sentence-splitted and annotated with morpho-grammatical information, the alignment for tokens, sentences, words, UPOS, UFeats and lemmas should be perfect (i. e. 100.00).[11] Then, CLAS (Content-Word Labeled Attachment Score)[12] and LAS (Labeled Attachment Score)[13] are evaluated in terms of Precision, Recall, F1 both with and without subtypes.[14]

As for the baseline, we provide the scores obtained on the test data using UDPipe 2 (Straka et al., 2016) with the models trained on the UD_Latin-CIRCSE treebank v. 2.17, and the *Index Thomisticus* Treebank (UD_Latin-ITTB)[15] v. 2.17, respectively, as they are available from the tool's web interface.[16] Baseline results are given in Table 1

[5] To ensure a balanced corpus, only a portion of the *Summa Theologiae* was included in the test data.

[6] https://www.lasla.uliege.be/cms/c_8508894/fr/lasla

[7] https://github.com/UniversalDependencies/UD_Latin-CIRCSE

[8] https://itreebank.marginalia.it.

[9] https://universaldependencies.org/conll18/evaluation.html

[10] https://circse.github.io/LT4HALA/2026/EvaLatin

[11] Although UPOS values were provided in the test data, the alignment is not always perfect. On poetry data, no system achieves 100.00 on UPOS. On prose data, only THIVLVC_2 and OmnesFlores_2 do not achieve 100.00 on UPOS. Evaluation on UPOS is not performed; it is worth noting that participants experimented, among others, also on the impact of UPOS on parsing; see Section 5 and (Matsuda and Asahara, 2026); (Stymne, 2026).

[12] CLAS is the labeled F1-score over all relations except those involving function words (`aux`, `case`, `cc`, `clf`, `cop`, `det`, `mark`) and punctuation (`punct`) (Nivre and Fang, 2017).

[13] LAS is the percentage of tokens assigned both the correct `DEPREL` and `HEAD` (Buchholz and Marsi, 2006).

[14] The scorer computes also the Unlabeled Attachment Score (UAS), that is the percentage of tokens assigned the correct HEAD; the Morphology-aware Labeled Attachment Score (MLAS), that is CLAS extended with evaluation of POS tags and morphological features; the Bi-Lexical dependency score (BLEX) that combines content-word relations with lemmatization, but not with POS tags and features. These 3 metrics are not taken into account for this shared task.

[15] https://github.com/UniversalDependencies/UD_Latin-ITTB

[16] http://lindat.mff.cuni.cz/services/udpipe/

Metric	Precision	Recall	F1
CLAS	56.14	57.76	56.94
LAS	57.22	57.22	57.22

Metric	Precision	Recall	F1
CLAS	57.29	57.19	57.24
LAS	59.74	59.74	59.74

Table 1: Baseline results on poetry with subtypes on the left and without subtypes on the right.

Metric	Precision	Recall	F1
CLAS	80.33	78.51	79.41
LAS	82.17	82.17	82.17

Metric	Precision	Recall	F1
CLAS	83.00	83.14	83.07
LAS	85.21	85.21	85.21

Table 2: Baseline results on prose with subtypes on the left and without subtypes on the right.

for poetry, and 2 for prose. As the results in the Tables show, the performance is substantially higher on prose than on poetry. In addition, removing dependency subtypes consistently improves results, suggesting that fine-grained label distinctions increase task difficulty, especially in structurally complex poetic texts.

5. Participants

Four teams took part in the task. One team did not submit the report; therefore, it will not be included in this overview. Details on the participating teams and their systems are given below:

- Omnes Flores. This team submitted two runs. Their system is an extended version of an existing UD-based NLP framework that relies on multilingual Large Language Models (LLMs) with Latin data taken from the six UD_Latin treebanks. The difference in the two runs lies on how the adopted models were trained. The model used for OmnesFlores_1 has been trained with the default configuration, that takes a list of word FORM values as input. Conversely, the model used for OmnesFlores_2 has been trained taking also the values in the UPOS column of the CoNLL-U as input. Adding UPOS in the training has a positive impact on the system's performance on prose, while negatively affecting performance on poetry.
- THIVLVC. This team submitted two runs, both based on LLMs. The difference in the two runs lies in the adoption of Retrieval-Augmented Generation (RAG) on UD_Latin-CIRCSE treebank in the run THIVLVC_2, which improves system's performance both on poetry and prose.
- UppsalaNLP. This team submitted two runs, experimenting on improvement of out-of-the-box parser, via cross-lingual and cross-treebank training, with data taken from UD treebanks of historical languages. The difference in the two runs lies in using only Latin data for UppsalaNLP_2, which yields slightly lower results than UppsalaNLP_1 for prose, while performing slightly better for poetry.

6. Results and Discussion

Tables 3, 4, 5 and 6 show the final ranking for poetry (POE.) and prose (PRO.). Results are provided in terms of F1, including the baseline discussed in Section 4.

Overall, performances of all participating systems are consistently higher on prose than on poetry, confirming the tendency observed for the baseline. The best results for poetry are achieved by the THIVLVC team leveraging LLMs, particularly the run adopting RAG on UD_Latin-CIRCSE treebank (i.e., THIVLVC_2). Conversely, the same run by the THIVLVC team shows lower performances on prose. The run achieving the best results on prose relies on transfer learning and fine-tuning of historical language data (i.e., UppsalaNLP_1).

In what follows, we provide an overview of the general trends emerging from the analysis, and discuss selected results displayed in Tables 7, 8, 9, and 10. Systems achieve low performances on correctly identifying the head node of adverbial clauses (`advcl`) – ranging from roughly 25 to slightly above 45 on poetry and from nearly 50 to slightly above 70 on prose – and of adnominal clauses (`acl`) – ranging from about 23 to 60 on poetry and from slightly above 55 to nearly 75 on prose – (see Table 7).

The following example from prose data highlights the challenges involved in identifying adverbial clauses:

> **Example 1**
> Th. Aquinas *ST*. *Prooemium*
> *haec igitur et alia huiusmodi euitare studentes, tentabimus, [...] prosequi*
> Endeavoring to avoid these and other like faults, we shall try, [...] to set forth[17]

In Example 1, the token *studentes* is a present participle which is the head node of the adverbial clause in this sentence. In the gold data

[17] English translations of *Summa Theologiae* are available at: https://aquinas.cc/la/en/~ST.I.S1.

TEAM	F1 POE.
THIVLVC_2	74.03
THIVLVC_1	72.71
UppsalaNLP_2	68.47
UppsalaNLP_1	67.34
Omnes Flores_1	60.79
BASELINE	56.94
Omnes Flores_2	55.37

TEAM	F1 PRO.
UppsalaNLP_1	83.14
UppsalaNLP_2	82.17
Omnes Flores_2	81.35
THIVLVC_2	80.92
Omnes Flores_1	80.60
BASELINE	79.41
THIVLVC_1	74.04

Table 3: Dependency Parsing results in terms of CLAS with subtypes.

TEAM	F1 POE.
THIVLVC_2	76.08
THIVLVC_1	76.00
UppsalaNLP_2	69.16
UppsalaNLP_1	68.20
Omnes Flores_1	63.17
Omnes Flores_2	58.37
BASELINE	57.24

TEAM	F1 PRO.
UppsalaNLP_1	86.88
THIVLVC_2	86.60
UppsalaNLP_2	86.45
Omnes Flores_2	85.28
Omnes Flores_1	84.17
BASELINE	83.07
THIVLVC_1	81.52

Table 4: Dependency Parsing results in terms of CLAS without subtypes.

TEAM	F1 POE.
THIVLVC_2	72.88
THIVLVC_1	70.36
UppsalaNLP_2	67.56
UppsalaNLP_1	66.63
OmnesFlores_1	60.83
BASELINE	57.22
OmnesFlores_2	54.74

TEAM	F1 PRO.
UppsalaNLP_1	84.41
UppsalaNLP_2	83.74
OmnesFlores_2	83.74
THIVLVC_2	83.26
OmnesFlores_1	83.26
BASELINE	82.17
THIVLVC_1	75.72

Table 5: Dependency Parsing results in terms of LAS with subtypes.

TEAM	F1 POE.
THIVLVC_2	77.60
THIVLVC_1	76.97
UppsalaNLP_2	70.55
UppsalaNLP_1	69.59
OmnesFlores_1	65.64
OmnesFlores_2	60.20
BASELINE	59.74

TEAM	F1 PRO.
THIVLVC_2	87.93
UppsalaNLP_1	87.55
UppsalaNLP_2	87.26
OmnesFlores_2	87.09
OmnesFlores_1	86.10
BASELINE	85.21
THIVLVC_1	82.78

Table 6: Dependency Parsing results in terms of LAS without subtypes.

(see Figure 2), it is annotated as the head of an adverbial clause carrying secondary predication (`advcl:pred`). Only runs from the THIVLVC team converged with the gold annotation as `advcl:pred`. The other four runs exhibit each a different annotation: *studentes* has been annotated as

- `advcl` depending from the root identified in the token *tentabimus*;
- clausal subject (`csubj`) depending from the root identified in the token *tentabimus*;
- nominal subject (`nsubj`) depending from the token *euitare* annotated as `advcl`;
- `nsubj` depending from the root identified in the token *tentabimus*.

Such a variation exemplifies the general tendency of systems as in Table 7. Zooming in on the dependency relations occurring in annotations that diverge from gold data, indeed, we find adverbial clause, that is, the dependency relation in the gold data without the indication of the subtype. The other annotations emerging from the runs, namely nominal subject and clausal subject, are, interestingly, dependency relations for which systems exhibit overall good performances, though some challenges remain. Exploring in greater details dependency relations identifying subjects, both nominals

	acl		advcl	
	PRO.	POE.	PRO.	POE.
OmnesFlores_1	64.10	27.45	50.47	25.90
OmnesFlores_2	60.87	22.97	49.48	25.68
THIVLVC_1	56.47	39.02	64.38	46.59
THIVLVC_2	74.07	37.24	71.49	42.55
UppsalaNLP_1	70.89	60.34	52.53	40.00
UppsalaNLP_2	69.88	56.52	50.46	45.12

Table 7: F1 on selected deprel with very low scores.

	nsubj		nsubj:outer		nsubj:pass	
	PRO.	POE.	PRO.	POE.	PRO.	POE.
OmnesFlores_1	90.45	68.14	NotUsed	NotUsed	86.73	36.00
OmnesFlores_2	89.91	62.23	NotUsed	NotUsed	87.38	33.90
THIVLVC_1	90.67	84.02	NotUsed	NotUsed	84.00	59.65
THIVLVC_2	91.36	82.98	NotUsed	NotUsed	90.29	56.72
UppsalaNLP_1	92.68	77.84	0.00	0.00	92.23	65.38
UppsalaNLP_2	91.87	77.67	0.00	0.00	91.75	56.25

Table 8: F1 on the nsubj deprel and related subtypes.

	csubj		csubj:cleft		csubj:pass		csubj:relcl	
	PRO.	POE.	PRO.	POE.	PRO.	POE.	PRO.	POE.
OmnesFlores_1	85.71	23.08	NotUsed	NotUsed	89.16	NotUsed	0.00	0.00
OmnesFlores_2	85.71	35.71	NotUsed	NotUsed	89.41	NotUsed	0.00	0.00
THIVLVC_1	62.79	36.84	NotUsed	NotUsed	27.45	NotUsed	0.00	0.00
THIVLVC_2	65.85	33.33	NotUsed	NotUsed	71.43	NotUsed	0.00	13.33
UppsalaNLP_1	66.67	33.33	NotUsed	NotUsed	76.71	0.00	33.33	50.00
UppsalaNLP_2	65.79	20.00	NotUsed	0.00	72.46	NotUsed	22.22	42.11

Table 9: F1 on the csubj deprel and related subtypes.

	case		amod		det		obj	
	PRO.	POE.	PRO.	POE.	PRO.	POE.	PRO.	POE.
OmnesFlores_1	97.94	90.04	98.00	71.39	75.63	95.35	91.51	74.79
OmnesFlores_2	97.14	90.62	97.52	64.70	61.13	95.95	90.20	69.05
THIVLVC_1	96.16	97.64	89.53	78.91	69.71	74.29	95.61	89.73
THIVLVC_2	96.82	98.02	98.51	84.08	85.08	96.79	95.61	87.96
UppsalaNLP_1	96.98	92.13	96.53	78.17	77.42	96.79	94.63	76.33
UppsalaNLP_2	97.00	91.76	97.76	80.30	80.21	97.09	95.10	76.60

Table 10: F1 on selected deprel with very high scores.

(`nsubj`) and clausal (`csubj`), with their respective subtypes, the results show that they are mostly correctly identified on prose data, whereas all systems achieve lower performances for these deprels on poetry (see Tables 8 and 9, respectively).

Example 2 is instead taken from poetry data:

Example 2
Sen. *Phoen*. 214-215
turba fortunae prior abscessit a te iussa
Those who thronged around your former fortunes have left at your command[18]

[18]English translations from Seneca's *Phoenissae* are drawn by (Fitch, 2018).

As displayed in Figure 3, the token *iussa* is the head node of the adnominal clause. Only two runs from two different systems converged with the gold annotation as `acl`. In the other runs, *iussa* has been annotated differently. Three runs converge on identifying *iussa* as `advcl` depending from the root identified in the token *abscessit*. In one run *iussa* has been annotated as `advcl:pred` depending from the root identified in the token *abscessit*.

Moving away from examples towards general tendencies emerging from the data, it is worth noting that systems predict deprels or subtypes not used in gold data. This is the case, for instance, of nominal subjects of a copular clause whose predicate

Example 1
Th. Aquinas *ST*. *Prooemium*

haec igitur et alia huiusmodi euitare studentes, tentabimus, [...] prosequi
Endeavoring to avoid these and other like faults, we shall try, [...] to set forth

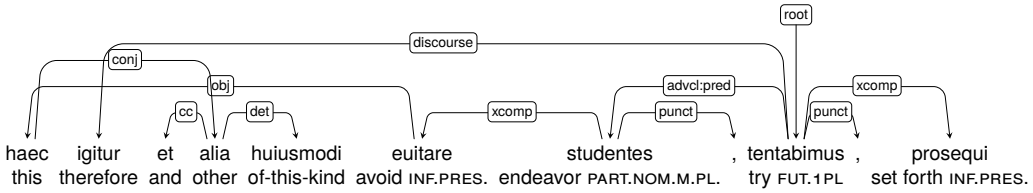


Figure 2: Syntactic tree of the first example: adverbial clause in prose text.

Example 2
Sen. *Phoen*. 214-215

turba fortunae prior abscessit a te iussa
Those who thronged around your former fortunes have left at your command

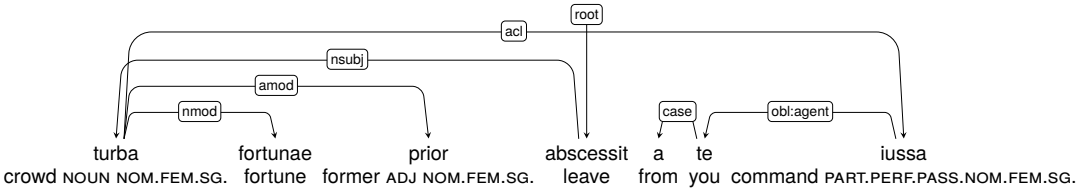


Figure 3: Syntactic tree of the second example: adnominal clause in poetry.

is itself a clause (`nsubj:outer`) in Table 8: this subtype does not occur in gold data, is predicted by UppsalaNLP systems – scoring 0 –, and is not used by other systems – appearing as NotUsed.

Conversely, some deprels occur in gold data, but are not identified by systems. This is the case, for instance, of `ccomp:reported`, used to describe reported speech, both in poetry and in prose data. Very low performances emerge particularly also with identifying `parataxis` on prose texts: the highest score is 22.22, achieved by THIVLVC_1, whereas all other systems score 0. On poetry data, results are still very low – the highest score being nearly 30 achieved by OmnesFlores_2 – but only one system exhibits 0 (OmnesFlores_1). Finally, a run from systems leveraging LLMs (i.e., THIVLVC_1) predict a deprel not used in Latin data, namely `cc:preconj`.

The best performances are achieved on deprels describing case marking (`case`) – all above 90 –, objects (`obj`) – ranging from nearly 70 for poetry to around 95 for prose –, adjectival modifiers (`amod`) – ranging from around 65 on poetry to nearly 100 on prose –, determiners (`det`) – showing a range of results slightly lower to what emerges for adjectival modifiers –, as Table 10 shows.

Overall, some deprels exhibit a moderate variation in performance (within a range of 5 percentage points): they are `nsubj`, `obj`, `amod`, `det`, `case`. A higher variation (approximately 15 percentage points) is observed for `obl`, `advmod`, and `nmod`. A consistent drop (more than 15 percentage points) emerges for `advcl` and `acl`.

7. Conclusions

The Dependency Parsing task within EvaLatin 2026 has contributed to advancing the previous state of the art: indeed, almost all partecipating systems outperform the baseline. Furthermore, in the 2024 edition, the best participant performances in the same task are lower than those achieved in the current edition. The comparison is particularly evident for poetry, as in both campaigns the test data is drawn from Senecan tragedies. More specifically, in 2024, the best system on poetry reached 74.53 F1 on CLAS without subtypes, and 75.75 F1 on LAS without subtypes (see Tables 6 and 7 in (Sprugnoli et al., 2024)). In the current edition, the best system reaches 76.08 on CLAS without subtypes, and 77.60 F1 on LAS without subtypes (see Table 4 and Table 6, respectively), with an increase of about 2 percentage points.

Nevertheless, there is still room for improvement. More specifically, as (Pommeret et al., 2026) observe, Latin UD treebanks show a degree of variation in annotation, leading to suboptimal performance of automatic models. The diversity of the available data remains an issue – that is raised for several other languages in the UD community;

currently, no strategies are able to fully address this obstacle to the optimal training of automatic systems.

From this overview, poetry emerges as more difficult than prose for parsing, as Stymne (2026) and Matsuda and Asahara (2026) also remark. A thorough investigation on the reasons for this discrepancy lies beyond the scope of this report. Future work might start exploring the impact of metrics on this aspect. In the present and in the previous edition of EvaLatin, only (Senecan) tragedy has been proposed as test data. Whether other genres within poetry do prove to be difficult for automatic parsers remains object for future research.

8. Acknowledgements

The authors thank Giovanni Moretti for customizing the scorer.

9. Bibliographical References

Valeria Boano, Eleonora Litta, and Matteo Romanello. 2026. Overview of the Named Entity Recognition Task at EvaLatin 2026. In *Proceedings of the Fourth Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2026)*, Palma, Mallorca (Spain). ELRA.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 149–164, New York City. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308. Retrievable at https://direct.mit.edu/coli/article/47/2/255/98516/Universal-Dependencies.

Joseph Denooz. 2004. Opera Latina : une base de données sur internet. *Euphrosyne*, 32:79–88.

J. G. Fitch. 2018. *Tragedies, Volume I: Hercules. Trojan Women. Phoenician Women. Medea. Phaedra*. Loeb classical library 62. Harvard University Press, Cambridge (MA).

Federica Gamba and Daniel Zeman. 2023a. Latin morphology through the centuries: Ensuring consistency for better language processing. In *Proceedings of the Ancient Language Processing Workshop*, pages 59–67, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Federica Gamba and Daniel Zeman. 2023b. Universalising Latin Universal Dependencies: a harmonisation of Latin treebanks in UD. In *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*, pages 7–16, Washington, D.C. Association for Computational Linguistics.

Hiroshi Matsuda and Masayuki Asahara. 2026. Extending omnes flores for the EvaLatin 2026 Dependency Parsing Tasks. In *Proceedings of the Fourth Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2026)*, Palma, Mallorca (Spain). ELRA.

Joakim Nivre and Chiao-Ting Fang. 2017. Universal Dependency evaluation. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 86–95, Gothenburg, Sweden. Association for Computational Linguistics.

Marco Passarotti. 2019. The project of the index thomisticus treebank. In Monica Berti, editor, *Digital Classical Philology*, pages 299–320. De Gruyter Saur, Berlin, Boston.

Luc Pommeret, Thibault Wagret, and Jules Deret. 2026. THIVLVC: Retrieval Augmented Dependency Parsing for Latin. In *Proceedings of the Fourth Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2026)*, Palma, Mallorca (Spain). ELRA.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Rachele Sprugnoli, Federica Iurescia, and Marco Passarotti. 2024. Overview of the EvaLatin 2024 evaluation campaign. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 190–197, Torino, Italia. ELRA and ICCL.

Rachele Sprugnoli, Marco Passarotti, Flavio Massimiliano Cecchini, Margherita Fantoli, and Giovanni Moretti. 2022. Overview of the EvaLatin 2022 evaluation campaign. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 183–188, Marseille, France. European Language Resources Association.

Rachele Sprugnoli, Marco Passarotti, Flavio Massimiliano Cecchini, and Matteo Pellegrini. 2020. Overview of the EvaLatin 2020 evaluation campaign. In *Proceedings of LT4HALA 2020 -*

1st Workshop on Language Technologies for Historical and Ancient Languages, pages 105–110, Marseille, France. European Language Resources Association (ELRA).

Milan Straka, Jan Hajič, and Jana Straková. 2016. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297, Portorož, Slovenia. European Language Resources Association (ELRA). Retrievable at https://aclanthology.org/L16-1680.

Sara Stymne. 2026. UppsalaNLP at EvaLatin 2026: Multilingual parsing for Latin. In *Proceedings of the Fourth Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2026)*, Palma, Mallorca (Spain). ELRA.

THIVLVC: Retrieval Augmented Dependency Parsing for Latin

Luc Pommeret[1], **Thibault Wagret**[2], **Jules Deret**
[1]Université Paris-Saclay, CNRS, LISN, 91400, Orsay, France
[2]École Normale Supérieure de Lyon, HISOMA, 69007, Lyon, France
pommeret@lisn.fr, thibault.wagret@ens-lyon.fr, deret.jules@gmail.com

Abstract

We describe THIVLVC, a two-stage system for the EvaLatin 2026 Dependency Parsing task. Given a Latin sentence, we retrieve structurally similar entries from the CIRCSE treebank using sentence length and POS n-gram similarity, then prompt a large language model to refine the baseline parse from UDPipe using the retrieved examples and UD annotation guidelines. We submit two configurations: one without retrieval and one with retrieval (RAG). On poetry (Seneca), THIVLVC improves CLAS by +17 points over the UDPipe baseline; on prose (Thomas Aquinas), the gain is +1.5 CLAS. A double-blind error analysis of 300 divergences between our system and the gold standard reveals that, among unanimous annotator decisions, 53.3% favour THIVLVC, showing annotation inconsistencies both within and across treebanks.

Keywords: Latin Dependency Parsing, Retrieval-Augmented Generation, Universal Dependencies, EvaLatin, Annotation Consistency

1. Introduction

The EvaLatin 2026 Dependency Task (Iurescia et al., 2026) invites participants to parse Latin texts for two genres: Classical poetry with Seneca, and philosophical prose of Thomas Aquinas.

Previous systems have relied on supervised neural models trained on existing treebanks. Such models learn whatever patterns the training data contains, including, inevitably, annotation choices that predate current guidelines. We explore a complementary approach: explicit injection of UD rules into a Large Language Model, allowing it to refine the output of a traditional parser, combined with a retrieval component. The method is simple. Whether it generalizes beyond Latin remains to be seen.

2. Description of the System

Our system is a two-stage pipeline[1]: (1) retrieval of structurally similar sentences from CIRCSE, and (2) generation, where an LLM refines a baseline parse using the retrieved examples and UD guidelines.

Information Retrieval. Given an input sentence, we retrieve the $k = 5$ most similar sentences from the training set of the CIRCSE treebank (762 sentences) using the *structural* retriever. This retriever is based on sentence length and POS n-grams provided in the input data. The similarity function is a weighted combination of normalized length difference and Jaccard similarities over POS bigrams and trigrams (Equation 1).

[1]Code available at https://github.com/l-pommeret/THIVLVC

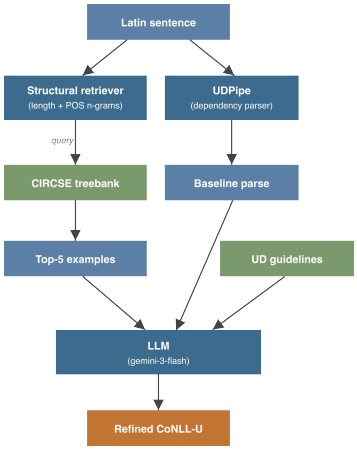


Figure 1: Overview of the THIVLVC pipeline. The input sentence is processed in parallel by the structural retriever (which selects similar examples from CIRCSE) and by UDPipe (which produces a baseline dependency parse). Both outputs, together with the UD guidelines, are passed to the LLM for refinement.

Generation. The retrieved examples, together with the official UD annotation guidelines and the baseline parse from UDPipe, are passed to `gemini-3-flash` (Google DeepMind, 2026). The LLM is prompted to act as a "Latin Chief Annotator": it compares the baseline parse with the retrieved examples and the guidelines, then outputs a refined CoNLL-U block with corrected HEAD and DEPREL columns. The full prompt is given in Appendix A.

We submit two configurations: THIVLVC_1 (LLM + UDPipe + UD guidelines, without retrieval) and THIVLVC_2 (same, but with RAG on CIRCSE). Figure 1 gives an overview of the pipeline.

Proceedings of LT4HALA 2026 @ LREC 2026, pages 219–225.
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

3. Evaluation Protocol

Retrieval strategies. We compared three retrieval strategies. In each case, q denotes the query sentence and s a candidate from the knowledge base.

1. TF-IDF (Baseline). Cosine similarity where $\mathbf{v}_q$ and $\mathbf{v}_s$ are the TF-IDF vectors of word forms for q and s:

$$\text{sim}_{\text{tfidf}}(q, s) = \frac{\mathbf{v}_q \cdot \mathbf{v}_s}{\|\mathbf{v}_q\| \, \|\mathbf{v}_s\|}$$

2. Structural (Length + POS n-grams). A weighted combination of sentence length similarity and POS n-gram Jaccard overlap:

$$\text{sim}_{\text{struct}}(q, s) = 0.33\, f_{\text{len}} + 0.33\, f_{\text{bi}} + 0.34\, f_{\text{tri}} \quad (1)$$

   where $f_{\text{len}} = 1 - \frac{||q|-|s||}{\max(|q|,|s|)}$ is the normalized length similarity ($|q|$ and $|s|$ denote sentence lengths in tokens), $f_{\text{bi}} = J(\text{bigrams}(\text{POS}_q), \text{bigrams}(\text{POS}_s))$ is the Jaccard coefficient $J(A, B) = |A \cap B|/|A \cup B|$ over POS bigrams, and f_{tri} is defined analogously for trigrams.
3. Morphological. Cosine similarity over TF-IDF vectors of concatenated POS and morphological features (`POS|FEATS` per token).

Retrieval metrics. Let $Q = \{q_1, \ldots, q_M\}$ be the test set. For each query q_i, we retrieve $k = 5$ examples $s_{i,1}, \ldots, s_{i,k}$ from the knowledge base. We evaluate retrieval quality with two metrics:

Length Difference: the average absolute difference in sentence length (in tokens):

$$\text{LenDiff} = \frac{1}{Mk} \sum_{i=1}^{M} \sum_{j=1}^{k} \big||q_i| - |s_{i,j}|\big|$$

A small difference (< 2 tokens) indicates similar syntactic complexity.

POS Overlap: the average Jaccard coefficient J (as defined above) over POS tag sets. For a sentence s, let $P(s) = \text{unique}(\text{POS}(s))$:

$$\text{POSOverlap} = \frac{1}{Mk} \sum_{i=1}^{M} \sum_{j=1}^{k} J\big(P(q_i),\, P(s_{i,j})\big)$$

Higher is better ($[0, 1]$).

Generation. We tested three recent LLMs of varying scale and provider: `gemini-3-flash`, `claude-4.5-sonnet`, and `qwen3-72B`.

Benchmarks. We use the CIRCSE treebank for retrieval evaluation and the EvaLatin 2026 test set for generation evaluation. The official metrics are CLAS and LAS (F1), reported both with and without relation subtypes.

Dataset	Strategy	Length Diff	POS Overlap
Prose	TF-IDF	13.24 ± 12.09	0.421 ± 0.150
	Morphological	12.86 ± 14.78	$\mathbf{0.519} \pm 0.138$
	Structural	$\mathbf{0.76} \pm 1.07$	0.512 ± 0.142
Poetry	TF-IDF	11.60 ± 15.95	0.454 ± 0.177
	Morphological	14.51 ± 20.63	0.556 ± 0.163
	Structural	$\mathbf{1.13} \pm 3.02$	$\mathbf{0.601} \pm 0.196$
Combined	TF-IDF	12.26 ± 14.52	0.441 ± 0.167
	Morphological	13.88 ± 18.56	0.541 ± 0.154
	Structural	$\mathbf{0.98} \pm 2.44$	$\mathbf{0.565} \pm 0.177$
CIRCSE	TF-IDF	11.63 ± 14.34	0.471 ± 0.185
	Morphological	15.55 ± 20.49	0.568 ± 0.175
	Structural	$\mathbf{1.06} \pm 1.73$	$\mathbf{0.611} \pm 0.198$

Table 1: Retrieval strategies comparison. Knowledge base: CIRCSE train (762 sentences). $k = 5$ retrievals per query.

4. Results and Analysis

IR results. Table 1 shows that the structural strategy strongly outperforms TF-IDF and morphological retrieval on length difference (< 1.2 tokens on average vs. > 11), while maintaining competitive POS overlap. This confirms that sentence length and POS n-grams are sufficient features for retrieving structurally similar examples.

System results. Table 2 compares our two configurations with the UDPipe baseline (Straka and Straková, 2020). On poetry, `THIVLVC_2` improves CLAS by +17 points over UDPipe (with subtypes). On prose, the gain is more modest (+1.5 CLAS), as UDPipe already performs well on this genre. The RAG component (`THIVLVC_2` vs. `THIVLVC_1`) brings a consistent improvement, especially on prose (+6.9 CLAS with subtypes).

		With subtypes		No subtypes	
System	Genre	CLAS	LAS	CLAS	LAS
`UDPipe`	Poetry	56.94	57.22	57.24	59.74
	Prose	79.41	82.17	83.07	85.21
`THIVLVC_1`	Poetry	72.71	70.36	76.00	76.97
	Prose	74.04	75.72	81.52	82.78
`THIVLVC_2`	Poetry	**74.03**	**72.88**	**76.08**	**77.60**
	Prose	**80.92**	**83.26**	**86.60**	**87.93**

Table 2: `THIVLVC` system comparison. `THIVLVC_1` = LLM + UDPipe + UD Guidelines. `THIVLVC_2` = `THIVLVC_1` + RAG on CIRCSE.

5. Error Analysis

Not all divergences from the gold standard are errors. To better understand our system's behaviour, we conducted a qualitative analysis of cases where predictions differed from reference annotations.

Annotation protocol. We designed a double-blind annotation comparing Gold and `THIVLVC`

outputs (see the interface in Figure 2). Annotators were presented with divergent annotations without knowing which came from which system, and chose among five categories: “Gold is better”, “System is better”, “both wrong”, “undecidable”, and “don’t know”.

Verdict	Ann. 1		Ann. 2	
	n	%	n	%
Gold (human)	137	45.7	110	36.7
THIVLVC (system)	126	42.0	143	47.7
Both wrong	7	2.3	11	3.7
Undecidable	11	3.7	21	7.0
Don’t know	19	6.3	15	5.0

Table 3: Blind evaluation verdicts per annotator (300 items).

Table 3 shows that out of 300 divergences, annotators unanimously agreed on 167 cases, of which 89 (53.3%) favour THIVLVC and 78 (46.7%) the gold standard (Table 4). Tables 5 and 6 in the Appendix break these down by error type and most frequent label confusions. Drawing on UD guidelines and recent work on Latin treebank harmonization (Gamba and Zeman, 2023), we organise the discussion into five categories.

6. Taxonomy of Disagreement

Contradictions between CIRCSE and EvaLatin. According to Table 6, errors of the type `advmod:lmod` instead of `advmod` account for 7 out of 37 (18%) of the main THIVLVC errors. The adverb *unde* illustrates a case of legitimate annotation divergence between the CIRCSE corpus and the EvaLatin 2026 corpus. In CIRCSE,[2] *unde* is annotated `advmod:lmod`, signalling a spatial reference in the underlying description. In EvaLatin 2026, by contrast, *unde* is annotated as a simple `advmod`, without the subtype `:lmod`.[3] Both analyses are linguistically defensible: *unde* may simultaneously carry a spatial origin reading (justifying `:lmod`) and function as a logical connector, and neither interpretation excludes the other. The two corpora simply reflect different (but individually coherent) annotation conventions regarding the scope of the `:lmod` subtype.

Our system reproduces the `advmod:lmod` pattern learned from the CIRCSE training data. When evaluated against EvaLatin 2026, this behaviour is penalized, although it reflects a valid annotation choice. This type of divergence suggests that evaluation metrics should ideally accept both annotations as correct when two conventions are independently defensible, rather than treating the test set annotation as the sole ground truth.

This case illustrates a broader methodological issue in dependency parsing: an annotation shift between training and evaluation datasets. Similar inconsistencies across treebanks have been documented in previous work on UD harmonization, especially among subtypes.[4] When a training corpus uses a subtype that the evaluation corpus omits, parsers may be penalized for faithfully reproducing the annotation patterns present in their training data.

Internal gold inconsistency: `obl` vs `obl:arg`. According to Table 6, errors of the type `obl` instead of `obl:arg` and `obl` instead of `obl:lmod` account for 12 out of 26 (46.2%) of the main EvaLatin 2026 errors. Some discrepancies between THIVLVC and EvaLatin 2026 arise from internal inconsistencies within EvaLatin 2026 itself, particularly in the use of the `obl:arg` subtype (the same goes for `obl:lmod`). EvaLatin 2026 contains divergent annotations for the verb *pertineo* and its prepositional complement introduced by *ad*.[5] The two examples receive different annotations despite their identical syntactic configuration. UD guidelines recommend `obl:arg` for oblique arguments selected by a predicate.[6]

THIVLVC, by contrast, annotates `obl:arg` in both cases, consistently with its training on CIRCSE and with the UD guidelines. The discrepancy therefore appears to result from variation in annotation practice rather than from a parsing error.

Such intra-corpus inconsistencies are distinct from the inter-corpus divergences discussed above. They suggest that EvaLatin 2026 contains a degree of internal annotation noise. Similar annotation fluctuations within the same dataset have been documented in previous work on UD harmonization (Gamba and Zeman, 2023). The distinction between arguments and oblique modifiers is widely recognized as difficult to apply consistently in dependency annotation (de Marneffe et al., 2014).

[2] Sentence `Latin_Tacitus_Ger_prose-163`: *unde annum quoque ipsum non in totidem digerunt species hiems et uer* [...] (“Hence they do not divide the year itself into the same number of seasons: winter, spring [...]”).

[3] E.g. sentence s142: *unde sacra doctrina maxime dicitur sapientia* (“Hence sacred doctrine is called wisdom in the highest sense”).

[4] “The most widespread issues are the `tmod` and `lmod` relation subtypes, as well as comparative clauses.” (Gamba and Zeman, 2023).

[5] In sentence S2, *propositum nostrae intentionis in hoc opere est, ea quae ad christianam religionem pertinent* [...], *religionem* is annotated `obl:arg` of *pertinent*. However, in sentence s147, *cum iudicium ad sapientem pertineat* [...], *sapientem* governed by the same verb with the same preposition, is annotated as bare `obl`.

[6] https://universaldependencies.org/u/dep/obl-arg.html

Clear-cut gold error: `mark` for `case`. Sentence s127 of EvaLatin 2026 illustrates a likely human annotation error.[7] The relation `mark` is reserved in the UD guidelines for subordinating conjunctions and clause-introducing function words (de Marneffe et al., 2014), whereas prepositions governing noun phrases are annotated as `case`. `THIVLVC` correctly annotates `case` in both instances, demonstrating consistency with the UD guidelines.

Similar misattributions of functional relations have been observed in Latin treebanks during harmonization and conversion processes. For instance, (Gamba and Zeman, 2023) report that nominal obliques were sometimes misannotated as `advmod` in PROIEL, while (Cecchini et al., 2018) show that the conversion of the *Index Thomisticus* from Prague Dependency style to UD introduced systematic errors in prepositional functional relations.

Undecidable ambiguity: *hic* (`det` or `advmod:lmod`?). Not all divergences between the parser and EvaLatin 2026 can be attributed to annotation error. Some reflect genuine linguistic ambiguity that the UD guidelines do not fully resolve.[8] Such cases illustrate a structural limitation of single-analysis treebank annotation. Because a gold standard records only one syntactic analysis, alternative interpretations are necessarily excluded, and parsers are penalized when they select a different but linguistically plausible structure. Cases of this kind have long been identified as a general challenge for treebank evaluation (Gamba and Zeman, 2023).

Errors linked with punctuation. Some groups of words presented as sentences in the corpus appear to contain several sentences.[9] If some editions read two sentences, the dependency between the root (*omitte*) and the head of the second clause (*traho*) could be analyzed as `parataxis` rather than `conj`. However, editorial punctuation reflects interpretive choices rather than established facts about the text, and the corpus annotation is not necessarily wrong for diverging from a given printed edition. EvaLatin 2026, however, reads `conj(omitte, traho)`, while `THIVLVC` has `parataxis(omitte, traho)`.

In another sentence, both EvaLatin 2026 and `THIVLVC` appear to make the same interpretation.[10] The lack of punctuation in the corpus can mislead both human annotators and automatic systems. A solution to this problem could be to better segment sentences in the corpus.

Adjectival misclassification: `amod` for `acl`. According to Table 6, errors of the type `amod` instead of `acl` account for 8 out of the 37 most frequent gold-correct cases (21.6%), making it one of the most frequent `THIVLVC` error types. In these instances, the system annotates past participles such as *subiecti*, *scissa* or *lapsi* as `amod` rather than `acl`, which is the annotation expected by the UD guidelines for participial constructions retaining verbal properties. The confusion is not entirely surprising from a linguistic standpoint: in traditional grammar, the participle is conventionally described as the adjectival form of the verb. The case of *notus* illustrates this ambiguity well: originally a participle of *nosco*, it has undergone a degree of lexicalisation and functions in many contexts as a plain adjective ("well-known"), making either analysis defensible. Nevertheless, the systematic nature of the misinterpretation (affecting multiple participles across different sentences) remains difficult to account for. The CIRCSE training data consistently distinguishes `acl` from `amod` in comparable configurations, and the UD guidelines passed to `THIVLVC` explicitly define `acl` as the appropriate relation for participial modifiers of nominals.

7. Limitations

Our approach has several limitations. First, the selection of the LLM was based on informal manual comparison rather than a systematic ablation study. We tested `gemini-3-flash`, `claude-4.5-sonnet`, and `qwen3-72B` on a small set of sentences and selected `gemini-3-flash` on the basis of output quality and cost, but we did not conduct a controlled evaluation across models. This limits the reproducibility and generalizability of our LLM choice. Second, the system relies on a single

[7] In *sed haec doctrina supponit principia sua aliunde, ut ex dictis patet*, the word *ex* is annotated as `mark`. The same preposition is correctly annotated as `case` in sentence s47: *omnis enim scientia procedit ex principiis per se notis*.

[8] In the CIRCSE sentence *hic rapax torrens cadit* [...] (SenPhoen-P-17-8), the form *hic* can be analyzed either as an anaphoric determiner modifying *torrens* (`det`: "this violent torrent here falls") or as a locative adverb attached to *cadit* (`advmod:lmod`: "here a violent torrent falls"). Both readings are syntactically and philologically defensible.

[9] For example, *omitte poenae languidas longae moras mortemque totam admitte quid segnis traho quod uiuo* (SenPhoen-P-17-22, Loeb: "Away with the slow delays of thy long-due punishment; receive death wholly. Why do I sluggishly drag on this life?"). Both the Latin text and the translation distinguish two sentences, separated by a full stop, also in the Teubner edition.

[10] In *flammas potius et uastum aggerem compone in altos ipse me immittam rogos* [...] (SenPhoen-P-17-61), both annotate `conj`(*compone*, *immittam*). The Loeb edition, however, distinguishes the two clauses with a semicolon, suggesting `parataxis`(*compone*, *immittam*).

knowledge base (CIRCSE, 762 sentences). The retriever cannot return useful examples for syntactic constructions absent from this small corpus, which may explain the more modest gains on prose, where UDPipe already performs well. Third, LLM-based parsing is inherently non-deterministic: the same prompt can yield different outputs across runs. We did not measure this variance. A related concern is cost: each sentence requires an API call, making the approach substantially more expensive than a fine-tuned neural parser. Finally, our error analysis, while informative, is limited to 300 items annotated by two annotators with fair inter-annotator agreement ($\kappa = 0.49$ on the binary Gold/`THIVLVC` decision). A larger annotation campaign would be needed to draw firmer conclusions.

8. Conclusion

We presented `THIVLVC`, a retrieval-augmented LLM system for Latin dependency parsing that combines a structural retriever, UD annotation guidelines, and a baseline parse to refine syntactic analysis. The system achieves substantial improvements over UDPipe on poetry and competitive results on prose. Our error analysis highlights a finding that goes beyond system performance: a significant proportion of divergences between parser output and gold annotations stem from annotation inconsistencies, both across corpora (CIRCSE vs. EvaLatin 2026) and within a single dataset. These results underscore the need for continued harmonization efforts in Latin treebanks, in line with previous work (Gamba and Zeman, 2023). More broadly, our analysis suggests that evaluation protocols for dependency parsing could benefit from accepting multiple valid annotations in cases where legitimate annotation conventions diverge or where genuine linguistic ambiguity makes a single gold standard inadequate. In future work, we plan to conduct a systematic comparison of LLMs for this task, expand the retrieval knowledge base to multiple treebanks, and investigate whether fine-tuning a smaller model on LLM-generated corrections could reduce inference cost while preserving quality.

9. Acknowledgements

We thank the EvaLatin 2026 organizers for making the shared task data available, and the CIRCSE Research Centre for the treebank used as knowledge base. This work was carried out at LISN (CNRS, Université Paris-Saclay) and HISOMA (ENS of Lyon).

10. Bibliographical References

Flavio Massimiliano Cecchini, Marco Passarotti, Paolo Ruffolo, Marinella Testori, Lia Draetta, Martina Fieromonte, Annarita Liano, Costanza Marini, and Giovanni Piantanida. 2018. Enhancing the Latin morphological analyser LEMLAT with a medieval Latin glossary. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, pages 88–93, Turin, Italy. CEUR Workshop Proceedings.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4585–4592, Reykjavik, Iceland. European Language Resources Association (ELRA).

Federica Gamba and Daniel Zeman. 2023. Universalising Latin Universal Dependencies: a harmonisation of Latin treebanks in UD. In *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*, pages 7–16, Washington, D.C. Association for Computational Linguistics.

Google DeepMind. 2026. Gemini 3 flash: A large multimodal model. `https://deepmind.google/technologies/gemini/`. Consulté le 10 mars 2026.

Federica Iurescia, Marco Passarotti, and Rachele Sprugnoli. 2026. Overview of the Dependency Parsing Task at EvaLatin 2026. In *Proceedings of the Fourth Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2026)*, Palma, Mallorca (Spain). ELRA.

Milan Straka and Jana Straková. 2020. UDPipe at EvaLatin 2020: Contextualized embeddings and treebank embeddings. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 124–129, Marseille, France. European Language Resources Association (ELRA).

A. Prompt

The LLM receives the following prompt (the variables are filled at runtime):

```
You are the Latin Chief Annotator.
Your goal: Refine and improve the syntax
(HEAD/DEPREL) of the Input Sentence using
best practices.

=== OFFICIAL ANNOTATION GUIDELINES ===
{guidelines_text}
======================================

Here are 5 SIMILAR examples from the
training data.
{example_str}

--- BASELINE (from automatic parser) ---
{latinpipe_context}

--- Input Sentence ---
{conllu_str}

Task:
1. Compare the baseline parse with the
   examples and guidelines
2. Identify any improvements needed in
   HEAD/DEPREL
3. Output your refined version
4. If uncertain, add a comment line
   # needs_council = true

Output ONLY the CoNLL-U block for the
Input Sentence (not the baseline).
```

B. Annotation Interface

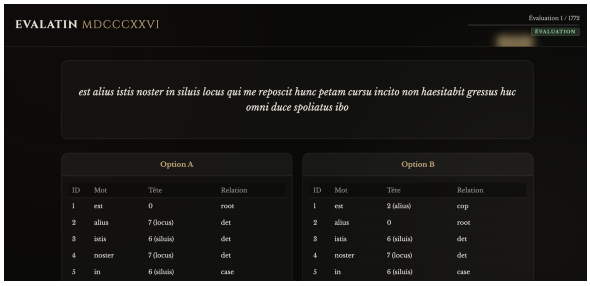


Figure 2: Annotation interface: overview. The Latin sentence is displayed at the top; two anonymized parse options (A and B) are shown side by side with their ID, word form, head, and relation columns.

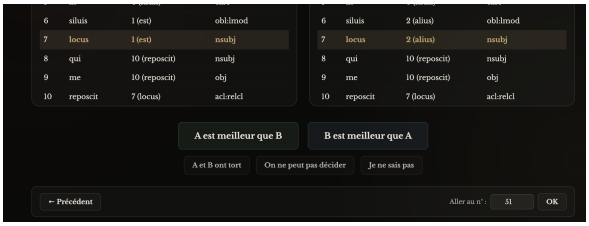


Figure 3: Annotation interface: verdict buttons. Divergent rows are highlighted. Annotators choose among five categories.

C. Detailed Annotation Results

We provide detailed statistics to support the comparison between THIVLVC and the gold standard.

Table 4 shows that, overall, THIVLVC is more often correct than the gold in head-to-head comparisons, at least on items where both annotators reach a consensus.

Table 4: Unanimous consensus: items where both annotators agree (173/300, 57.7%).

Verdict	n	%
THIVLVC better than Gold	89	53.3
Gold better than THIVLVC	78	46.7
Decided	*167*	
Undecidable	3	1.7
Don't know	3	1.7
Total	**173**	

When examining the different error types in Table 5, we observe that the most frequent errors are distributed across head attachment, relation labelling, and subtype confusion.

Table 5: Taxonomy of errors by source (Gold vs. THIVLVC), based on 167 unanimous decided cases.

Error type	**Gold err.**		**THIVLVC err.**	
	n	%	n	%
Wrong head + relation	26	29.2	14	17.9
Subtype confusion	23	25.8	29	37.2
Wrong head only	20	22.5	8	10.3
Wrong relation only	17	19.1	24	30.8
Head + subtype	3	3.4	3	3.8
Total	**89**		**78**	

The label confusions are also informative: Table 6 shows that the most frequent gold errors involve obl, while the most frequent THIVLVC errors involve amod and advcl.

Table 6: Most frequent label confusions by error source.

Gold errors (THIVLVC correct)			**THIVLVC errors (Gold correct)**		
Wrong	Correct	n	Wrong	Correct	n
obl	obl:arg	6	amod	acl	8
obl	obl:lmod	6	advcl	advcl:cmp	8
amod	root	4	advmod	advmod:tmod	7
obl	nummod	4	advmod:lmod	advmod	7
conj	parataxis	3	advmod	discourse	4
obl:agent	obl:arg	3	conj	conj:expl	3

Regarding genre, Table 7 shows that the consensus is fairly evenly distributed between poetry and prose.

Table 7: Consensus results by genre (167 decided cases).

Genre	**Gold**	**THIVLVC**	**Total**	**THIVLVC %**
Poetry	49	50	99	50.5
Prose	29	39	68	57.4
All	**78**	**89**	**167**	**53.3**

Table 8 shows that inter-annotator agreement is reasonably solid, with a Cohen's κ of 0.493 on head-to-head Gold vs. THIVLVC items.

Table 8: Inter-annotator agreement on 300 blind evaluation items. κ is Cohen's kappa.

Metric	**All categories**	**Gold/THIVLVC only**
Items evaluated	300	224
Observed agreement (p_o)	0.577	0.746
Expected agreement (p_e)	0.374	0.499
Cohen's κ	0.324	0.493

Table 9 presents the full confusion matrix between annotators, showing that most disagreements arise when one annotator selects Gold and the other selects THIVLVC.

Table 9: Confusion matrix between annotators (300 items). Rows = Ann. 1, Columns = Ann. 2.

	Gold	**THIVLVC**	**Both**	**Undec.**	**DK**	**Total**
Gold	**78**	37	6	8	8	137
THIVLVC	20	**89**	3	10	4	126
Both	3	4	0	0	0	7
Undec.	3	5	0	**3**	0	11
DK	6	8	2	0	**3**	19
Total	110	143	11	21	15	300

Overview of the Named Entity Recognition Task at EvaLatin 2026

Valeria Irene Boano, Eleonora Litta, Matteo Romanello
KU Leuven, Università Cattolica del Sascro Cuore, Milano, University of Zurich
Oude Markt 13, 3000 Leuven, Largo Gemelli 1, 20123 Milan, Rämistrasse 71, 8006 Zürich
valeria.boano@kuleuven.be, eleonoramaria.litta@unicatt.it, matteo.romanello@uzh.ch

Abstract

This paper describes the organisation and results of the Named Entity Recognition and Classification (NERC) shared task, conducted as part of EvaLatin 2026. The fourth edition of this evaluation campaign for Natural Language Processing on Latin features two shared tasks, i.e. Dependency Parsing and NERC. After introducing the objective of the task and presenting the Ancient Named Entities Special Interest Group, which aims to address the specific challenges that this task presents, this overview details the annotation tagset, the data provided to the participants and their format. The evaluation metrics and the scorer are also described. Finally, the methodology used by each participating team and their results are presented and discussed.

Keywords: Latin, evaluation, named entities recognition

1. Introduction

EvaLatin 2026 marks the fourth iteration of the evaluation campaign dedicated to the benchmarking and advancement of Natural Language Processing (NLP) tools for the Latin language. Continuing the trajectory established by the inaugural edition in 2020 (Sprugnoli et al., 2020) and subsequent campaigns in 2022 (Sprugnoli et al., 2022) and 2024 (Sprugnoli et al., 2024), EvaLatin 2026 was organised as a core component of the Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA2026), co-located with the 2026 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2026).[1] The 2026 edition of EvaLatin was structured around two primary scientific challenges: Dependency Parsing (Iurescia et al., 2026) and Named Entity Recognition and Classification (NERC).[2] This report specifically details the organisation, dataset characteristics, and performance results of the NERC task, providing a critical analysis of the current state-of-the-art in identifying and categorising entities within Latin corpora. The shared task provided participants with standardised test data, which have been made publicly available to promote the iterative improvement of language technologies for Latin, together with comprehensive annotation guidelines, through a dedicated repository.[3] This NERC task also leverages the frameworks of the HIPE (Identifying Historical People, Places and other Entities) 2020 and 2022 campaigns,[4] reusing its file format and evaluation scorer as detailed in Section 3 and Section 4.

2. NERC

Named Entity Recognition and Classification (NERC) is a fundamental task in NLP that involves the automatic identification and classification of proper names in running text into predefined categories such as Persons, Places, Organisations. Although commercial NERC systems have achieved good performance in modern languages such as English, largely driven by massive datasets drawn from news corpora and the web, the application of these technologies to historical languages remains an active and critical frontier.

2.1. Ancient Named Entities

To address specific domain requirements, the Ancient Named Entities Special Interest Group has developed a Named Entity tagset tailored for Ancient Greek and Latin textual materials and their translations.[5] This schema offers labels that are specific to ancient literature while remaining mappable onto standard concepts employed in modern NLP datasets, supporting the creation of consistent annotated datasets of Named Entities in ancient texts by minimising annotator disagreement and ambiguity. The primary applications of this collaborative work is the development of robust machine learning methods and a basic level of interoperability and exchange across projects involving Named

[1] https://lrec2026.info/.

[2] EvaLatin is an initiative organised by the CIRCSE research centre at the Università Cattolica del Sacro Cuore in Milan, Italy.https://centridiricerca.unicatt.it/circse/en.html

[3] https://github.com/CIRCSE/LT4HALA/tree/master/2026/data_and_doc

[4] https://hipe-eval.github.io/HIPE-2022/about.

[5] Available from the SIG - Ancient NER github page.

Proceedings of LT4HALA 2026 @ LREC 2026, pages 226–233.
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

Entities in Ancient Greek and Latin. The compilation of these guidelines aims at facilitating the integration of historical data into broader linguistic infrastructures. From a machine learning perspective, significant challenges arise from the high degree of inflexion of the target language, flexible word order, and complexities regarding tokenisation and lemmatisation. These difficulties are further compounded by the potential absence of capitalisation in some manuscript traditions, inconsistent spelling conventions, and the extreme heterogeneity of historical domains across different eras of Latinity.

The EvaLatin NERC shared task aims to improve a state of the art that is not optimal. With regard to ancient languages, NERC systems currently show different degrees of harmonisation, and Latin is not an exception in this respect. The diversity of the data currently available is an issue we are aware of and that needs to be addressed. This evaluation campaign aims to address this issue, and find strategies to deal with it successfully.

2.2. Tagset

The tagset developed by the Ancient Named Entities Special Interest Group provides a set of semantic labels to classify Ancient Named Entities. The tags provided include two hierarchical levels. Primary first-level tags include *person* (individuals/deities), *place* (geopolitical/geographical), *collective* (groups), *creature* (non-human), and *non-consecutive-entity* (split entity strings). The second-level tags capture nuanced semantic functions, such as *.ancestry*[6] (for familial relationships and lineage, e.g. "sons of Priamus"), *.ethnic* (identifying individuals or groups via membership in a geographically defined group), *.derivative* (used for adjectives derived from names/toponyms where the common noun is not a rigid designator, e.g. "Iberian" or "Platonic"), or *.epithet* (for capitalised nicknames or titles used as unambiguous identifiers, e.g. "Phoebus" for Apollo). First-level tags represent the type of object being designated by the named entity. The usage of first-level tags is intended to be strict, to ensure interoperability. Second-level tags are designed with more consideration for the semantic function of the name: they may point to the way in which a referent is designated (for example, an individual identified with an epithet rather than their own name) or the specifically identifiable subgroup to which an entity belongs (e.g. a political or ethnic collective). They are designed to be used, expanded, or selected for project-specific goals and contexts.

[6]The structural formula for a complete fine-grained annotation is defined by the concatenation of the primary category and the secondary sub-tag, separated by a full stop.

3. Data

This section describes the data provided for the NERC shared task. The EvaLatin data format, for both the sample and test subsets, is based on the one developed for the HIPE 2020 and 2022 NERC shared tasks.[7] Detailed counts by entity type can be found in Tables 1a and 1b.

3.1. HIPE format and tagging scheme

The HIPE format consists of a simple tab-separated column textual format, comparable to the CoNLL format. The annotation scheme follows the IOB (Inside-Outside-Beginning) format.

3.2. Sample Data

The Sample Data consisted of approximately 2,900 annotated tokens from Vergil (*Aeneid*) and Sallust (*Bellum Catilinae*). This set contained 88 mentions, primarily distributed across the *collective* (24), *person* (32), and *place* (27) categories, serving as a baseline for formatting and preliminary model alignment.

3.3. Test Data

The Test Data, totalling 26,308 tokens and 2,665 entity mentions, introduced significant complexity through three authors with distinct stylistic and thematic profiles: Tacitus, Pliny the Elder, and Ovid. These specific authors were selected to ensure the representation of both prose and poetic genres, each offering a rich and diverse array of named entities. A selection from Pliny the Elder's *Naturalis Historia* (Book IV, 1–85), an encyclopaedic treatise covering various aspects of the natural world, was included due to its high density of named entities pertaining to geography and demographic populations. Tacitus's *Germania*, an ethnographic treatise on the Germanic peoples and their territories, was incorporated to evaluate the model's capacity to identify not merely standard locations and persons, but also complex references to ethnicity and geographic affiliation. This serves as a critical benchmark for fine-grained named entity classification. Finally, Books IV and XIV of Ovid's *Metamorphoses* were specifically selected to challenge the system's ability to accurately distinguish between human entities and mythological or divine figures' epithets.

[7]https://hipe-eval.github.io/HIPE-2022/about.

(a) Coarse-grained types

Entity type	Sample	Test
collective	24	422
creature	0	23
event	1	0
language	0	4
misc	4	1
object	0	0
person	32	509
place	27	1704
time	0	1
work	0	1
Total	**88**	**2665**

(b) Fine-grained types

Entity type	Sample	Test
collective	3	7
collective.ancestry	0	13
collective.animal	0	0
collective.astronomy	0	0
collective.derivative	0	11
collective.epithet	0	0
collective.ethnic	20	391
collective.organization	1	0
creature	0	17
creature.animal	0	2
creature.astronomy	0	4
event	1	0
language	0	4
misc	4	1
object	0	0
person	27	364
person.ancestry	2	31
person.author	0	46
person.derivative	1	22
person.epithet	1	31
person.ethnic	0	15
place	25	1647
place.astronomy	0	1
place.derivative	2	56
place.epithet	0	0
time	0	1
work	0	1
Total	**87**	**2665**

Table 1: Entity counts by type in the released sample and test sets.

3.4. File structure

The test file encodes the annotations needed for the NER task, and contains the following lines:

- empty lines, which mark the boundaries between different documents;
- comment lines, starting with the character '#' followed by a space, which provide metadata about the document;
- annotated lines, which contain a token followed by its tab-separated annotations.

The annotation scheme originates from the CLEF-HIPE-2022 Shared Task and consists of 10 different columns, 8 of which contain detailed named entity annotation tags. However, for the evaluation phase of this specific shared task, only two columns are relevant, i.e. NE-COARSE-LIT and NE-FINE-LIT (as detailed in the list below). The tagset used corresponds to the one described in Section 2.2. In the HIPE format, each column contains specific information. The columns used for the task are indicated in bold:

1. **TOKEN**: the annotated token;
2. **NE-COARSE-LIT**: the coarse type (IOB-type) of the entity mention token, according to the literal sense;
3. NE-COARSE-METO: the coarse type (IOB-type) of the entity mention token, according to the metonymic sense;
4. **NE-FINE-LIT**: the fine-grained type (IOB-type.subtype) of the entity mention token, according to the literal sense;
5. NE-FINE-METO: the fine-grained type (IOB-type.subtype) of the entity mention token, according to the metonymic sense;
6. NE-FINE-COMP: the component type of the entity mention token;
7. NE-NESTED: the coarse type of the nested entity (if any);
8. NEL-LIT: the Wikidata Qid of the literal sense, or 'NIL' if an entity cannot be linked. Rows without link annotations have value '_';
9. NEL-METO: the Wikidata Qid of the metonymic sense, or 'NIL';
10. **MISC**: a flag which can take the following values:
    - NoSpaceAfter, indicating the absence of white space after the token; this information can be used to rebuild the original dataset;
    - EndOfLine, indicating the end of a layout line;
    - EndOfSentence, indicating the end of a sentence;

- Partial-START:STOP, marking the character range of a mention that covers only part of a token (most commonly the enclitic *-que*). The offsets are zero-based and follow Python slice semantics, where "abcd"[1:3] yields "bc";
- Non-specified values are marked by the underscore character "_".

For the evaluation phase, the test data have been provided in one single file adhering to the IOB format presented above, where the true NE values were "masked" and represented by underscores in every annotation column.

4. Tasks and Evaluation

4.1. Tasks

The EvaLatin NERC Shared Task focuses on two task types:

- Task 1 - NERC Coarse-grained: this task includes the recognition and classification of entity mentions according to coarse-grained types, where only first level categories should be recognised;
- Task 2 - NERC Fine-grained: this task includes the recognition and classification of entity mentions according to fine-grained types, where also second level categories should be recognised.

4.2. Metrics

NERC is evaluated in terms of macro and micro Precision, Recall, F1-measure. Two evaluation settings are considered: strict (exact boundary matching) and fuzzy (relaxed boundary matching, e.g. at least one token overlap). Each column is independently evaluated according to the following metrics:

- Micro average P, R, F1 at entity level (not at token level), i.e. consideration of all true positives, false positives, true negatives, and false negatives over all documents:
  - strict and fuzzy;
  - separately per type and cumulative for all types.
- Document-level Macro average P, R, F1 at entity level (not on token level), i.e. average of separate micro evaluation on each individual document:
  - strict and fuzzy;
  - separately per type and cumulative for all types.

Our definition of "macro" differs from the usual one, and macro measures are computed as aggregates on document-level instead of entity-type level. Specifically, macro measures average the corresponding micro scores across all the documents. This means that they reflect variance in document length but are insensitive to class imbalances. Note that in the strict scenario, predicting wrong boundaries leads to heavy penalty of one false negative (entity present in the gold standard but not predicted by the system) and one false positive (entity predicted by the system but not present in the gold standard). Although this may be severe, we keep this metric in line with CoNLL standards and refer to the fuzzy scenario if the boundaries of an entity are considered as less important.

4.3. Scorer

EvaLatin NERC evaluations were conducted using the CLEF-HIPE-2020-scorer,[8] a Python module for evaluating NERC and Entity Linking (EL) systems, developed and used in the context of the HIPE shared tasks on NE processing on historical documents. The scorer evaluates at the entity level, whereby entities (most often multi-words) are considered as the reference units, with a specific type as well as a token-based onset and offset. There are different evaluation regimes depending on how strictly entity type and boundary correctness is judged. The scorer provides strict and fuzzy evaluation regimes, where "strict" requires exact match of both entity type and entity boundaries, and "fuzzy" requires exact match of entity type and at least one token overlap. System performance has been computed, reported and published in terms of micro and macro P, R, and F1 for each Task. For each task, the systems have been ranked according to their F1 scores.

5. Results and Discussion

5.1. Team results

Three teams took part in the task: **argo-navis**, **KULeuven** and **uOttawa**. Argo-navis and uOttawa both submitted two runs for each task, while KULeuven submitted one run.

The **argo-navis** team approached the task with a zero-shot cross-lingual strategy (Ripoll-Alberola, 2026). The openly available GliNER2 model[9] (Zaratiana et al., 2025), an extension of the original GliNER model (Zaratiana et al., 2024), was used. GliNER concatenates task specifications with input text, jointly computes contextualised span and task

[8] https://github.com/hipe-eval/HIPE-scorer

[9] https://huggingface.co/collections/fastino/gliner2-family.

```
TOKEN	NE-COARSE-LIT	NE-COARSE-METO	NE-FINE-LIT	NE-FINE-METO	NE-FINE-COMP	NE-NESTED	NEL-LIT	NEL-METO	MISC
# evalatin2026:version = v0.1
# evalatin2026:title = Aeneid
# evalatin2026:original_source = data/preparation/corpus/la/curated/aeneid1-110_lat.xmi
# evalatin2026:license = CC-BY
# evalatin2026:document_id = aeneid1-110_lat
# evalatin2026:author = Vergil
# evalatin2026:applicable_columns = TOKEN NE-COARSE-LIT NE-FINE-LIT MISC
Arma	O	_	O	_	_	O	_	_	_
virumque	O	_	O	_	_	O	_	_	_
cano	O	_	O	_	_	O	_	_	NoSpaceAfter
,	O	_	O	_	_	O	_	_	_
Troiae	B-place	_	B-place	_	_	O	_	_	_
qui	O	_	O	_	_	O	_	_	_
primus	O	_	O	_	_	O	_	_	_
ab	O	_	O	_	_	O	_	_	_
oris	O	_	O	_	_	O	_	_	NoSpaceAfter
Italiam	B-place	_	B-place	_	_	O	_	_	NoSpaceAfter
```

Figure 1: Example of output file format.

embeddings via a shared encoder and feedforward networks, then scores each span-task pair with a dot-product sigmoid to predict entity probabilities. The team submitted two runs for each task, using two variants of the model: the multilingual[10] and the monolingual large.[11] Since GliNER accepts optional natural language descriptions for each task, the model was provided with the tag descriptions extracted from the annotation guidelines, after removing the examples to adapt them to the zero-shot approach. As GliNER is not trained on Latin texts, a cross-lingual approach was required: the input data was first translated into English using the deep-translator Python package[12] to query the Google Translate API. The results were then aligned back to the original Latin text using simalign (Jalili Sabet et al., 2020), using the embeddings of UGARIT/grc-alignment (Yousef et al., 2022). The team achieved third place in Task 1 (NERC Coarse-grained) and second place in Task 2 (NERC Fine-grained). The results showed different performance according to which variant of the model was used. The large model achieves higher precision, whereas the multilingual model delivers higher recall and a better performance on the fine-grained task, which in any case remains more critical.

The **KU_Leuven** team participated proposing a methodology that relies on a substitution-based probing strategy to identify entities within a text (Maria Mihaela Trusca and de Cruys, 2026). The core procedure involves taking a sentence and sequentially replacing individual words with predefined "probes", representative candidate entities of specific target types, such as "person" or "place". To quantify compatibility, the system calculates a Pseudo-Log-Likelihood (PLL) score, which measures how naturally a probe fits into the surrounding textual context. If the PLM predicts a natural fit, the original word is inferred to be functionally similar and likely an entity of that predefined type. These scores are then normalised across entity types. The primary language model utilised was LaBERTa, a RoBERTa-based encoder pretrained on Latin corpora, as preliminary tests showed it outperformed LatinBERT. To train the NER models, the team combined five datasets, one of which was generated by translating existing multilingual HIPE datasets into Latin using the Google Translate API. To align the NER labels, the researchers used a two-step procedure that balanced semantic similarity (via LaBERT embeddings) and orthographic similarity (via Levenshtein distance). The KULeuven team sent one run for each task. In Task 1 (NERC Coarse-grained), the team used a fine-tuned LaBERTa model integrated with the PLL-based auxiliary decoding signal, ranking second among competitors both in the strict and in the fuzzy evaluation. In Task 2 (NERC Fine-grained), the team applied their PLL-based scoring method in a purely zero-shot setting, utilising defined lists of probes for all category pairs, and achieved third place.

Team **uOttawa** addressed the task by leveraging commercial Large Language Models (LLMs), specifically Gemini-2.5-pro and Claude-sonnet-4-5, via prompt engineering (Chan, 2026). Their methodology involved creating prompt templates that contained a task overview, specific requirements, descriptions of the entity classes, and output guidelines mandating the Inside-Outside-Beginning (IOB) format. The team experimented with both zero-shot and few-shot learning strategies. They determined that few-shot prompting consistently outperformed zero-shot approaches because it provided essential contextual depth, grammatical cues, and domain knowledge. For their few-shot templates, they embedded three randomly sampled Latin sentences alongside their correct entity anno-

[10] https://huggingface.co/fastino/gliner2-multi-v1.

[11] https://huggingface.co/fastino/gliner2-large-v1.

[12] https://github.com/nidhaloff/deep-translator.

tations, drawn from a supplementary dataset. The target text was processed sentence by sentence, and light post-processing was applied to the LLM outputs to clean up formatting and automatically classify unlabelled tokens as non-entities. During preliminary testing, uOttawa discovered a distinct divergence in the strengths of their chosen models. They found that Claude-sonnet-4-5 performed best on the coarse-grained task, likely because it is adept at confidently assigning probabilities across a limited set of broad options. Conversely, Gemini-2.5-pro excelled at the fine-grained subtask, as it seemingly utilizes a more cautious predictive approach that is advantageous when navigating large label spaces with subtle semantic distinctions. Consequently, they submitted their official few-shot runs as 'nerc_1' (utilising Gemini) and 'nerc_2' (utilising Claude). A notable characteristic of uOttawa's results was that recall consistently exceeded precision across all models, prompting techniques, and subtasks. The team attributed this to an over-generation tendency, where the English-dominant pre-training data of the LLMs caused them to mistakenly flag Latin common nouns as proper entities if those words function as names in modern derivative languages (e.g., misidentifying the Latin common noun "gloria" as a name). Furthermore, while their few-shot method proved highly effective overall, the researchers acknowledged that the performance drop observed during the fine-grained task indicates that highly specialised, historically contextualised classical labels remain a significant hurdle for generalised commercial LLMs.

5.2. Comparative results

Limited availability of data shaped the approaches adopted by participants to this shared task. The absence of a sizeable training set meant that approaches relying on fine-tuning of transformer-based models were not applicable. As a result, the submitted systems use techniques like zero-shot, few-shot prompting, as well as cross-language transfer, to circumvent this challenge of limited data availability. With the exception of KULeuven, none of the other participating teams made substantial usage of data augmentation and synthetic data generation. The official results reveal that LLM-based few-shot prompting currently defines the upper bound for Latin NER, while probing-based and cross-lingual systems provide valuable, more computationally efficient alternatives.[13]

[13]Only Coarse and Fine Strict ranked results are reported here, for a comprehensive list of results for all teams check the results document in the LT4HALA repository.

System	P	R	F1
uOttawa_nerc-coarse_2	0.899	0.917	0.908
uOttawa_nerc-coarse_1	0.876	0.914	0.895
KULeuven_nerc-coarse_1	0.736	0.694	0.714
argo-navis_nerc-coarse_2	0.575	0.498	0.534
argo-navis_nerc-coarse_1	0.552	0.5	0.525

Table 2: NERC coarse strict

System	P	R	F1
uOttawa_nerc-fine_1	0.841	0.89	0.865
uOttawa_nerc-fine_2	0.841	0.871	0.856
argo-navis_nerc-fine_1	0.466	0.327	0.384
argo-navis_nerc-fine_2	0.491	0.218	0.302
KULeuven_nerc-fine_1	0.131	0.128	0.129

Table 3: NERC fine strict

The uOttawa system dominated both subtasks, with the Claude-sonnet-4-5 run (nerc_2) achieving the highest Coarse-Strict score. This performance confirms that generative LLMs, when provided with explicit formatting and semantic guidance, can effectively bridge the low-resource gap. The KULeuven results highlight a critical methodological distinction: while they were highly competitive in the Coarse task, their Fine Strict score (0.129) was a zero-shot result using only the PLL method without fine-tuning, as no guideline-compliant training data was available for the fine-grained subcategories. Similarly, argo-navis demonstrated that cross-lingual pipelines can achieve a viable baseline (F1 0.534), although they face precision drops in fine-grained tasks. Specifically, this model tends to favour the coarse categories, while rarely committing to the second-level distinctions. The results of the evaluation campaign show that the fine-grained task remains the most challenging. Persons and places are the best recognised categories across all three teams' attempts, reflecting their predominance in most NER multilingual pretraining data. The *creature* category was consistently poorly identified. Table 4 lists F1 performances per team per highest frequency first level entities.

Fine-grained categories such as *.epithet*, *.ancestry*, and *.author* posed a significant hurdle. On many occasions, models yielded zero scores here, illustrating that identifying that a token is a name is significantly easier than determining its semantic role (e.g., distinguishing a person mentioned as an author from a person mentioned as an ancestor). The guidelines defined the *.ethnic* tag for persons or collectives identified by their membership in a geographically or ethnically defined group, and the *.derivative* tag for adjectives derived from toponyms, personal names, or group names. The performance on these categories varied drastically

Entity Label	KULeuven	argo-navis_1	argo-navis_2	uOttawa_1	uOttawa_2
PERSON	0.663	0.43	0.568	0.839	0.855
PLACE	0.774	0.539	0.549	0.923	0.942
COLLECTIVE	0.576	0.684	0.507	0.899	0.88
CREATURE	0	0.475	0.333	0.333	0.318

Table 4: Strict F1 Scores by Team on First Level Entities

depending on the computational methodology employed by the participating teams. Team uOttawa achieved the most robust performance on derived adjectives by employing a few-shot prompting strategy using commercial Large Language Models (LLMs). Table 5 lists F1 performances per team per highest frequency second level entities.

A critical reflection on this campaign involves the tension between high performance and scientific reproducibility. The participating systems represent two distinct architectural philosophies: commercial LLMs, with uOttawa system achieving the highest scores by leveraging commercial models like Claude-sonnet-4-5 and Gemini-2.5-pro, versus open and specialised architectures with the argo-navis and KULeuven teams using openly available models such as GliNER2 and LaBERTa. While highly effective, the commercial LLM models operate as "black boxes" with proprietary training data and non-static versions, which still represents a challenge for the long-term reproducibility of results in a scientific context. Conversely, the open source approaches, while yielding lower absolute scores in this iteration, offer a higher degree of transparency and allow for consistent auditing and local deployment by the research community.

6. Conclusions

The EvaLatin 2026 NERC shared task demonstrates that the landscape of Natural Language Processing for historical languages is continuing to evolve. The results indicate that LLM-based few-shot prompting could potentially represent the state-of-the-art for this domain, effectively bridging the gap created by the absence of large-scale, domain-specific training sets. While coarse-grained entity recognition has reached a high level of maturity, the fine-grained classification of semantic roles remains a significant challenge. Specifically, distinguishing between the nuanced functions of entities continues to be a hurdle for both generalised and specialised models. Given the performance trends observed in the EvaLatin 2026 results, it is evident that the low score problem for fine-grained categories is the next major hurdle for Latin NERC. While models can reliably identify that a token is a name, they often fail to assign specific semantic roles like .epithet or .ancestry due to a lack of representative examples in existing corpora. To move beyond the limitations of existing manual annotations, more strategies involving Synthetic Data Generation can be used to "teach" models these subtle distinctions.

7. Acknowledgements

Our sincere thanks go to Margherita Fantoli and Laura Soffiantini for helping out with the compilation and gold standard annotation of the test data for this task. A special mention goes to all the other members of the Ancient Named Entities Special Interest Group, for their continuous feedback on how to better the tagset.

8. Bibliographical References

Callum Chan. 2026. Transfer learning for named entity recognition of classical latin through llm prompting. In *Proceedings of the Fourth Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2026)*, Palma, Mallorca (Spain). ELRA.

Federica Iurescia, Marco Passarotti, and Rachele Sprugnoli. 2026. Overview of the dependency parsing task at evalatin 2026. In *Proceedings of the Fourth Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2026)*, Palma, Mallorca (Spain). ELRA.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Violet Soen Ine de Daele Kevin Verbruggen Maria Mihaela Trusca, Mark Depauw and Tim Van de Cruys. 2026. Contextual probing for low-resource named entity recognition in latin. In *Proceedings of the Fourth Workshop on Language Technologies for Historical and Ancient*

Fine-Grained Category	KULeuven_1	argo-navis_1	argo-navis_2	uOttawa_1	uOttawa_2
COLLECTIVE.ETHNIC	0.016	0.129	0.027	0.910	0.852
PERSON.AUTHOR	0.000	0.125	0.346	0.624	0.538
PLACE.DERIVATIVE	0.000	0.000	0.000	0.739	0.623
PERSON.ANCESTRY	0.056	0.000	0.000	0.638	0.358
PERSON.ETHNIC	0.000	0.000	0.000	0.529	0.200
PERSON.EPITHET	0.030	0.000	0.000	0.500	0.382
COLLECTIVE.ANCESTRY	0.000	0.000	0.000	0.296	0.071
CREATURE.ANIMAL	0.000	0.000	0.000	1.000	0.000

Table 5: F1 Scores by Team for Fine-Grained NERC Categories

Languages (LT4HALA 2026), Palma, Mallorca (Spain). ELRA.

Ripoll-Alberola. 2026. Classificatio sine iactu – that is, zero-shot nerc in latin. In *Proceedings of the Fourth Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2026)*, Palma, Mallorca (Spain). ELRA.

Rachele Sprugnoli, Federica Iurescia, and Marco Passarotti. 2024. Overview of the EvaLatin 2024 evaluation campaign. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 190–197, Torino, Italia. ELRA and ICCL.

Rachele Sprugnoli, Marco Passarotti, Flavio Massimiliano Cecchini, Margherita Fantoli, and Giovanni Moretti. 2022. Overview of the EvaLatin 2022 evaluation campaign. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 183–188, Marseille, France. European Language Resources Association.

Rachele Sprugnoli, Marco Passarotti, Flavio Massimiliano Cecchini, and Matteo Pellegrini. 2020. Overview of the EvaLatin 2020 evaluation campaign. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 105–110, Marseille, France. European Language Resources Association (ELRA).

Tariq Yousef, Chiara Palladino, David J. Wright, and Monica Berti. 2022. Automatic translation alignment for Ancient Greek and Latin. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 101–107, Marseille, France. European Language Resources Association.

Urchade Zaratiana, Gil Pasternak, Oliver Boyd, George Hurn-Maloney, and Ash Lewis. 2025. Gliner2: An efficient multi-task information extraction system with schema-driven interface. *arXiv preprint arXiv:2507.18546*.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. Gliner: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376.

Transfer Learning for Named Entity Recognition of Classical Latin through LLM Prompting

Callum Chan
University of Ottawa
Ottawa, Canada
cchan073@uottawa.ca

Abstract

With the increase in digitized resources of Classical Latin texts and modern breakthroughs of Large Language Models (LLMs), I contribute to ancient language research by participating in EvaLatin 2026. This paper describes Team uOttawa's system description and results for the Named Entity Recognition (NER) shared task. The task is divided into two subtasks: coarse-grained NER with 11 classes and fine-grained NER with 28 classes, each evaluated under strict and fuzzy regimes. Through prompt engineering of commercial LLMs gemini-2.5-pro and claude-sonnet-4-5, I show that the underrepresented ancient Latin language can take advantage of cross-lingual transfer learning by using advancements made by the wider LLM development community. Overall, the methods discussed in this report demonstrate very strong results, placing first in both NER subtasks and achieving the best scores across all evaluation metrics and regimes among all submissions.

Keywords: Named Entity Recognition, Statistical and Machine Learning Methods

1. Introduction

The study of computational linguistics for ancient languages is hampered from the lack of large annotated datasets (Sprugnoli et al., 2020). However, Ancient Natural Language Processing (ANLP) researchers have identified that modern state-of-the-art general purpose Large Language Models (LLMs) trained on vast amounts of multilingual data have great potential to bridge this low-resource gap (Zhong et al., 2026). The ANLP community has contributed resources for Classical Latin NER (Romanello and Najem-Meyer, 2024), alongside work applying modern deep learning methods to automate the task (Erdmann et al., 2019). ANLP practitioners have also leveraged contemporary BERT (Devlin et al., 2019) and embedding-based NLP techniques, yielding promising results. Nevertheless, automated NER of Classical Latin texts remains an open research problem (Bamman and Burns, 2020; Riemenschneider and Frank, 2023; Torres Aguilar, 2022).

Through transfer learning, generative autoregressive LLMs have shown strong capabilities in natural language understanding and generation across languages for which there are limited resources (Zhong et al., 2026). It has also been discussed that tasks involving ancient Latin texts benefit from these advancements (Volk et al., 2024). In particular, prompting ChatGPT has recently been examined as a potential approach to NER for ancient languages (Woodrum et al., 2025).

These studies, as well as my own experiences, lead me to explore prompting strategies on generative LLMs for this shared task. Examining literature on LLM prompting, it is widely understood that few-shot techniques help guide these models to understand given assignments by providing context and examples (Brown et al., 2020). In particular, thoughtful examples lead to improved results in terms of both output format and content across a wide range of tasks through in-context learning (Liu et al., 2021).

By participating in this shared task as Team uOttawa, I present the ANLP community with another avenue for automated NER of ancient texts. My submissions show that commercial LLMs have the potential to accurately identify named entities of Latin text through principled prompting strategies. Team uOttawa places first for all subtasks and achieves the best scores across all evaluation methods for EvaLatin's 2026 NER shared task.

2. Shared Task

The EvaLatin 2026 NER shared task consists of two subtasks and two evaluation regimes. Each subtask charged participants to identify and classify named entities within a corpus of Classical Latin text. The first subtask focused on coarse-grained named entity recognition; a multi-class classification task consisting of 11 broad classes. The second subtask increased the granularity of the classification by introducing subcategories of the 11 classes mentioned previously. This fine-grained multi-class classification consisted of 28 labels. These classes are shown in Table 1. Per the task guidelines, all text is tokenized and each token is given a named entity label in the Inside-Outside-Beginning (IOB) format.

Proceedings of LT4HALA 2026 @ LREC 2026, pages 234–243.
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

Coarse	Fine	
Person	Person	Collective.ancestry
Place	Person.author	Collective.organization
Collective	Person.ancestry	Collective.epithet
Creature	Person.epithet	Collective.derivative
Event	Person.ethnic	Creature
Language	Person.derivative	Creature.animal
NCE	Place	Creature.astronomy
Object	Place.astronomy	Event
Misc	Place.epithet	Language
Time	Place.derivative	NCE
Work	Collective	Object
	Collective.animal	Misc
	Collective.astronomy	Time
	Collective.ethnic	Work

Table 1: Classes of coarse and fine-grained classification subtasks for EvaLatin 2026 Named Entity Recogintion. NCE stands for Non-Consecutive-Entity.

3. Evaluation Metrics

Each subtask was evaluated under a strict and fuzzy evaluation regime. Strict evaluation considered exact matching of entity types and token boundaries, whereas the fuzzy evaluation loosened token boundaries and allowed for mis-identification of named entities so long as one token overlapped with the gold standard. With respect to metrics, entity level, micro precision, recall and F1 measure were used to score submissions.

4. Dataset

A dataset consisting of around 2900 annotated tokens was provided by the shared task organizers as sample data. This text was taken from poetic and prose textual sources and contained 86 named entities belonging to the 'Collective', 'Misc', 'Person' and 'Place' classes.

In addition to the data provided by the organizers, I also collected a corpus of annotated named Latin entities from the works of (Torres Aguilar, 2022) and (Erdmann et al., 2019). In all, this supplemental dataset consisted of 1.48M classified Latin tokens. Due to differences in class naming, this text was pre-processed to align with the tagset used by the shared task. The mapping between supplementary and EvaLatin 2026 tags is illustrated in Table 2. However, the supplemental dataset suffers from limited tagset coverage, containing only 3 of the 11 coarse-grained tags and lacks more detailed, fine-grained classes. This dataset was used both in examples for my prompting strategies from Section 5 and in additional experiments described in Appendix A.1.

The test dataset for official scoring consisted of around 26,000 tokens which were sampled from works from Tacitus, Pliny the Elder and Ovid. This selection was, again, chosen to give systems a diverse array of poetic and prose texts. This test set also contained a wider array of named entities, including more complex references to geographical, demographical, ethnical, mythological and humanities-related entities when compared to the sample data.

Supplementary Tag	EvaLatin 2026 Tag
PERS, PRS	Person
LOC, GEO	Place
GRP	Collective

Table 2: Tagset alignment between supplementary dataset and EvaLating 2026 NER tagset.

5. Description of the system

To tackle each of the subtasks of the EvaLatin 2026 NER shared task, I employed prompting techniques on public instances of gemini-2.5-pro and claude-sonnet-4-5. These models were chosen for their predecessor's strong generative performance and multilingual understanding (Ahuja et al., 2024; Jayakody and Dias, 2024). For each subtask (coarse and fine) two distinct prompting strategies were used: zero-shot learning and few-shot learning. Prompt templates were created for each strategy and each strategy was run against both subtasks. In all, 4 prompt templates were created to perform experiments to address this shared task. Each of the prompt templates included the following sections:

- Task overview, which consisted of the broad task guidelines.
- Task specifics, which consisted of the specific task requirements.
- Types of entities, which listed and described each of the named entity classes.
- Guidelines for output, which specified the desired format of responses.

The primary difference between the subtask prompts lies in the scope of their entity descriptions, with the coarse-grained prompts covering 11 classes and the fine-grained prompts providing explanations for all 28.

With respect to prompting strategies, I employed zero-shot and few-shot techniques. Both techniques contained the same sections, however few-shot prompts also included 3 example sentences of Latin NER in the desired output format. These examples were randomly sampled from the larger supplementary corpus.

The input text was split into sentences and each sentence was given its own prompt. Then, each prompt was fed into the models under consideration

and responses were collected. Some light post-processing of responses was conducted to clean up the formatting and, if a token was not given a label by the model, it was automatically classified as not belonging to a named entity. All 4 prompt templates used in these methods are listed in Appendix A.2.

5.1. Additional Experiments

Additional experiments were performed after the official scoring period of EvaLatin 2026. These methods explore fine-tuning Latin BERT-based models for token classification. They were not submitted for official scoring due to time-constraints of the evaluation period and the limitation of 2 submissions per team. Their details and results are briefly described in Appendix A.1.

6. Results

The results of my experimentation and official submissions show that LLMs and thoughtful prompting can be effectively applied to the task of textual Latin NER. Before the official scoring period of EvaLatin 2026, I performed preliminary assessments of these methods on the sample data provided by the shared task organizers. These results are shown in in Table 3 and were scored using the official HIPE 2020 scorer[1].

Overall, comparing prompt strategies across models, few-shot techniques consistently outperform zero-shot prompts. Furthermore, gemini-2.5-pro generally performs better at fine-grained classification, whereas claude-sonnet-4-5 outperforms in the coarse-grained classification subtask.

Using these observations, I submitted official predictions on the test set using the few-shot prompting strategy for both models, across both subtasks, for evaluation. Submission uOttawa_nerc_1 contains the responses from gemini-2.5-pro and uOttawa_nerc_2 contains the responses from claude-sonnet-4-5. All official scoring was performed by the shared task organizers and the results of my submissions are reported in Table 4. Consistent with preliminary experimental results, gemini-2.5-pro demonstrates superior performance in fine-grained classification, whereas claude-sonnet-4-5 excels at the coarse-grained level.

Both uOttawa_nerc_1 and uOttawa_nerc_2 outperformed all other submissions across precision, recall and F1 measure. This disparity is also presented in Table 4 where next best scores by submissions KULeuven_nerc-coarse_1 and argonavis_nerc-fine_1 are reported. While the gap in performance is modest, around 0.1 to 0.2 for the coarse-grained subtask, it is much more pronounced for the fine-grained subtask with recall and F1 measures lagging by up to 0.5.

These findings demonstrate that careful prompting of commercial LLMs can provide ANLP researchers with powerful ways to automate NER for Classical Latin text.

7. Discussion

Overall, we can see that the results of Team uOttawa's methods across subtasks and evaluation regimes is competitive.

With respect to intra-model performance over the sample data, I note that few-shot techniques outperformed its zero-shot counterparts in nearly every single metric, across both subtasks and for all evaluation regimes. This is consistent across both models. It is shown that the performance gained from the technique of few-shot learning is model agnostic and clearly benefits both LLMs for the task of Latin NER. Zero-shot prompting does not provide enough context to meaningfully activate the models' latent understanding of Latin text when compared to few-shot techniques. The presence of examples that include output format, Latin text, and their annotated named entities guides the model in its understanding of the task. The increase in contextual depth to the prompt (i.e. the starting point) is crucial in the iterative generation loop performed by modern auto-regressive LLMs. Since Classical Latin is a low-resource language, when compared to English, the explicit guidance given by examples helps to alleviate the gap in understanding brought about by the models' weaker internal representation of Latin grammar, morphology and named entity conventions. Examples provide the model with domain cues to help in its classification with respect to understanding what language it should use to perform this task.

Another observation warranting further discussion are the high recall scores. I notice that across both sample and test datasets, coarse and fine-grained classification, gemini and sonnet models, prompting techniques and evaluation regimes, recall scores are consistently higher than precision. These results illustrate my method's capacity to identify named entities with few false negatives. However, this also demonstrates a tendency toward over-generation or over-classification. By disproportionately labeling Latin tokens as named entities, my method minimizes missed detections at the cost of incorrectly flagging non-named entities as named entities. This phenomenon stems from the tendency of English and other modern languages to adopt Latin common nouns as proper names. A primary example is the word 'Gloria'; while modern English categorizes this exclusively as a name, in

[1] https://github.com/hipe-eval/HIPE-scorer

		Coarse			Fine		
Regime	**Method**	**P**	**R**	**F1**	**P**	**R**	**F1**
Strict	gemini zero-shot	0.716	**0.839**	0.772	0.619	**0.805**	0.700
	gemini few-shot	0.737	**0.839**	0.785	0.650	0.770	**0.705**
	sonnet zero-shot	0.783	0.828	0.804	0.667	0.621	0.643
	sonnet few-shot	**0.849**	**0.839**	**0.844**	**0.698**	0.690	0.694
Fuzzy	gemini zero-shot	0.755	0.885	0.815	0.628	**0.816**	0.710
	gemini few-shot	0.788	**0.897**	0.839	0.689	**0.816**	0.747
	sonnet zero-shot	0.815	0.862	0.838	0.679	0.632	0.655
	sonnet few-shot	**0.895**	0.885	**0.890**	**0.756**	0.747	**0.751**

Table 3: Preliminary results of Team uOttawa's methods on the sample annotated dataset provided by the shared task organizers. Top results have been **bolded**, and the next-best scores have been underlined.

		Coarse			Fine		
Regime	**Submission**	**P**	**R**	**F1**	**P**	**R**	**F1**
Strict	uOttawa_nerc_1	0.876	0.914	0.895	**0.841**	**0.890**	**0.865**
	uOttawa_nerc_2	**0.899**	**0.917**	**0.908**	**0.841**	0.871	0.856
	next_best_strict	*0.736*	*0.694*	*0.714*	*0.466*	*0.327*	*0.384*
Fuzzy	uOttawa_nerc_1	0.917	**0.956**	0.936	0.862	**0.913**	**0.887**
	uOttawa_nerc_2	**0.932**	0.950	**0.941**	**0.863**	0.893	0.878
	next_best_fuzzy	*0.794*	*0.749*	*0.771*	*0.523*	*0.367*	*0.432*

Table 4: Official results for Team uOttawa's submissions. Top results have been **bolded**. Each submission contained predictions for both coarse and fine-grained subtasks. The uOttawa_nerc_1 submission employed few-shot prompting on gemini-2.5-pro and the uOttawa_nerc_2 submission employed few-shot prompting on claude-sonnet-4-5. The next best scores from other teams' submission to the shared task are also reported as next_best_strict and next_best_fuzzy from submissions KULeuven_nerc-coarse_1 and argo-navis_nerc-fine_1.

Latin it functions strictly as a common noun meaning 'glory.' Because modern commercial LLMs are trained on vast English and Romance-language corpora, their named entity recognition is inevitably biased toward these derivative languages, leading to the misapplication of modern entity structures over those native to Latin.

Finally, I notice that scores for the fine-grained classification task are poorer across both models, prompting techniques and evaluation regimes when compared to the coarse-grained subtask. This suggests that while few-shot prompting yields measurable improvements, highly nuanced classification tasks involving many labels remain a significant challenge for LLMs. This is an indication that these highly generalized commercial language models are able to classify broad categories of named entities, which may have been present in their training data. However, their understanding of natural language does not translate as well when presented with new categories or unseen finer-grained labels.

In summary, I have discussed that my methods demonstrate the effectiveness of prompt engineering with commercial LLMs for the task of Latin NER. From my results, it has been shown that few-shot techniques guide these models to accurately and precisely identify and classify named entities in Latin.

8. Practical Adoption

The methods described in this paper presents powerful tools to the greater ANLP community. They both achieve impressive results and are widely accessible to those unfamiliar with machine learning or computational linguistic technicalities. A pipeline consisting of few-shot LLM prompting over non-classified Latin text along with a second pass of humans-in-the-loop (HITL) could prove valuable to researchers studying these documents. My results also show that human correctors should concentrate their efforts on identifying tokens misidentified as named entities (because of the high recall scores) and on instances of rare or unusual labels.

9. Future Work

My work demonstrates that there exists powerful tools for automatically classified Latin NER. I also recognize that this task is far from solved and there is room to improve these results. Latin is morphologically complex and textual fine-grained multi-label classification over its tokens remains challenging.

My main methods focus on prompting auto-regressive commercial LLMs for token classification. Future work could include exploring LLM hy-

perparmeter tuning of temperature or top_p. Probing various other open source models and their fine-tuning over Latin text could also lead to improved results. With respect to the prompting techniques, a wider investigation of prompting strategies should be performed. In particular, I express that there could be more intentional ways of choosing examples that are given within prompts. A cosine similarity-based approach as described in (Liu et al., 2022) could benefit classification performance by identifying the most relevant examples for a given input sentence, more so than a random selection as was performed in my methods.

Finally, there has been progress in training BERT based models over Latin text (Bamman and Burns, 2020; Riemenschneider and Frank, 2023; Torres Aguilar, 2022). It has also been shown that the fine-tuning of these architectures still perform competitively for ancient language classification tasks (Sommerschield et al., 2023). Given the increase in digitized annotated Latin resources, more concentrated efforts than my additional experiments to fine-tune these embedding-based models could yield better results at the fraction of the cost it would take to train or fine-tune a generative LLM.

10. Conclusion

To conclude, I have showcased that the underrepresented ancient Latin language benefits from transfer learning of commercial generative LLMs gemini-2.5-pro and claude-sonnet-4-5. Through few-shot prompting of these modern models, Team uOttawa achieves strong performance when compared to all submissions made to the EvaLatin 2026 NER shared task. I encourage ANLP researchers to continue exploring advancements made by the wider computational linguistic community that bridge the low-resource gap for tasks involving Classical Latin.

11. Limitations

My results are limited to the text provided by the shared task organizers. Although the text has been chosen to vary with respect to genre and author, it is far from representative of the greater body of works that comprise all known Latin corpora.

Furthermore, I recognize that my experiments have been performed on a limited number of LLMs, gemini-2.5-pro and claude-sonnet-4-5. Ablation studies on prompt variations or hyperparameter selection and tests for statistical significance were not performed due to the limited time frame of this shared task.

Finally, my prompt formulations were influenced by my prior experience working with LLMs across different domains and for different purposes. These experiences could introduce biases into the prompting strategy and skew the results of my experiments.

12. Bibliographical References

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Mohamed Ahmed, Kalika Bali, et al. 2024. Megaverse: Benchmarking large language models across languages, modalities, models and tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2598–2637.

David Bamman and Patrick J. Burns. 2020. Latin bert: A contextual language model for classical philology.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Alexander Erdmann, David Joseph Wrisley, Benjamin Allen, Christopher Brown, Sophie Cohen-Bodénès, Micha Elsner, Yukun Feng, Brian Joseph, Béatrice Joyeux-Prunel, and Marie-Catherine de Marneffe. 2019. Practical, efficient, and customizable active learning for named entity recognition in the digital humanities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2223–2234, Minneapolis, Minnesota. Association for Computational Linguistics.

Ravindu Jayakody and Gihan Dias. 2024. Performance of recent large language models for a low-resourced language. In *2024 International Conference on Asian Language Processing (IALP)*, pages 162–167. IEEE.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing.

Frederick Riemenschneider and Anette Frank. 2023. Exploring large language models for classical philology. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15181–15199, Toronto, Canada. Association for Computational Linguistics.

Matteo Romanello and Sven Najem-Meyer. 2024. A named entity-annotated corpus of 19th century classical commentaries. *Journal of Open Humanities Data*, 10(1).

Thea Sommerschield, Yannis Assael, John Pavlopoulos, Vanessa Stefanak, Andrew Senior, Chris Dyer, John Bodel, Jonathan Prag, Ion Androutsopoulos, and Nando de Freitas. 2023. Machine learning for ancient languages: A survey. *Computational Linguistics*, 49(3):703–747.

Rachele Sprugnoli, Marco Passarotti, Flavio Massimiliano Cecchini, and Matteo Pellegrini. 2020. Overview of the EvaLatin 2020 evaluation campaign. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 105–110, Marseille, France. European Language Resources Association (ELRA).

Sergio Torres Aguilar. 2022. Multilingual named entity recognition for medieval charters using stacked embeddings and bert-based models. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 119–128, Marseille, France. European Language Resources Association.

Martin Volk, Dominic Philipp Fischer, Lukas Fischer, Patricia Scheurer, and Phillip Benjamin Ströbel. 2024. LLM-based machine translation and summarization for Latin. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 122–128, Torino, Italia. ELRA and ICCL.

Charles Woodrum, Brian Joseph, Antoine Haaker, and Christopher Brown. 2025. Named entity recognition on ancient languages using large language models: An exploratory study. In *International Conference on Human-Computer Interaction*, pages 430–439. Springer.

Tianyang Zhong, Zhenyuan Yang, Zhengliang Liu, Ruidong Zhang, Weihang You, Yiheng Liu, Haiyang Sun, Yi Pan, Yiwei Li, Yifan Zhou, Hanqi Jiang, Junhao Chen, and Tianming Liu. 2026. Opportunities and challenges of large language models for low-resource languages in humanities research.

A. Appendix

A.1. Additional Experiments

A brief exploration of fine-tuning BERT-based models for token classification was performed. Building on the works of (Riemenschneider and Frank, 2023), I fine-tuned their LaBERTa model on the supplemental dataset described in Section 4. This dataset contained 15,593 sentences and an 80-20 train-validation split was performed. A learning rate of 2e-5 was used with a weight decay of 0.01 over 10 epochs. From these epochs the best model was saved for inference. This training was conducted using HuggingFace's Trainer framework [2].

Since the tags of the supplemental dataset only contains 'Person', 'Place' and 'Collective' classes, this method was only used to generate predictions for the coarse-grained classification subtask.

Regime	P	R	F1
Strict	0.735	0.763	0.749
Fuzzy	0.770	0.800	0.785

Table 5: Results of the extra experiments performed by fine-tuning laBERTa on the supplemental dataset. Scores are from the coarse-grained classification subtask.

While not as strong as the results for my main methods, these scores are still quite competitive. With respect to F1 measure, this submission would have placed third behind uOttawa_nerc_2 and uOttawa_nerc_1. In this instance, prompting large commercial LLMs outperforms embedding-based classification.

[2] https://huggingface.co/docs/transformers/en/main_classes/trainer

A.2. Prompt Templates

The following appendix section presents the four prompt templates that were used for my submissions to the EvaLatin 2026 NER shared task. These are shown in Tables 6, 7, 8 and 9.

```
### Task Outline
Your task is to perform coarse grained named entity recognition on latin text. The latin text is already tokenized and your job is to identify tokens that belong to named entities.

—

### Task Specifics
- Named entity recognition is a technique that identifies and classifies entities in written language into categories such as names, organizations, locations, time expressions, quantites and more.
- You are to use the IOB (Inside Outside Beginning) format to clasify each token as either B - the first token of a multi-token entity, I - any subsequent token within the same entity, or O - a token that does not belong to an entity.
- If the token is idenified as I or B, you are then tasked with identifying the type of entity it is. I or B should be followed by a dash '-' then by the type label.
- The input text will contain each token of the sentence on a newline.

—

### Types of Entities (Labels)
The following is the list of the entity labels and their definition:
- Person: Any identifiable single individual, including deities and anthropomorphic mythological figures.
- Place: A politically, culturally, or geographically defined location, including fictional spaces and structures like temples, buildings, specific urban areas (e.g., gymnasia), and houses.
- Collective: A named group of people or other creatures with shared identifiable characteristics on social, intellectual, political, national, family, mythical, or ethnic basis.
- Creature: Mythical or real precisely identifiable non-human, non-anthropomorphic creatures.
- Event: Significant named events identified by a string with a precise boundary.
- Language: Languages and dialects clearly identified as such.
- Non-consecutive-entity: Non-consecutive entities are strings that contain one named entity but are split across the text. These are provisionally tagged as non-consecutive-entity.
- Object: Artifacts or groups of artifacts clearly identified with a name, such as ships, weapons, statues, columns, dedications, etc.
- Miscellaneous: Entities that do not (yet) have a specific first-level tag among those provided.
- Time: Any absolute date or time expression.
- Work: Titles of literary or non-literary works, in any form.

—

### Guidelines for Output:
- For each word in the sentence, output the word followed by its IOB label and entity type where appropriate.
- Use only the set of labels as defined earlier.
- Each word should be on a new line in the same way it was input.
- Do not add any other explanations or text.

—

### Input:
{INPUT_SENTENCE}

### Output:
```

Table 6: Zero-Shot prompt template for coarse-grained NER.

Task Outline
Your task is to perform coarse grained named entity recognition on latin text. The latin text is already tokenized and your job is to identify tokens that belong to named entities.

—

Task Specifics
- Named entity recognition is a technique that identifies and classifies entities in written language into categories such as names, organizations, locations, time expressions, quantites and more.
- You are to use the IOB (Inside Outside Beginning) format to clasify each token as either B - the first token of a multi-token entity, I - any subsequent token within the same entity, or O - a token that does not belong to an entity.
- If the token is idenified as I or B, you are then tasked with identifying the type of entity it is. I or B should be followed by a dash '-' then by the type label.
- The input text will contain each token of the sentence on a newline.

—

Types of Entities (Labels)
The following is the list of the entity labels and their definition:
- Person: Any identifiable single individual, including deities and anthropomorphic mythological figures.
- Place: A politically, culturally, or geographically defined location, including fictional spaces and structures like temples, buildings, specific urban areas (e.g., gymnasia), and houses.
- Collective: A named group of people or other creatures with shared identifiable characteristics on social, intellectual, political, national, family, mythical, or ethnic basis.
- Creature: Mythical or real precisely identifiable non-human, non-anthropomorphic creatures.
- Event: Significant named events identified by a string with a precise boundary. - Language: Languages and dialects clearly identified as such.
- Non-consecutive-entity: Non-consecutive entities are strings that contain one named entity but are split across the text. These are provisionally tagged as non-consecutive-entity.
- Object: Artifacts or groups of artifacts clearly identified with a name, such as ships, weapons, statues, columns, dedications, etc.
- Miscellaneous: Entities that do not (yet) have a specific first-level tag among those provided.
- Time: Any absolute date or time expression.
- Work: Titles of literary or non-literary works, in any form.

—

Guidelines for Output:
- For each word in the sentence, output the word followed by its IOB label and entity type where appropriate.
- Use only the set of labels as defined earlier.
- Each word should be on a new line in the same way it was input.
- Do not add any other explanations or text.

—

Example:

-> Example 1
Input:
{INPUT_EXAMPLE_1}

Output:
{OUTPUT_EXAMPLE_1}

-> Example 2
Input:
{INPUT_EXAMPLE_2}

Output:
{OUTPUT_EXAMPLE_2}

-> Example 3
Input:
{INPUT_EXAMPLE_3}

Output:
{OUTPUT_EXAMPLE_3}

—

Input:
{INPUT_SENTENCE}

Output:

Table 7: Few-Shot prompt template for coarse-grained NER.

Task Outline
Your task is to perform fine grained named entity recognition on latin text. The latin text is already tokenized and your job is to identify tokens that belong to named entities.

—

Task Specifics
- Named entity recognition is a technique that identifies and classifies entities in written language into categories such as names, organizations, locations, time expressions, quantites and more.
- You are to use the IOB (Inside Outside Beginning) format to clasify each token as either B - the first token of a multi-token entity, I - any subsequent token within the same entity, or O - a token that does not belong to an entity.
- If the token is idenified as I or B, you are then tasked with identifying the first level type of entity it is. I or B should be followed by a dash '-' then by the first level tag.
- Furthermore, if the token belongs to a set of second level tags (defined below), the first level tag should be followed by a pont "." then by the second level tag.
- The input text will contain each token of the sentence on a newline.

—

First Level Tags
The following is the list of the first level entity labels, definitions and available second level tags:
- Person: Any identifiable single individual, including deities and anthropomorphic mythological figures.
– .author
– .ancestry
– .epithet
– .ethnic
– .derivative
- Place: A politically, culturally, or geographically defined location, including fictional spaces and structures like temples, buildings, specific urban areas (e.g., gymnasia), and houses.
– .astronomy
– .epithet
– .derivative
- Collective: A named group of people or other creatures with shared identifiable characteristics on social, intellectual, political, national, family, mythical, or ethnic basis.
– .ancestry
– .animal
– .astronomy
– .ethnic
– .organization
– .epithet
– .derivative
- Creature: Mythical or real precisely identifiable non-human, non-anthropomorphic creatures.
– .animal
– .astronomy
- Event: Significant named events identified by a string with a precise boundary.
- Language: Languages and dialects clearly identified as such.
- Non-consecutive-entity: Non-consecutive entities are strings that contain one named entity but are split across the text. These are provisionally tagged as non-consecutive-entity.
- Object: Artifacts or groups of artifacts clearly identified with a name, such as ships, weapons, statues, columns, dedications, etc.
- Miscellaneous: Entities that do not (yet) have a specific first-level tag among those provided.
- Time: Any absolute date or time expression.
- Work: Titles of literary or non-literary works, in any form.

Second Level Tags
The following is the list of second level labels and their description:
- .ancestry (collective.ancestry, person.ancestry): A designation or expression that refers unambiguously to one individual or group of individuals by using a family name, patronymic, matronymic, or other indication of lineage or familial relationships.
- .animal (collective.animal, creature.animal): A type of creature or collective of creatures clearly identifiable with an animal or animal species.
- .astronomy (creature.astronomy, collective.astronomy, place.astronomy): Named stars, groups of stars, constellations, and planets.
- .author (person.author): A person clearly mentioned in relation to works they have authored. This tag may be modified or even omitted for project-specific goals.
- .derivative (collective.derivative, person.derivative, place.derivative): An adjective derived from a toponym, personal name, or group name, used to identify things that are not individuals or collectives (for individuals or collectives, see .ethnic). Only the derivative is annotated, as the common noun in the expression does not act as a rigid designator. The first-level tag depends on the name from which the adjective derives (e.g. “Iberian” will be a place, “Platonic” a person, etc.).
- .ethnic (collective.ethnic, person.ethnic): An ethnonym, demonym, or other word used to identify persons or collectives by means of their membership to a geographically or ethnically defined group. This tag is exclusively used with persons or collectives, as ethnics are mainly used in the ancient world to identify individuals via ethnic memberships (see also our rationale below). For all other uses of adjectives derived from places, use the .derivative subtag.
- .epithet (collective.epithet, person.epithet, place.epithet): A capitalized epithet used to refer unambiguously to one individual, location, or collective, including nicknames, titles, and other appellatives.
- .organization (collective.organization): Collectives identified by precise organizational structures, such as priesthoods, legions, religious, intellectual, or political groups and institutions, and so on.

—

Guidelines for Output:
- For each word in the sentence, output the word followed by its IOB label and entity type where appropriate.
- Use only the set of labels as defined earlier.
- Each word should be on a new line in the same way it was input.
- Do not add any other explanations or text.

—

Input:
{INPUT_SENTENCE}

Output:

Table 8: Zero-Shot prompt template for fine-grained NER.

Task Outline
Your task is to perform fine grained named entity recognition on latin text. The latin text is already tokenized and your job is to identify tokens that belong to named entities.

—

Task Specifics
- Named entity recognition is a technique that identifies and classifies entities in written language into categories such as names, organizations, locations, time expressions, quantites and more.
- You are to use the IOB (Inside Outside Beginning) format to clasify each token as either B - the first token of a multi-token entity, I - any subsequent token within the same entity, or O - a token that does not belong to an entity.
- If the token is idenified as I or B, you are then tasked with identifying the first level type of entity it is. I or B should be followed by a dash '-' then by the first level tag.
- Furthermore, if the token belongs to a set of second level tags (defined below), the first level tag should be followed by a pont "." then by the second level tag.
- The input text will contain each token of the sentence on a newline.

—

First Level Tags
The following is the list of the first level entity labels, definitions and available second level tags:
- Person: Any identifiable single individual, including deities and anthropomorphic mythological figures.
– .author
– .ancestry
– .epithet
– .ethnic
– .derivative
- Place: A politically, culturally, or geographically defined location, including fictional spaces and structures like temples, buildings, specific urban areas (e.g., gymnasia), and houses.
– .astronomy
– .epithet
– .derivative
- Collective: A named group of people or other creatures with shared identifiable characteristics on social, intellectual, political, national, family, mythical, or ethnic basis.
– .ancestry
– .animal
– .astronomy
– .ethnic
– .organization
– .epithet
– .derivative
- Creature: Mythical or real precisely identifiable non-human, non-anthropomorphic creatures.
– .animal
– .astronomy
- Event: Significant named events identified by a string with a precise boundary.
- Language: Languages and dialects clearly identified as such.
- Non-consecutive-entity: Non-consecutive entities are strings that contain one named entity but are split across the text. These are provisionally tagged as non-consecutive-entity.
- Object: Artifacts or groups of artifacts clearly identified with a name, such as ships, weapons, statues, columns, dedications, etc.
- Miscellaneous: Entities that do not (yet) have a specific first-level tag among those provided.
- Time: Any absolute date or time expression.
- Work: Titles of literary or non-literary works, in any form.

Second Level Tags
The following is the list of second level labels and their description:
- .ancestry (collective.ancestry, person.ancestry): A designation or expression that refers unambiguously to one individual or group of individuals by using a family name, patronymic, matronymic, or other indication of lineage or familial relationships.
- .animal (collective.animal, creature.animal): A type of creature or collective of creatures clearly identifiable with an animal or animal species.
- .astronomy (creature.astronomy, collective.astronomy, place.astronomy): Named stars, groups of stars, constellations, and planets.
- .author (person.author): A person clearly mentioned in relation to works they have authored. This tag may be modified or even omitted for project-specific goals.
- .derivative (collective.derivative, person.derivative, place.derivative): An adjective derived from a toponym, personal name, or group name, used to identify things that are not individuals or collectives (for individuals or collectives, see .ethnic). Only the derivative is annotated, as the common noun in the expression does not act as a rigid designator. The first-level tag depends on the name from which the adjective derives (e.g. “Iberian” will be a place, “Platonic” a person, etc.).
- .ethnic (collective.ethnic, person.ethnic): An ethnonym, demonym, or other word used to identify persons or collectives by means of their membership to a geographically or ethnically defined group. This tag is exclusively used with persons or collectives, as ethnics are mainly used in the ancient world to identify individuals via ethnic memberships (see also our rationale below). For all other uses of adjectives derived from places, use the .derivative subtag.
- .epithet (collective.epithet, person.epithet, place.epithet): A capitalized epithet used to refer unambiguously to one individual, location, or collective, including nicknames, titles, and other appellatives.
- .organization (collective.organization): Collectives identified by precise organizational structures, such as priesthoods, legions, religious, intellectual, or political groups and institutions, and so on.

—

Guidelines for Output:
- For each word in the sentence, output the word followed by its IOB label and entity type where appropriate.
- Use only the set of labels as defined earlier.
- Each word should be on a new line in the same way it was input.
- Do not add any other explanations or text.

—

Example:

-> Example 1
Input:
{INPUT_EXAMPLE_1}

Output:
{OUTPUT_EXAMPLE_1}

-> Example 2
Input:
{INPUT_EXAMPLE_2}

Output:
{OUTPUT_EXAMPLE_2}

-> Example 3
Input:
{INPUT_EXAMPLE_3}

Output:
{OUTPUT_EXAMPLE_3}

—
Input:
{INPUT_SENTENCE}

Output:

Table 9: Few-Shot prompt template for fine-grained NER.

Overview of EvaHan2026: The First International Evaluation of Ancient Chinese OCR and Layout Analysis*

Dongbo Wang[1], Dongmei Zhu[1], Jieqiong Li[1], Chang Liu[1], Ruifeng Wu[1], Fan Yang[1], Yue Zhu[2,3,4], Xin Gao[1], Zhixiao Zhao[1], Xue Zhao[1], Zhongzheng Wu[5], Liu Liu[1], Bin Li[2,3,4]

[1] School of Information Management, Nanjing Agricultural University, Nanjing, China
[2] School of Chinese Language and Literature, Nanjing Normal University, China,
[3] Center for Language Big Data and Computational Humanities, Nanjing Normal University, China
[4] Lab of Artificial Intelligence and Innovative Application for Chinese Classics, Nanjing Normal University, China
[5] Artificial Intelligence and Human Languages Lab, Beijing Foreign Studies University, China
{db.wang, liuliu}@njau.edu.cn {dongmei.zhu, lijieqiong, doclc, wuruifeng, yangf, gaoxin, zhaozhixiao, zhaoxue}@stu.njau.edu.cn {wufidel}@bfsu.edu.cn {zhuyue, lib}@njnu.edu.cn

Abstract

Ancient Chinese documents are vital for historical research, necessitating high-precision character recognition and layout analysis for digitization. This paper introduces EvaHan2026, the inaugural international shared task for simultaneous optical character recognition and layout parsing of ancient texts. The evaluation framework comprehensively assesses model performance across diverse calligraphic styles and complex structures, including main body text, interlinear annotations, and illustrations. Among thirteen participating teams, four successfully completed all tasks within the closed track. Experimental results reveal that character recognition accuracy reached 97.36% on engraved texts (Test Set A) and 95.71% on handwritten texts (Test Set C) when accounting for character variants. For layout recognition in complex layouts (Test Set B), the best team achieved a peak mean Average Precision (mAP) of 59.41% and an Intersection over Union (IoU) of 76.38%. Our analysis indicates that calligraphic variability, layout density, and character variants significantly modulate system performance. Consequently, enhancing robustness within complex layouts and developing synergistic models that integrate textual and structural information remain primary challenges for intelligent interpretation of ancient writings .

Keywords: Ancient Chinese, Optical Character Recognition (OCR), Layout Analysis, Shared Task

1. Introduction

Ancient Chinese documents are pivotal for historical and philological research. Their large-scale digitization relies on high-precision optical character recognition (OCR) and layout analysis, which convert scanned images into searchable, analyzable text.

As a flagship international shared task for ancient Chinese language processing, EvaHan has built a full-stack NLP evaluation framework in collaboration with the LT4HALA/ALP workshop over four years. It started with word segmentation and POS tagging in 2022 (Li et al., 2022), expanded to machine translation in 2023 (Wang et al., 2023), added sentence segmentation and punctuation in 2024 (Li et al., 2024), and focused on named entity recognition in 2025 (Li et al., 2025). EvaHan now supports end-to-end evaluation for ancient document processing.

Massive ancient Chinese document archives have been digitized as scanned images, creating a strong demand for robust visual transcription tools (Shi et al., 2023). In this context, EvaHan 2026 for the first time centers its evaluation on OCR and layout analysis, two critical components that determine the upper bound of downstream performance. Standardized evaluation of these modules carries both academic and practical importance.

EvaHan 2026 comprises three decoupled tasks to enable fine-grained assessment: Task A (Engraved Text Recognition), Task B (Complex Layout Analysis), and Task C (Handwritten Text Recognition). Major technical challenges include pervasive character variants, print degradation, free-form handwriting, dense and interleaved layouts, paper aging, and ink bleed-through. Among these, character variants are the primary accuracy bottleneck, as they cause frequent confusions and hinder feature extraction. Accurate variant handling is essential for preserving textual fidelity and ensuring reliable downstream historical research.

Against these challenges, EvaHan 2026 aims to build a standardized, reproducible benchmark for ancient document OCR. It uses automated, objective scoring to ensure fair and consistent model comparison, thereby supporting the long-

* The official datasets and detailed guidelines for the EvaHan evaluation campaigns are available at: https://github.com/GoThereGit/EvaHan.

Proceedings of LT4HALA 2026 @ LREC 2026, pages 244–262.
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

term preservation and scholarly use of cultural heritage.

2. Related Work

2.1 Deep Learning-Based OCR Methodologies

Deep learning has fundamentally advanced OCR systems. Early gradient‑based learning and Graph Transformer Networks (GTN) laid the foundation for end‑to-end trainable document recognition (LeCun et al., 1998). The Convolutional Recurrent Neural Network (CRNN) integrated CNN and RNN features to enable direct sequence recognition without explicit character segmentation, becoming a de facto standard for text recognition (Shi et al., 2016). More recently, OCR‑free frameworks such as Donut directly transformed document images into structured information, alleviating error propagation and excessive computation in traditional pipelines (Kim et al., 2021). These technical breakthroughs provide a solid basis for visual analysis and information extraction from ancient documents.

2.2 Visual Processing for Historical Documents

Research on historical document visual processing has advanced rapidly across pipelines, benchmarks, layout analysis, and multi-task modeling. OCR4all delivered a semi‑automated workflow for 15th–19th century engraved books, achieving low CER but still requiring manual proofreading for complex layouts (Reul et al., 2019). M^{5}HisDoc constructed a large‑scale benchmark with eight thousand images and a multi‑attribute difficulty definition system, yet performance degrades on seriously distorted samples (Shi et al., 2023). HierText built hierarchical layout parsing standards across word, line, and paragraph levels (Long et al., 2022). Xu et al. (2018) designed a multi‑task FCN for pixel‑level segmentation of pre‑modern Chinese handwritten manuscripts. AncientDoc pioneered VLM evaluation for ancient documents, covering tasks from page‑level OCR to knowledge reasoning (Yu et al., 2025). Despite these advances, a unified, international benchmark for simultaneous OCR and layout analysis remains lacking.

2.3 Shared Tasks and Evaluation Platforms

Specialized competitions and platforms have accelerated technological development. The Guangdong‑Hong Kong‑Macao Greater Bay Area (Huangpu) International Algorithm Competition has launched a dedicated track for Ancient Document Image Recognition and Analysis, with supporting resources available at https://github.com/SCUT-DLVCLab/MCS-Bench. However, existing evaluations rarely support decoupled yet unified testing of Engraved Text, Handwritten Text, and Complex Layout Analysis. Few systematically address character variants or compare model performance under closed versus open settings. Building on the established EvaHan series, EvaHan 2026 fills these gaps with a holistic, standardized, and internationally accessible evaluation framework for ancient Chinese document visual processing.

3. Task Description

To systematically address the technical bottlenecks inherent in digitizing heterogeneous ancient Chinese texts, EvaHan 2026 structures its evaluation into three distinct sub-tasks based on visual characteristics and target objectives (summarized in Table 1). This decoupled task design is specifically formulated to quantify model performance across the distinct challenges of block-engraved character recognition, complex layout parsing, and manuscript transcription.

Task A evaluates the model's recognition capability on engraved classical texts across different manuscript styles. Engraved classical texts not only involve character evolution and complex layout arrangements (such as marginal annotations and double-line footnotes), but also commonly suffer from degradation phenomena like missing pages, ink stains, and damaged or missing characters, making their recognition significantly more challenging than standardized modern texts. This task employs Character Error Rate (CER) as the primary evaluation metric, supplemented by F1-score and Normalized Edit Distance (NED). This approach aims to comprehensively evaluate models across multiple dimensions, including recognition accuracy, sequence integrity, and transcription cost.

Task B focuses on Document Layout Analysis, aiming to achieve automated extraction of key semantic elements within complex classical text layouts. The identified objects encompass four core components: text blocks, illustrations, book margins, and seals. The evaluation employs Mean Average Precision (mAP) as the primary metric, while incorporating Intersection over Union (IoU) and F1 score to scientifically quantify the model's overall performance in regional localization accuracy and multi-class recognition tasks.

Task C targets recognizing handwritten ancient texts. Manuscripts exhibit highly individualized characteristics and are rife with random disturbances such as hasty connected strokes, colloquial/variant characters, missing strokes, and later alterations, creating significant technical barriers. This task comprehensively employs CER, F1 score, and NED metrics to overcome algorithmic bottlenecks in automatic transcription of handwritten text, providing foundational

technical support for deep mining of rare unique copies and manuscripts, as well as knowledge graph construction.

4. Dataset

4.1 Data Sources

The EvaHan2026 evaluation task utilizes datasets curated from highly representative classical Chinese texts, aiming to comprehensively cover diverse textual formats and layout characteristics. Both Datasets A and B are sourced from the Siku Quanshu, a monumental collection of immense scholarly value. Dataset A features extensive sampling from the Four Divisions: Confucian Classics, History, Philosophy, and Literature, encompassing diverse fields such as philosophy, history, and literature, and providing abundant engraved text samples. Concurrently, Dataset B is derived from complex pages in these classical texts that feature a mix of text and images.

Dataset C focuses on the more challenging task of handwriting text recognition, featuring sources with significant heterogeneity and cross-genre distribution. These materials can be broadly categorized into four main types. First, historical and governmental archives, such as the Chronicle of Foreign Affairs Arrangements, document Qing dynasty diplomatic affairs. Second, private literary collections and poetry anthologies—exemplified by the Poems of the Escaping Void Master and Collected Works of Master Renjie—showcase highly individualized writing styles. Third, calligraphic manuscripts and ink tracings, such as the Bu Shang Reading Manuscript, represent highly artistic and challenging handwriting. Finally, core Confucian and religious literature encompasses the Thirteen Classics and multiple volumes of Chinese Buddhist Scriptures. This corpus, spanning engraved and handwritten materials as well as official and folk sources, ensures the evaluation task can assess the robustness of OCR algorithms across multiple dimensions in complex real-world scenarios.

4.2 Data Annotation

To ensure the model can learn deep semantic and visual features, this evaluation conducted targeted and manually refined annotation on three sub-datasets. Examples from the three datasets are shown in Table 1.

The annotation team performed character-level annotation and variant character processing for Datasets A and C, achieving bounding box localization and text transcription for images. Addressing the prevalent issue of variant characters in ancient texts, the annotation strategy balanced “character form priority” with “semantic consideration.” The annotation guidelines also standardized and consolidated exceptionally rare variant characters. For example, when the character 𣍐 appeared with its bottom component written as 壬 instead of 王, it was consistently annotated as 呈 in accordance with our unified transcription specifications.

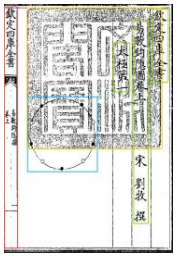
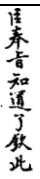
Dataset	Example	Text
Dataset A		{"image_path": "Dataset_A/a_016.jpg", "text": "管民官管匠人打捕諸頭目及諸軍馬使臣人等宣聖廟國家歲"},
Dataset B		{"image_path": "Dataset_B/b_0002.jpg","regions": [{"label": "book_edge", "text": "" "points": [[2, 12], [42, 12], [42, 776], [2, 776]]}, {"label": "image"……
Dataset C		{"image_path": "Dataset_C/c_015.jpg","text": "臣奉旨知道了欽此"},

Table 1: Dataset Sample

For complex layout annotation, Dataset B employs a rectangular region annotation strategy to delineate boundaries for text, images, book margins, and seals. Addressing the unique layout complexity of ancient texts, this study focuses on analyzing the proportion of text-image interleaved data. Approximately 27.46% of the dataset features highly integrated image regions with main text and flexible layouts. Such high-complexity annotations require models to possess not only fundamental object detection capabilities but also the ability to handle hierarchical layout relationships.

For each subtask, we have set up 5,000 images for training and 200 images for test. During the training phase, participants can only access the annotated training set for parameter optimization and internal validation (as detailed in Table 2).

Dataset	Images	Char tokens	Char types
Dataset A training	5,000	83,832	3,241
Dataset A test	200	4,589	1,415
Dataset B training	5,000	-	-
Dataset_B test	200	-	-
Dataset C training	5,000	69,183	4,729
Dataset C test	200	4,728	681
Total	-	162,332	5,551

Table 2: Dataset Statistics

5. Evaluation Metrics

This evaluation encompasses three core tasks: Task A (Engraved Text Recognition), Task B (Complex Layout Analysis), and Task C (Handwritten Text Recognition). To comprehensively and objectively assess model performance, we adopt a suite of well-established metrics tailored to text recognition and object detection tasks.

For text recognition tasks (Tasks A and C), we use five complementary metrics: CER as the primary core metric, alongside Precision, Recall, F1-Score, and NED to capture recognition accuracy, completeness, and overall edit distance efficiency. For layout analysis (Task B), we employ mean Average Precision (mAP@[.5:.95]) as the primary detection metric, supplemented by Micro-average F1, Macro-average F1, and Average Matching IoU to evaluate category-level detection performance, class balance, and pixel-level localization precision, respectively. Detailed definitions of auxiliary metrics and the final comprehensive ranking score can be found in the Appendix.

5.1 Text Recognition Metrics (Task A and Task C)

In the text recognition phase of Tasks A and C, we adopted a character-level evaluation approach. This system primarily quantifies recognition errors by calculating the Edit Distance between the model's predicted text and the ground truth labels. The specific evaluation metrics include the following dimensions:

(1) Character Error Rate (CER)

CER is the core metric of this evaluation, defined as the ratio of the Edit Distance to the length of the reference text:

$$\mathrm{CER}=\frac{\mathrm{Edit\backslash Distance(ref,hyp)}}{|\mathrm{ref}|} \qquad (5\text{-}1)$$

where *ref* is the reference text (ground truth), h*yp* is the predicted text (OCR output), and |*ref*| represents the number of characters in the reference text.

The calculation of the Edit Distance includes three basic operations:

$$\mathrm{Edit(ref,hyp)}=\mathrm{N_{del}}+\mathrm{N_{ins}}+\mathrm{N_{sub}} \qquad (5\text{-}2)$$

Deletion: N_{del}, the number of characters present in the reference text but missing in the predicted text.

Insertion: N_{ins}, the number of characters present in the predicted text but not in the reference text.

Substitution: N_{sub} , the number of characters present in both but different.

(2) Variant Character Matching

Considering the existence of a large number of variant characters in ancient texts, Tasks A and C provide two evaluation modes. The standard mode counts a prediction as correct only when the predicted character matches the reference character exactly; the variant character mode is based on a predefined variant character mapping table, where a predicted character is also counted as correct if it is a visually similar variant of the reference character. Under the variant character mode, the CER calculation will automatically match equivalent characters within the edit distance algorithm. The equivalence relation sim is defined as:

$$c_1 \sim c_2 \Leftrightarrow c_1=c_2 \vee (c_1,c_2)\in \mathrm{V} \qquad (5\text{-}3)$$

where *V* is the set of variant character mappings. During the edit distance calculation, if *ref*[*i*]∼h*yp*[*j*], it is considered a successful match. Specifically, the evaluation code uses a dynamic programming algorithm to calculate the edit distance, introducing variant character judgment during the character alignment phase: when a mapping relationship exists in *V* between the reference character and the predicted character, the two are considered equivalent, and no insertion, deletion, or substitution operations are required. Regarding the selection of variant characters, researchers manually filtered high-frequency variant character pairs from the test set files of datasets A and C to build the variant character mapping table used for the evaluation calculation.

5.2 Text Recognition Metrics (Task B)

Task B adopts object detection evaluation standards to assess the model's ability to localize and classify layout elements (text blocks, illustrations, book margins, and seals). The primary metric used is the Mean Average Precision (mAP).

(1) Mean Average Precision (mAP)

We adopt the mAP@[0.5:0.95] of the COCO evaluation standard (Lin et al., 2014). Specifically, we sequentially calculate the AP at IoU thresholds ranging from 0.5 to 0.95 (with a step size of 0.05), and finally calculate the mean. During this process, the AP calculation at a single threshold refers to the classic 11-point interpolation method, ensuring evaluation smoothness and objectivity by sampling on the precision-recall curve. That is, calculating the mean of average precisions when the IoU threshold goes from 0.5 to 0.95 (step size 0.05):

$$\mathrm{mAP_{[0.5:0.95]}}=\frac{1}{10}\sum_{\mathrm{t}\in\{0.50,0.55,\ldots,0.95\}}\mathrm{AP\ (t)} \qquad (5\text{-}4)$$

The Average Precision (*AP*) at each IoU threshold *t* is calculated via the 11-point interpolation method on the precision-recall curve:

$$\mathrm{AP(t)}=\frac{1}{11}\sum_{\mathrm{r}\in\{0.0,0.1,\ldots,1.0\}}\max_{\mathrm{r'}\geq\mathrm{r}}\mathrm{P\ (r')} \qquad (5\text{-}5)$$

where $P(r')$ is the precision at recall r'.

(2) Intersection over Union (IoU)

For any predicted bounding box B_p and ground truth bounding boxB_g, IoU is defined as:

$$\mathrm{IoU}(\mathrm{B_p},\mathrm{B_g})=\frac{\mathrm{Area}(\mathrm{B_p}\cap\mathrm{B_g})}{\mathrm{Area}(\mathrm{B_p}\cup\mathrm{B_g})} \quad (5\text{-}6)$$

The matching rule uses a greedy algorithm: labels are grouped for each image, the IoU matrix between predicted boxes and ground truth boxes is calculated, and the predicted-ground truth box pair with the maximum IoU is preferentially matched (IoU ≥0.5 is considered a valid match).

5.3 Evaluation Modalities

This evaluation sets up two participation modalities: the Closed Modality and the Open Modality, to adapt to the needs of different technical routes and practical application scenarios.

(1) **Closed Modality**

Under the Closed Modality, participants are strictly limited to using the officially provided pre-trained models and the fixed training set data. This modality aims to evaluate the generalization ability of models under restricted resource conditions, simulating the real-world constraints of limited annotated data and computing resources in actual ancient document digitization scenarios. The authorized pre-trained models allowed in this modality include two VLMs, Qwen2.5-7B-VL and Xunzi-Qwen2-7B-VL, as well as the YOLO series object detection models. Qwen2.5-7B-VL is a multi-modal large language model launched by Alibaba (Bai et al., 2023), possessing strong image understanding and text generation capabilities, and performing excellently in general multi-modal data processing tasks; Xunzi-Qwen2-7B-VL (Wang et al., 2024) is a specialized model optimized for ancient text recognition scenarios, featuring targeted advantages in ancient character recognition and layout understanding; the YOLO series models are classic object detection architectures, widely applied in layout element localization tasks. By restricting model selection and data sources, the closed modality facilitates a fair benchmarking of the effectiveness of different methods, and its evaluation results better reflect the generalization ability of models under data-constrained conditions.

(2) Open Modality

Under the Open Modality, the evaluation does not restrict the use of models and training data, encouraging participants to explore state-of-the-art technical methods. Participants are free to use various multi-modal large models including GPT-4o-mini, LLaVA (Liu et al., 2024), Deepseek-OCR (Lu et al., 2024), etc. They can also utilize publicly available ancient book datasets and synthetic data for training, and may introduce external knowledge bases such as variant character dictionaries and ancient character sets to enhance model performance. This modality aims to push the technical boundaries of ancient book OCR and layout analysis, evaluating the optimal performance of current methods under unconstrained conditions. It is expected that the open modality will achieve higher evaluation scores, but it may also be accompanied by higher computational costs. By comparing the evaluation results of the two modalities, we hope to comprehensively reveal the current technical level of ancient book OCR and layout analysis from the dual dimensions of algorithmic effectiveness and technical upper limits, providing a reference basis for actual deployment.

6. Participants and Results

6.1 Participating Teams

This section statistically analyses the institutional affiliations, participation scale, and model submission features of teams in the EvaHan 2026 competition, based on official application and final submission data. Initially, 41 teams registered, with core data drawn from the organizing committee's registration forms and technical review database. Ultimately, 13 teams successfully submitted their final evaluation results across the three core tasks (Task A for Engraved OCR, Task B for Layout Analysis, and Task C for Handwritten OCR). Among them, 4 teams participated in both the closed and open modalities, while the remaining teams focused solely on the closed track. It is worth emphasizing that, with the singular exception of SL-ISTIS–which completed the model evaluation but failed to provide the corresponding documentation–all other 12 participating teams successfully submitted their required final papers. Participating entities consist primarily of universities and research teams, with the detailed information of each team's model and method submissions presented in Table 3.

6.2 Results

The performance evaluation of this competition was structured around three core tasks: Task A (Engraved Text Recognition), Task B (Complex Layout Analysis), and Task C (Handwritten Text Recognition). To comprehensively assess the systems, we established two evaluation tracks (closed and open modalities) and applied two distinct character matching rules (inclusive and exclusive of variant characters). Owing to the large number of participating teams in this EvaHan edition and the page limit of the paper, only the top-five results under different tracks and tasks are presented for the evaluation results of each participating team. Complete rankings are available in Appendix A.

Visualizing the performance metrics across modalities reveals critical technical insights as

shown in Figure 1. In the closed modality, TJU's specialized architectural design such as the integration of HistLayout-DETR achieved state-of-the-art results in the OCR tasks (Tasks A and C). This demonstrates that task-specific structural innovations can effectively compensate for strict resource constraints. Conversely, WHU-SAI relied on comprehensive multi-stage optimization (LoRA SFT, DPO, GRPO), maintaining dominant positions across the open modality. However, as indicated by the tight convergence of the open and closed trend lines for top-performing systems, leveraging unrestricted external datasets and larger model ensembles yielded only marginal gains in peak performance—and even a slight degradation in Task B for the leading team. Notably, while the closed modality exhibits severe performance drop-offs (indicated by the sharp upward spikes in CER) among lower-ranked teams, the open track maintains a remarkably flat trajectory. This suggests that scaling data and model size primarily functions to elevate the performance floor and ensure system robustness, rather than significantly advancing the absolute technical ceiling. Ultimately, these patterns highlight a diminishing return on simply scaling general multimodal models without introducing corresponding, domain-specific algorithmic innovations.

ID	Team	Affiliation	Close/ Open	Core Models	Key Methods & Data
1	BNU	Beijing Normal University	2/0	Xunzi-VL, Qwen-VL, Doubao-1.5-pro (Vision)	LoRA, linguistic rule post-processing; No external data
2	ENC-PSL	École nationale des chartes, PSL University	2/2	CRNN, PaddleOCR, YOLO series, RT-DETR, Qwen-VL series, GPT-4o, Claude, Gemini	UltraGlyph pipeline, ICL, manual re-annotation; External data used
3	EPHE-PSL	École pratique des hautes études, PSL University	1/1	ANANDASKY VLM, ViT, Qwen3-0.6B, YOLO26	Global attention encoder, line-level transcription, full fine-tuning; External data used
4	FDU	Fudan University	1/0	Xunzi-VL, DocLayout-YOLOv10	SFT, DoRA, DPO, multi-round inference; No external data
5	NEFU	Northeast Forestry University	1/0	Qwen-VL, YOLOv8/11	LoRA, dual-stream YOLO ensemble, VLM semantic filter; No external data
6	NJU	Nanjing University	1/0	Qwen-VL	LoRA, QLoRA, adaptive prompt engineering; No external data
7	QLNU	Qilu Normal University	1/0	Qwen-VL	LoRA, data augmentation, annotation refinement; No external data
8	RUC-MIDU	Renmin University of China; Midu Technology	2/2	Qwen-VL, Xunzi-VL, DocLayout-YOLO	LoRA, background embedding augmentation; No external data
9	SL-ISTIS	Shanghai Library (Institute of Scientific and Technical Information of Shanghai)	1/0	-	No valid paper submitted
10	SYSU	Sun Yat-sen University	2/0	Qwen-VL	SFT, LoRA, full-parameter FT, GRPO, curriculum learning; No external data
11	TJU	Tongji University	2/0	Qwen-VL, Xunzi-VL, HistLayout-DETR, ResNet-50	LoRA, coordinate normalization, curriculum learning; No external data
12	WHU	Wuhan University	2/0	Qwen-VL	LoRA, multi-task joint training; No external data
13	WHU-SAI	Wuhan University (School of Artificial Intelligence)	2/2	Qwen-VL	SFT, LoRA, DPO, GRPO, curriculum learning; External data used

Table 3: Overview of participating teams by modality ("Close/Open" denotes the number of evaluation result submissions by participating teams in the closed and open tracks, respectively.)

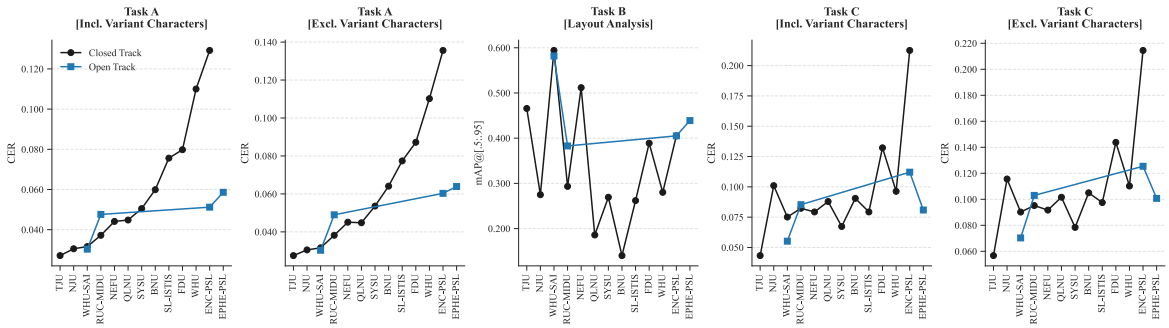


Figure 1: Performance by Task and Modality

6.2.1 Evaluation Results under Closed Modality

Table 4 and Table 5 summarize the top-five results under variant-included and variant-excluded rules.

ID	Team	Task A (inclusive of variant chars)			
		CER	NED	F1	Comp
1	TJU	2.71%	2.70%	97.54%	97.36%
2	NJU	3.05%	3.02%	97.19%	97.03%
3	WHU--SAI	3.16%	3.14%	97.02%	96.90%
4	RUC-MIDU	3.72%	3.69%	96.63%	96.39%
5	NEFU	4.41%	4.35%	95.89%	95.69%
ID	**Team**	**Task A (with variant chars excluded)**			
		CER	**NED**	**F1**	**Comp**
1	TJU	2.75%	2.74%	97.50%	97.32%
2	NJU	3.05%	3.02%	97.19%	97.03%
3	WHU--SAI	3.16%	3.14%	97.02%	96.90%
4	RUC-MIDU	3.82%	3.79%	96.53%	96.29%
5	QLNU	4.48%	4.40%	95.89%	95.65%

Table 4: Top5 results of Task A closed

TJU achieved state-of-the-art performance across both engraved and handwritten recognition tasks. For Task A, TJU attained a Comprehensive Score of 0.9736 (variant-included) with a CER of 0.0271, leveraging an innovative integration of Qwen2.5-VL with HistLayout-DETR (Carion et al., 2020). This layout-aware framework effectively handled complex typographical structures inherent in block-engraved text. For Task C, TJU secured first place with a score of 0.9571, outperforming the second-place team by 2.4 percentage points. NJU and WHU-SAI followed closely in Task A, utilizing base Qwen2.5-VL-7B and multi-stage optimization strategies such as Lora SFT (Hu et al., 2021), DPO (Rafailov et al., 2024), and GRPO (Shao et al., 2024), respectively.

ID	Team	Task C (with variant chars included)			
		CER	NED	F1	Comp
1	TJU	4.33%	4.33%	95.80%	95.71%
2	SYSU	6.73%	6.72%	93.39%	93.31%
3	WHU-SAI	7.51%	7.50%	92.67%	92.55%
4	NEFU	7.93%	7.91%	92.22%	92.12%
5	SL-ISTIS	7.93%	7.93%	92.16%	92.10%
ID	**Team**	**Task C (exclusive of variant chars)**			
		CER	**NED**	**F1**	**Comp**
1	TJU	5.67%	5.66%	94.46%	94.37%
2	SYSU	7.84%	7.83%	92.28%	92.20%
3	WHU-SAI	9.02%	9.01%	91.16%	91.04%
4	NEFU	9.17%	9.15%	90.98%	90.88%
5	RUC-MIDU	9.52%	9.51%	90.68%	90.54%

Table 5: Top5 results of Task C closed

For layout analysis (Task B), metrics included mAP@[.5:.95], Micro/Macro F1-scores, and Avg Match IoU. Table 6 presents the top-five results.

ID	Team	Task B			
		mAP@ [.5:.95]	Micro F1	Macro F1	Avg Match IoU
1	WHU-SAI	59.41%	83.42%	86.86%	76.38%
2	NEFU	51.18%	75.53%	78.79%	77.40%
3	TJU	46.57%	57.42%	71.34%	74.30%
4	ENC-PSL	40.51%	70.83%	67.75%	74.03%
5	FDU	38.89%	64.01%	63.18%	81.19%

Table 6: Top5 results of Task B closed

WHU-SAI achieved optimal performance with an mAP of 0.5941, significantly surpassing the

second-place team (NEFU, 0.5118). Their success stemmed from a sophisticated multi-stage training strategy combining SFT, DPO, and GRPO on Qwen2.5-VL-7B, augmented with LoRA for efficient parameter tuning. NEFU employed a dual-stream architecture combining YOLO ensembles with Qwen2.5-VL semantic filtering. Notably, FDU achieved the highest Avg Match IoU (0.8119), indicating superior bounding box localization accuracy despite lower classification performance.

6.2.2 Evaluation Results under Open Modality

The open modality imposes no additional restrictions on the pre-trained models, external data, and training strategies used by the participating systems, which can better test the upper performance limit of current mainstream multimodal large models and related technologies in ancient document visual processing tasks. A total of four teams submitted valid results for the open track of this competition, and the core performance of each system is as follows.

ID	Team	Task B			
		mAP@ [.5:.95]	Micro F1	Macro F1	Avg Match IoU
1	WHU-SAI	58.18%	82.78 %	86.45 %	75.72%
2	EPHE-PSL	43.89%	72.32 %	66.57 %	81.14%
3	ENC-PSL	40.51%	70.83 %	67.75 %	74.03%
4	RUC-MIDU	38.28%	66.73 %	59.83 %	80.41%

Table 8: Evaluation results of Task B open

The evaluation results of the character recognition tasks under the open modality are shown in Table 7.

WHU-SAI demonstrated consistent superiority across all open modality tasks. For Task A, their system attained a comprehensive score of 0.9703 (CER: 0.0303), approaching the closed modality optimum (0.9736). This marginal gap (0.33%) suggests their closed-track optimization strategy was already near the technical frontier. For Task C, WHU-SAI maintained leadership with scores of 0.9456 (variant-included) and 0.9305 (variant-excluded), utilizing a multi-stage paradigm (SFT, LoRA, DPO, GRPO) effective for handling calligraphic variability.

The evaluation results of the layout analysis task under the open modality are shown in Table 8. In layout analysis (Task B), WHU-SAI achieved an mAP of 0.5818, only 2.1% lower than their closed modality result. EPHE-PSL secured second place (mAP: 0.4389) by leveraging a multiscript pretraining corpus and 4.04 million annotated lines of historical documents. The marginal performance gains overall suggest that carefully curated training strategies tailored for the closed track may be more pivotal than simply scaling data and model parameters.

ID	Team	Task A (with variant chars included)			
		CER	NED	F1	Comprehensive
1	WHU-SAI	3.03 %	3.01%	97.15 %	97.03%
2	RUC-MIDU	4.76 %	4.67%	95.60 %	95.37%
3	ENC-PSL	5.12 %	5.08%	95.32 %	95.02%
4	EPHE-PSL	5.86 %	5.64%	94.79 %	94.38%
ID	**Team**	**Task A (with variant chars excluded)**			
		CER	**NED**	**F1**	**Comprehensive**
1	WHU-SAI	3.03 %	3.01%	97.15 %	97.03%
2	RUC-MIDU	4.90 %	4.81%	95.46 %	95.23%
3	EPHE-PSL	6.03 %	5.99%	94.41 %	94.11%
4	ENC-PSL	6.39 %	6.17%	94.26 %	93.85%
ID	**Team**	**Task C (with variant chars included)**			
		CER	**NED**	**F1**	**Comprehensive**
1	WHU-SAI	5.52 %	5.49%	94.71 %	94.56%
2	EPHE-PSL	8.09 %	8.09%	92.00 %	91.94%
3	RUC-MIDU	8.53 %	8.49%	91.68 %	91.54%
4	ENC-PSL	11.21 %	10.90%	89.96 %	89.20%
ID	**Team**	**Task C (with variant chars excluded)**			
		CER	**NED**	**F1**	**Comprehensive**
1	WHU-SAI	7.03 %	7.00%	93.20 %	93.05%
2	EPHE-PSL	10.07 %	10.07%	90.02 %	89.96%
3	RUC-MIDU	10.30 %	10.26%	89.91 %	89.77%
4	ENC-PSL	12.54 %	12.22%	88.63 %	87.88%

Table 7: Top5 results of Tasks A and C open

7. Conclusion

The evaluation task focused on character recognition and structural parsing within complex classical documents that are characterized by heterogeneous calligraphic styles and intricate page formats. Despite formidable challenges

including orthographic variants, layout interference, and interleaved text-image arrangements, the majority of participating teams demonstrated notable robustness, and achieved consistent and reliable performance across various subtasks. The results indicate that performance on well-structured engraved texts generally surpasses that achieved on handwritten manuscripts. Although the latter presents greater challenges, the majority of participating systems substantially exceeded the baseline models. This demonstrates the robust efficacy of contemporary methodologies in addressing the complexities of ancient document corpora. Methodologically, most teams developed task-specific models for distinct subtasks, resulting in significant performance gains in both character recognition and layout region detection. However, a unified architecture capable of concurrently processing diverse calligraphic styles and complex structures remains elusive, as does the synergistic modeling of textual content and layout semantics. Furthermore, comparative analysis reveals that zero-shot inference relying solely on general-purpose large VLMs underperforms in specialized ancient document contexts, underscoring the necessity of domain-specific instruction fine-tuning and layout-aware architectural adaptations.

Future research should focus on two primary directions: advancing lightweight specialized models to capture the nuances of ancient character morphology, and integrating LLMs into hybrid frameworks to balance stability and generalization. We expect future work to achieve high-precision, integrated recognition across heterogeneous writing styles and layouts. Beyond the scope of this task, the field is evolving toward more complex challenges, including structured text generation, layout-to-text alignment, and multimodal understanding of ancient Chinese documents.

8. Acknowledgments

This work was supported by the Major Project of the National Social Science Fund of China (Grant No. 21&ZD331) and General Program of the Ministry of Education of China Humanities and Social Sciences Fund (Grant No. 24A10319028).

9. Ethical Considerations

The datasets utilized in this evaluation are sourced exclusively from historical Chinese documents. Any views, ideologies, or cultural norms reflected in the textual content are inherent to the original ancient corpora and do not represent the perspectives, values, or endorsements of the authors or the organizing committee.

10. Bibliographical References

Li, B., Y. Yuan, J. Lu, M. Feng, C. Xu, W. Qu, and D. Wang. 2022. The First International Ancient Chinese Word Segmentation and POS Tagging Bakeoff: Overview of the EvaHan 2022 Evaluation Campaign. Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages, pages 135–140.

Wang, D., L. Lin, Z. Zhao, W. Ye, K. Meng, W. Sun, L. Zhao, X. Zhao, S. Shen, W. Zhang, and B. Li. 2023. EvaHan2023: Overview of the First International Ancient Chinese Translation Bakeoff. Proceedings of ALT2023: First Workshop on Ancient Language Translation, pages 1–14.

Li, B., B. Chang, Z. Xu, M. Feng, C. Xu, W. Qu, S. Shen, and D. Wang. 2024. Overview of EvaHan2024: The First International Evaluation on Ancient Chinese Sentence Segmentation and Punctuation. Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2024) @ LREC-COLING-2024, pages 229–236.

Li, B., B. Chang, R. Liu, X. Zhao, S. Shen, L. Liu, Y. Zhu, Z. Xu, W. Qu, and D. Wang. 2025. Overview of EvaHan2025: The First International Evaluation on Ancient Chinese Named Entity Recognition. Proceedings of the Second Ancient Language Processing Workshop associated with NAACL 2025, pages 97–105.

Bai, J. Z., S. Bai, S. S. Yang, S. J. Wang, S. N. Tan, P. Wang, J. Y. Lin, C. Zhou, and J. R. Zhou. 2023. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966.

Carion, N., F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. 2020. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision (ECCV), pages 213–229, Cham. Springer.

Hu, E. J., Y. L. Shen, P. Wallis, Z. Y. Allen-Zhu, Y. Z. Li, S. A. Wang, L. Wang, and W. Z. Chen. 2021. LoRA: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.

Kim, G., T. Hong, M. Yim, J. Nam, J. Park, J. Yim, W. Hwang, S. Yun, D. Han, and S. Park. 2021. OCR-free document understanding transformer. In Proceedings of the European Conference on Computer Vision (ECCV).

LeCun, Y., L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324.

Li, B., B. L. Chang, R. L. Liu, X. Zhao, S. Shen, L. H. Liu, Y. Zhu, Z. X. Xu, W. G. Qu, and D. B. Wang. 2025. Overview of EvaHan2025: The

first international evaluation on ancient Chinese named entity recognition. In Proceedings of the Second Workshop on Ancient Language Processing, pages 156–164, Laguna. Association for Computational Linguistics.

Li, B., B. L. Chang, Z. X. Xu, M. X. Feng, C. Xu, W. G. Qu, S. Shen, and D. B. Wang. 2024. Overview of EvaHan2024: The first international evaluation on ancient Chinese sentence segmentation and punctuation. In Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024, pages 229–236, Torino, Italia. ELRA and ICCL.

Li, B., Y. G. Yuan, J. Y. Lu, M. X. Feng, C. Xu, W. G. Qu, and D. B. Wang. 2022. The first international ancient Chinese word segmentation and POS tagging bakeoff: Overview of the EvaHan 2022 evaluation campaign. In Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages, pages 135–140, Marseille, France. European Language Resources Association.

Lin, T. Y., M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár. 2014. Microsoft COCO: Common objects in context. In Proceedings of the European Conference on Computer Vision (ECCV), pages 740–755, Cham. Springer.

Liu, H., C. Li, Q. Wu, and Y. J. Lee. 2024. Visual instruction tuning. In Advances in Neural Information Processing Systems (NeurIPS), 36.

Long, S. B., S. Y. Qin, D. Panteleev, A. Bissacco, Y. Fujii, and M. Raptis. 2022. Towards end-to-end unified scene text detection and layout analysis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 1039–1049.

Lu, H., W. Liu, B. Zhang, B. Wang, K. Dong, B. Liu, J. X. Sun, T. Z. Ren, Z. S. Li, H. Yang, Y. F. Sun, C. Q. Deng, H. W. Xu, Z. D. Xie, and C. Ruan. 2024. DeepSeek-VL: Towards real-world vision-language understanding. arXiv preprint arXiv:2403.05525.

Rafailov, R., A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. In Advances in Neural Information Processing Systems (NeurIPS), 36.

Reul, C., D. Christ, A. Hartelt, N. Balbach, M. Wehner, U. Springmann, C. Wick, C. Grundig, A. Büttner, and F. Puppe. 2019. OCR4all - an open-source tool providing a (semi-)automatic OCR workflow for historical printings. Applied Sciences, 9(22):4853.

Shao, Z. H., P. Y. Wang, Q. H. Zhu, R. X. Xu, J. X. Song, X. Bi, H. W. Zhang, M. C. Zhang, Y. K. Li, Y. Wu, and D. Y. Guo. 2024. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300.

Shen, Z., R. Zhang, M. Dell, B. C. G. Lee, J. Carlson, and W. Li. 2021. LayoutParser: A unified toolkit for deep learning based document image analysis. In Proceedings of the International Conference on Document Analysis and Recognition (ICDAR), pages 131–146. Springer International Publishing.

Shi, B., X. Bai, and C. Yao. 2015. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39:2298–2304.

Shi, Y., C. Liu, D. Peng, C. Jian, J. Huang, and L. Jin. 2023. M5HisDoc: A large-scale multi-style Chinese historical document analysis benchmark. In Advances in Neural Information Processing Systems (NeurIPS), 36.

Wang, D. B., L. T. Lin, Z. X. Zhao, W. H. Ye, K. Meng, W. L. Sun, L. Z. Zhao, X. Zhao, S. Shen, W. Zhang, and B. Li. 2023. EvaHan2023: Overview of the first international ancient Chinese translation bakeoff. In Proceedings of the Ancient Language Translation Workshop (ALT2023), pages 1–14, Macao SAR, China. Asia-Pacific Association for Machine Translation.

Wang, T., Y. Wu, R. Rubino, and C. Baumgartner. 2019. Radical aggregation network for few-shot offline handwritten Chinese character recognition. Pattern Recognition Letters, 125:821–827.

Wang, X., Y. Wang, B. Li, and D. Wang. 2024. Xunzi: A Chinese ancient text large language model. In Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 11984–11993.

Xu, Y., F. Yin, Z. X. Zhang, and C. L. Liu. 2018. Multi-task layout analysis for historical handwritten documents using fully convolutional networks. In Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI'18), pages 1057–1063. AAAI Press.

Yu, H. Y., Y. C. Wu, F. Shi, L. Liao, J. H. Lu, X. D. Ge, H. Wang, M. H. Zhuo, X. C. Wu, X. Fei, H. Feng, G. Z. Tang, A. L. Wang, H. S. Zhu, Y. F. He, Q. H. Liang, L. Y. Meng, C. Feng, C. Huang, J. Q. Tang, and B. Li. 2025. Benchmarking vision-language models on Chinese ancient documents: From OCR to knowledge reasoning. arXiv preprint arXiv:2509.09731.

Appendix A: Scores of Participating Teams

A.1 Scores of Participating Teams under Closed Modality

For comprehensive reference, the complete ranking results for all participating teams across all tasks and modalities are provided below. These tables include all valid submissions received during the evaluation period, extending beyond the top-five rankings presented in the main text.

ID	Team	Task A (with variant characters included)			
		CER	NED	F1	Comprehensive
1	TJU	2.71%	2.70%	97.54%	97.36%
2	NJU	3.05%	3.02%	97.19%	97.03%
3	WHU--SAI	3.16%	3.14%	97.02%	96.90%
4	RUC-MIDU	3.72%	3.69%	96.63%	96.39%
5	NEFU	4.41%	4.35%	95.89%	95.69%
6	QLNU	4.48%	4.40%	95.89%	95.65%
7	SYSU	5.05%	5.02%	95.36%	95.08%
8	BNU	5.99%	5.95%	94.53%	94.17%
9	SL-ISTIS	7.56%	7.44%	93.06%	92.65%
10	FDU	7.98%	7.95%	92.70%	92.23%
11	WHU	11.00%	10.89%	89.36%	89.13%
12	ENC-PSL	12.92%	12.89%	87.73%	87.28%
ID	Team	Task A (with variant characters excluded)			
		CER	NED	F1	Comprehensive
1	TJU	2.75%	2.74%	97.50%	97.32%
2	NJU	3.05%	3.02%	97.19%	97.03%
3	WHU--SAI	3.16%	3.14%	97.02%	96.90%
4	RUC-MIDU	3.82%	3.79%	96.53%	96.29%
5	QLNU	4.48%	4.40%	95.89%	95.65%
6	NEFU	4.51%	4.46%	95.79%	95.59%
7	SYSU	5.36%	5.32%	95.05%	94.77%
8	BNU	6.41%	6.37%	94.11%	93.75%
9	SL-ISTIS	7.74%	7.62%	92.88%	92.47%
10	FDU	8.72%	8.69%	91.95%	91.49%
11	WHU	11.02%	10.91%	89.34%	89.11%
12	ENC-PSL	13.56%	13.52%	87.09%	86.64%

Table A.1: Evaluation results of Task A under the closed track

ID	Team	Task C (with variant characters included)			
		CER	NED	F1	Comprehensive
1	TJU	4.33%	4.33%	95.80%	95.71%
2	SYSU	6.73%	6.72%	93.39%	93.31%
3	WHU-SAI	7.51%	7.50%	92.67%	92.55%
4	NEFU	7.93%	7.91%	92.22%	92.12%
5	SL-ISTIS	7.93%	7.93%	92.16%	92.10%
6	RUC-MIDU	8.23%	8.22%	91.97%	91.83%
7	BNU	9.05%	9.03%	91.33%	91.33%
8	QLNU	8.79%	8.73%	91.40%	91.28%
9	WHU	9.63%	9.61%	90.51%	90.42%
10	NJU	10.10%	10.02%	90.18%	90.00%
11	FDU	13.22%	13.20%	87.23%	86.92%
12	ENC-PSL	21.25%	21.25%	80.12%	79.16%
ID	Team	Task C (with variant characters excluded)			
		CER	NED	F1	Comprehensive
1	TJU	5.67%	5.66%	94.46%	94.37%
2	SYSU	7.84%	7.83%	92.28%	92.20%
3	WHU-SAI	9.02%	9.01%	91.16%	91.04%
4	NEFU	9.17%	9.15%	90.98%	90.88%
5	RUC-MIDU	9.52%	9.51%	90.68%	90.54%
6	SL-ISTIS	9.75%	9.75%	90.34%	90.28%
7	QLNU	10.16%	10.10%	90.03%	89.91%
8	BNU	10.50%	10.49%	89.86%	89.61%
9	WHU	11.02%	11.00%	89.12%	89.03%
10	NJU	11.56%	11.47%	88.73%	88.54%
11	FDU	14.38%	14.36%	86.09%	85.76%
12	ENC-PSL	21.45%	21.45%	79.92%	78.96%

Table A.2: Evaluation results of Task C under the closed track

ID	Team	Task B			
		mAP@[.5:.95]	Micro F1	Macro F1	Avg Match IoU
1	WHU-SAI	59.41%	83.42%	86.86%	76.38%
2	NEFU	51.18%	75.53%	78.79%	77.40%
3	TJU	46.57%	57.42%	71.34%	74.30%
4	ENC-PSL	40.51%	70.83%	67.75%	74.03%
5	FDU	38.89%	64.01%	63.18%	81.19%
6	RUC-MIDU	29.33%	24.14%	37.52%	71.40%
7	WHU	28.02%	29.12%	36.09%	71.04%
8	NJU	27.49%	26.72%	41.62%	67.40%
9	SYSU	26.93%	26.50%	36.76%	68.34%
10	SL-ISTIS	26.19%	19.46%	33.14%	72.25%
11	QLNU	18.58%	4.29%	20.29%	63.90%
12	BNU	14.02%	7.14%	11.73%	70.48%

Table A.3. Evaluation results of Task B under the closed track

A.2 Baselines

To provide a unified performance reference for the participating systems of this competition, current mainstream multimodal large models were selected to construct the baseline systems, namely the general multimodal large model Qwen2.5_VL_7B_Instruction, and Xunzi_Qwen2_VL_7B_Instruction optimized for the ancient document domain. Two processing paradigms were set for the baseline systems: the zero-shot inference paradigm without fine-tuning, and the supervised learning paradigm with instruction fine-tuning based on the training set of this competition, which fully covers the mainstream technical paths that participating teams may adopt. The performance results of each baseline system under all tasks and evaluation rules are shown in Table A.4.

The baseline test results show that the Qwen2.5_VL_7B_Instruction model with instruction fine-tuning on the training set achieved the optimal baseline performance in all tasks. Under the variant-included rule for Task A, the comprehensive score of the fine-tuned model reached 0.9397, a 3.60% improvement compared with the non-fine-tuned state, and the CER decreased from 0.1014 to 0.0618, with a significant performance improvement. Under the variant-included rule for Task C, the comprehensive score of the fine-tuned model increased from 0.8812 to 0.9086, also showing a significant performance gain.

Task A (with variant characters included)					
Model Name	**Fine-tuning Stage**	**CER**	**NED**	**F1**	**Comprehensive**
Qwen2.5_VL_7B_Instruction	No Fine-tuning	10.14%	9.47%	91.10%	90.37%
	Instruction Fine-tuning on Training Set	6.18%	6.13%	94.30%	93.97%
Xunzi_Qwen2_VL_7B_Instruction	No Fine-tuning	17.86%	17.40%	84.09%	82.82%
	Instruction Fine-tuning on Training Set	12.14%	11.83%	89.93%	88.54%
Task A (with variant characters excluded)					
Model Name	**Fine-tuning Stage**	**CER**	**NED**	**F1**	**Comprehensive**
Qwen2.5_VL_7B_Instruction	No Fine-tuning	11.21%	10.54%	90.07%	89.31%
	Instruction Fine-tuning on Training Set	6.85%	6.79%	93.64%	93.31%
Xunzi_Qwen2_VL_7B_Instruction	No Fine-tuning	18.51%	18.02%	83.45%	82.18%
	Instruction Fine-tuning on Training Set	12.64%	12.32%	89.45%	88.05%
Task B					
Model Name	**Fine-tuning Stage**	**mAP@[.5:.95]**	**Micro F1**	**Macro F1**	**Avg Match IoU**
Qwen2.5_VL_7B_Instruction	No Fine-tuning	0.00%	0.00%	0.00%	0.00%
	Instruction Fine-tuning on Training Set	20.06%	5.13%	15.30%	66.00%
Xunzi_Qwen2_VL_7B_Instruction	No Fine-tuning	2.36%	0.03%	1.14%	59.43%
	Instruction Fine-tuning on Training Set	19.17%	4.03%	11.30%	66.54%

Task C (with variant characters included)					
Model Name	**Fine-tuning Stage**	**CER**	**NED**	**F1**	**Comprehensive**
Qwen2.5_VL_7B_Instruction	No Fine-tuning	12.07%	11.93%	88.49%	88.12%
	Instruction Fine-tuning on Training Set	9.20%	9.19%	90.99%	90.86%
Xunzi_Qwen2_VL_7B_Instruction	No Fine-tuning	14.97%	14.92%	85.38%	85.14%
	Instruction Fine-tuning on Training Set	13.83%	13.76%	86.73%	86.35%
Task C (with variant characters excluded)					
Model Name	**Fine-tuning Stage**	**CER**	**NED**	**F1**	**Comprehensive**
Qwen2.5_VL_7B_Instruction	No Fine-tuning	13.38%	13.24%	87.18%	86.81%
	Instruction Fine-tuning on Training Set	10.66%	10.65%	89.53%	89.40%
Xunzi_Qwen2_VL_7B_Instruction	No Fine-tuning	16.09%	16.04%	84.25%	84.02%
	Instruction Fine-tuning on Training Set	15.20%	15.12%	85.34%	84.98%

Table A.4: Baseline evaluation results

Task Name	**Evaluation Rule**	**Modality**	**CER (%)**	**NED (%)**	**F1-score (%)**	**Comprehensive Score (%)**
Task A	With variant characters included	Closed Modality	2.71(-3.47)	2.70(-3.43)	97.54(+3.24)	97.36(+3.39)
		Open Modality	3.03(-3.15)	3.01(-3.12)	97.15(+2.85)	97.03(+3.06)
	With variant characters excluded	Closed Modality	2.75(-4.10)	2.74(-4.05)	97.50(+3.86)	97.32(+4.01)
		Open Modality	3.03(-3.82)	3.01(-3.78)	97.15(+3.51)	97.03(+3.72)
Task C	With variant characters included	Closed Modality	4.33(-4.87)	4.33(-4.86)	95.80(+4.81)	95.71(+4.85)
		Open Modality	5.52(-3.68)	5.49(-3.70)	94.71(+3.72)	94.56(+3.70)
	With variant characters excluded	Closed Modality	5.67(-4.99)	5.66(-4.99)	94.46(+4.93)	94.37(+4.97)
		Open Modality	7.03(-3.63)	7.00(-3.65)	93.20(+3.67)	93.05(+3.65)
Task Name	**Evaluation Rule**	**Modality**	**mAP@[.5:.95]**	**Micro F1**	**Macro F1**	**Avg Match IoU**
Task B	-	Closed Modality	59.41(+39.35)	83.42(+78.29)	86.86(+71.56)	76.38(+10.38)
		Open Modality	58.18(+38.12)	82.78(+77.65)	86.45(+71.15)	75.72(+9.72)

Table A.5: Performance improvement of the optimal participating results wrt the baseline model (%)

Although the Xunzi_Qwen2_VL_7B_Instruction model has been pre-trained and optimized for the ancient document domain, its performance is lower than that of the general multimodal large model Qwen2.5_VL_7B_Instruction in both zero-shot and fine-tuning scenarios of this task. The core reason is that the pre-training focus of this model is concentrated on the semantic understanding of ancient document texts, and it has obvious shortcomings in the visual feature learning of ancient document images and character recognition ability.

To quantify the performance improvement of the participating systems relative to the baseline model, the optimal participating results under each task and each modality were selected to compare with the performance of the core baseline model. Based on the average value of the baseline data, the maximum improvement range and specific values of each core indicator were calculated, and the results are shown in T able A.5.

The performance improvement results show that the optimal performance of all participating systems in this competition is significantly better than that of the official baseline model. For the character recognition tasks, the overall performance improvement under the closed modality is higher than that under the open modality, with the improvement range of comprehensive score from 3.06% to 4.97%, and the maximum reduction range of CER from 3.15% to 4.99%. For the layout analysis task, the performance improvement of the participating systems is more significant, with the improvement range of mAP@[.5:.95] from 38.12% to 39.35%, and the improvement range of Macro F1-score from 71.15% to 71.56%. This reflects that the model optimization and adaptation strategies adopted by the participating teams, targeting the domain characteristics of ancient document visual processing, are extremely effective and can significantly break through the performance bottleneck of general multimodal large models in ancient document layout recognition tasks.

A.3 Analysis of Variant Character Recognition and Character Distribution Characteristics

Rank	Char	Frequency	%
1	無	2,694	1.66%
2	之	1,829	1.13%
3	不	1,786	1.10%
4	一	1,513	0.93%
5	為	1,473	0.91%
6	以	1,411	0.87%
7	空	1,305	0.80%
8	有	1,188	0.73%
9	二	1,179	0.73%
10	所	1,154	0.71%

Table A.6: Statistics of top 10 high-frequency characters across datasets

Table A.6 presents the frequency statistics of the top 10 high-frequency characters across all datasets in this evaluation. Collectively, these 10 characters amount to 15,532 tokens, accounting for 9.57% of the total 162,332 character tokens in the full corpus. This distribution explicitly reveals the pronounced long-tail feature and severe class imbalance inherent in ancient Chinese texts. While models achieve stable and high recognition accuracy on these prevalent characters with extensive pre-training coverage, low-frequency rare characters and non-standard variant forms, which are particularly concentrated in handwritten datasets, remain the core bottleneck restricting the overall performance of ancient Chinese OCR systems.

Appendix B: Character Distribution Statistics

The following tables present the full statistical distribution of Chinese characters across Unicode blocks for all dataset splits, referenced from Section 6.2.

Unicode Block	Train_A	Test_A	Train_C	Test_C
CJK Basic	81,626 (97.80%)	4,540 (99.65%)	67,302 (97.85%)	4,716 (99.75%)
Ext-B	1,664 (1.99%)	13 (0.29%)	695 (1.01%)	0 (0.00%)
Ext-A	165 (0.20%)	1 (0.02%)	469 (0.68%)	12 (0.25%)
Ext-F	1 (0.00%)	0 (0.00%)	213 (0.31%)	0 (0.00%)
Ext-E	2 (0.00%)	0 (0.00%)	36 (0.05%)	0 (0.00%)
Ext-D	0 (0.00%)	0 (0.00%)	35 (0.05%)	0 (0.00%)
Ext-C	2 (0.00%)	0 (0.00%)	33 (0.05%)	0 (0.00%)
Compati-bility Supple-ment	0 (0.00%)	2 (0.04%)	0 (0.00%)	0 (0.00%)
Sum	83,460 (100.00%)	4,556 (100.00%)	68,783 (100.00%)	4,728 (100.00%)

Table B.1: Statistical Distribution by Total Character Frequency

Unicode Block	Train_A	Test_A	Train_C	Test_C
CJK Basic	3,162 (98.20%)	1,406 (99.58%)	4,523 (96.01%)	680 (99.85%)
Ext-B	33 (1.02%)	3 (0.21%)	95 (2.02%)	0 (0.00%)
Ext-A	21 (0.65%)	1 (0.07%)	51 (1.08%)	1 (0.15%)
Ext-F	1 (0.03%)	0 (0.00%)	25 (0.53%)	0 (0.00%)
Ext-E	2 (0.06%)	0 (0.00%)	7 (0.15%)	0 (0.00%)
Ext-D	0 (0.00%)	0 (0.00%)	6 (0.13%)	0 (0.00%)
Ext-C	1 (0.03%)	0 (0.00%)	4 (0.08%)	0 (0.00%)
Compa-tibility Supple-ment	0 (0.00%)	2 (0.14%)	0 (0.00%)	0 (0.00%)
Total Sum	3,220 (100.00%)	1,412 (100.00%)	4,711 (100.00%)	681 (100.00%)

Table B.2: Statistical Distribution by Unique Character Vocabulary

Appendix C: Auxiliary Evaluation Metrics

This appendix details the auxiliary metrics used to provide a more granular analysis of the models' performance across Tasks A, B, and C, as well as the comprehensive score used for final rankings.

C.1 Auxiliary Metrics for Text Recognition (Tasks A and C)

(1)Macro and Micro CER:

To prevent a single metric from introducing evaluation bias, we further split CER into Macro CER and Micro CER:

$$\text{Macro CER}=\frac{1}{N}\sum_{i=1}^{N}\text{CER}_i \qquad (C\text{-}1)$$

$$\text{Micro CER}=\frac{\sum_{i=1}^{N}\text{Edit_Distance}(\text{ref}_i,\text{hyp}_i)}{\sum_{i=1}^{N}|\text{ref}_i|} \qquad (C\text{-}2)$$

(2) Precision, Recall, and F1-score

Based on the number of correctly matched characters, we calculate three auxiliary metrics that intuitively reflect the model's comprehensive ability in positive sample recognition:

$$\text{Precision}=\frac{N_{\text{correct}}}{|\text{hyp}|} \qquad (C\text{-}3)$$

$$\text{Recall}=\frac{N_{\text{correct}}}{|\text{ref}|} \qquad (C\text{-}4)$$

$$\text{F1}=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \qquad (C\text{-}5)$$

where $N_{correct}$ is the number of correctly matched characters between the predicted text and the reference text.

(3) Normalized Edit Distance (NED)

Considering that conventional edit distance can easily fail when there is a drastic length deviation between the predicted text and the ground truth label, we supplemented the Normalized Edit Distance (NED). NED ensures evaluation robustness under extreme length differences by using the maximum length of the two as the denominator:

$$\text{NED}=\frac{\text{Edit_Distance}(\text{ref},\text{hyp})}{\max(|\text{ref}|,|\text{hyp}|)} \qquad (C\text{-}6)$$

(4) Comprehensive Score

To provide an intuitive basis for the final ranking, we designed a weighted comprehensive score integrating the above dimensions. This score blends the character error rate, exact match rate, and normalized edit distance in different proportions, and its value is strictly mapped to the range [0,1]. The calculation formula is:

$$\text{Score}_{\text{comprehensive}}=$$

$$(1\text{-CER})\times 0.5+\text{F1}\times 0.3+(1\text{-NED})\times 0.2 \qquad (C\text{-}7)$$

A higher score indicates better comprehensive model performance. This metric is used to determine the final ranking of participating teams on datasets A and C.

C.2 Auxiliary Metrics for Object Detection (Task B)

(1)F1 Score

Both Macro-F1 and Micro-F1 calculation methods are provided. For the label set $\mathbb{L}$, they are defined as:

$$\text{Micro-F1}=\frac{2\times\sum_{l\in\mathbb{L}}\text{TP}_l}{2\times\sum_{l\in\mathbb{L}}\text{TP}_l+\sum_{l\in\mathbb{L}}\text{FP}_l+\sum_{l\in\mathbb{L}}\text{FN}_l} \qquad (C\text{-}8)$$

$$\text{Macro-F1}=\frac{1}{|\mathbb{L}|}\sum_{l\in\mathbb{L}}\frac{2\times\text{TP}_l}{2\times\text{TP}_l+\text{FP}_l+\text{FN}_l} \qquad (C\text{-}9)$$

where TP_l, FP_l, and FN_l are the number of true positives, false positives, and false negatives for label l, respectively.

(2) Mean IoU of Matched Boxes

When determining whether a predicted box and a ground truth annotation box are successfully matched, we use Intersection over Union (IoU) as the core basis. The algorithm employs a greedy strategy to group labels within the same image, preferentially matches the box pair with the maximum IoU value, and considers IoU 0.5 as a valid detection. Its calculation formula is:

$$\overline{\text{IoU}}=\frac{1}{N_{\text{match}}}\sum_{(B_p,B_g)\in\mathbb{M}}\text{IoU}(B_p,B_g) \qquad (C\text{-}10)$$

where $\mathbb{M}$ is the set of successfully matched box pairs, and $N_{match}=|\mathbb{M}|$.

Appendix D: Error Analysis and Discussion

D.1 Failures in Complex Layouts

The typesetting and layout of ancient Chinese documents differ significantly from modern printing. Among these differences, complex elements such as interlinear small-character annotations (Shuanghang Xiaozhi), raised items (Taixiang), seals, and book margins (Shukou) pose severe challenges to the model's layout analysis (Task B). Evaluation results indicate that even the best-performing system in the closed track achieved an average precision (mAP@[.5:.95]) of only 0.5941 in layout analysis, which is far below the performance level of character recognition tasks.

Among the numerous factors causing layout analysis failures, the confusion between double-line annotations and the main text is the most prominent. From the perspective of visual

features, interlinear small-character annotations in ancient books are tightly embedded between single columns of large-character main text, forming a unique "interlocking text-and-image nesting" phenomenon. Most existing object detection models or layout analysis paradigms are typically pre-trained on modern documents (e.g., left-to-right, horizontal layout, distinct line spacing). When processing ancient book images, models are highly prone to structural misjudgments: because the physical distance between the large main text characters and the small annotation characters is too close, the model easily groups the entire column (containing both main text and annotations) into a single Text Block, resulting in the failure of fine-grained layout element segmentation. Furthermore, although some models perform well on the Avg Match IoU metric for layout region localization, they fail to accurately separate “text” (main text) from "annotations" according to typographic logic, leading to a drastic decline in comprehensive classification metrics (such as Macro F1). Future structural parsing should integrate mature document layout frameworks, such as LayoutParser (Shen et al., 2021), to mitigate typographic logic errors.

D.2 Variant and Rare Characters

One of the core difficulties in optical character recognition (OCR) lies in the accurate recognition of long-tail rare characters and variant characters (Wang et al., 2019). Although variant characters exert a measurable influence on system performance, their relatively low occurrence frequency means that the overall impact on evaluation scores remains limited. To quantitatively examine this effect, this study conducted a statistical mapping of Chinese character Unicode blocks for Dataset A (block-engraved editions) and Dataset C (manuscripts) from two dimensions: Total Character Frequency and Unique Character Vocabulary (deduplicated non-repeating characters). The detailed statistical results are provided in Appendix B (Tables B1 and B2).

The statistical results reveal the extreme long-tailed distribution characteristics of ancient Chinese texts. In the dimension of character frequency (Table B1), the CJK Basic block (U+4E00–U+9FFF) maintains absolute dominance across all subsets (97.80% to 99.75%). This explains why even baseline models without domain-specific fine-tuning can often achieve a basic recognition rate of around 90%.

Regarding the recognition performance of characters in Unicode extension blocks, our analysis reveals a pronounced performance gap compared to CJK Basic characters. Characters from the CJK Basic block (U+4E00-U+9FFF) are recognized with high accuracy owing to their high frequency and broad coverage in model pre-training corpora. In contrast, characters belonging to extension blocks (Ext-A through Ext-F) exhibit substantially higher error rates. For instance, in the manuscript dataset (Dataset C), extension-block characters account for approximately 4% of the unique character vocabulary; yet their CER contribution is disproportionately large relative to their frequency. This disparity is particularly pronounced for Ext-B characters (U+20000-U+2A6DF), which represent obscure historical glyphs rarely encountered in modern digitized corpora. These findings confirm that improving the coverage of extension-block characters in model training data and vocabulary design is a critical direction for advancing ancient Chinese OCR systems.

However, in the dimension of unique character vocabulary (Table B2), different text carriers present significant variances and recognition challenges. Compared to the well-standardized block-engraved editions (Train_A), the unique character vocabulary (vocabulary diversity) of the manuscript texts (Train_C) expands sharply to 4,711. More importantly, in Train_C, variant and rare characters located in the extension blocks outside the CJK Basic set (Ext-A to Ext-F) contribute to nearly 4% of the character vocabulary diversity, with a distribution breadth far exceeding that of Train_A. For instance, Ext-B and Ext-F account for 2.02% and 0.53% of the unique characters in the manuscripts, respectively.

This data structural contrast—“extremely rich in unique characters but extremely low in occurrence frequency”—constitutes the core bottleneck for current large multimodal models(LMMs) in ancient text OCR tasks. Because the visual encoders of open-source LMMs (such as Qwen2.5-VL) rarely cover low-frequency rare characters in the Ext-B to Ext-F blocks during the pre-training phase, models inevitably encounter severe representation dimensionality reduction or Out-of-Vocabulary (OOV) truncation when processing such long-tail characters. This causes the Character Error Rate (CER) for the long-tail character segment to remain persistently high.

Beyond the Unicode distribution analysis, we further conducted a fine-grained evaluation of how each participating team’s system actually handled the specific variant character instances present in the test sets. Using the predefined variant character mapping table (19 mapping pairs in total), we identified all positions in the ground-truth annotations where a variant character occurred and examined whether each team's prediction was correct at those positions. Two accuracy metrics are reported: Strict Accuracy, which counts a prediction as correct only when the predicted character is identical to the ground-truth variant character; and Loose Accuracy, which additionally accepts the

equivalent standard-form counterpart as correct. Table D.1 summarizes the results across all thirteen participating teams.

Team	Task A Strict Acc.	Task A Loose Acc.	Task C Strict Acc.	Task C Loose Acc.
BNU	85.3%	94.1%	56.8%	79.3%
ENC-PSL	47.1%	77.9%	20.6%	47.1%
EPHE-PSL	94.1%	95.6%	49.7%	72.9%
FDU	83.8%	88.2%	61.9%	84.5%
NEFU	76.5%	89.7%	36.8%	78.7%
NJU	60.3%	92.1%	48.4%	48.4%
QLNU	95.6%	95.6%	46.5%	83.2%
RUC-MIDU	95.6%	95.6%	40.6%	71.6%
SL-ISTIS	94.1%	95.6%	63.2%	92.9%
SYSU	70.6%	80.9%	30.3%	65.8%
TJU	91.2%	94.1%	40.6%	68.4%
WHU	95.2%	95.2%	49.7%	79.3%
WHU-SAI	94.1%	94.1%	47.7%	77.4%

Table D.1: Variant character recognition accuracy

Several patterns emerge from Table 12. First, Task A variant accuracy is substantially higher than Task C across all teams, with most teams achieving loose accuracy above 88% on Task A, while Task C loose accuracy ranges widely from 47.1% (ENC-PSL) to 92.9% (SL-ISTIS). This performance gap is consistent with the broader finding that handwritten manuscripts pose greater recognition difficulties than block-engraved texts, and specifically reflects that the variant character forms encountered in handwriting are more visually irregular and diverse.

Second, the gap between strict and loose accuracy reveals how systems handle graphically equivalent forms. Teams such as NEFU (Task C: 36.8% strict vs. 78.7% loose, a gap of 41.9 percentage points) and QLNU (Task C: 46.5% vs. 83.2%, a gap of 36.7 points) show that their models frequently predict the standard-form counterpart of a variant character rather than the variant itself. This reflects the strong prior toward modern standardized character forms instilled during pre-training, which causes models to unconsciously "normalize" the visual input even when the ground truth requires the historical variant. The character 稅/税 is a representative example: the ground-truth character in Dataset C is 稅 (the variant form), yet the majority of teams predominantly predicted 税, the modern standard form, resulting in high loose accuracy but low strict accuracy.

Third, the character 鬨 proved to be the most challenging variant across all teams in Task C. It was frequently confused with visually similar but unrelated characters such as 鬩, 鬧, and 鬩, leading to low loose accuracy for many systems (e.g., ENC-PSL: 1/47, NJU: 11/47). Unlike the 稅/税 case - where the error pattern is systematic and semantically grounded - errors on 鬨 appear more perceptually driven, suggesting that models fail to reliably distinguish this character's distinctive graphemic features from surrounding similar glyphs.

D.3 Systematic Biases

The adoption of LMMs as the dominant technical backbone in this evaluation—employed by virtually all participating teams across both modalitie—brings into focus two systematic failure modes that are distinct in mechanism yet share a common root: the conflict between the model's intrinsic language prior, shaped by pre-training on modern standardized text, and the visual fidelity demanded by archival OCR. These two modes are hallucination and orthographic over-normalization, and the results of this evaluation provide concrete evidence for both.

Hallucination: perceptual failure driven by language prior. Hallucination in the OCR context refers to cases where the model generates a character that bears no faithful correspondence to the visual content of the image, where the language decoder, failing to obtain a reliable signal from the visual encoder, defaults to a contextually plausible character rather than the one actually present. The clearest evidence in our results comes from the character 鬨 in Dataset C. Across all thirteen participating teams, 鬨 was the most poorly recognized variant character, with loose accuracy as low as 1/47 (ENC-PSL) and 11/47 (NJU). Crucially, the errors were not random: the character was systematically confused with visually similar but semantically unrelated characters - 鬩, 鬧, 鬩 - all of which belong to the high-frequency CJK Basic block and appear extensively in classical Chinese pre-training corpora. The pattern is precisely what hallucination predicts: when the visual encoder produces an uncertain representation of an unfamiliar or low-frequency glyph, the language model component intervenes and "completes" the output toward a character it has seen many times in similar syntactic contexts, overriding the visual evidence. This mechanism is further amplified for characters in Unicode extension blocks (Ext-B through Ext-F), which are rarely encountered in modern digitized corpora and for which the visual decoder has no stable internal representation; in such cases, the language prior becomes the only signal effectively, producing outputs that may be linguistically coherent but graphically unfaithful.

Orthographic over-normalization: generation-stage bias toward standard forms. The second failure mode is subtler and, from the perspective of documentary fidelity, equally damaging. Over-

normalization does not involve a failure of visual perception: the model correctly identifies the semantic identity of the character in the image, but in the generation stage substitutes the modern standard form for the historical variant actually depicted. The gap between strict and loose accuracy observed across teams in Dataset C provides systematic evidence for this phenomenon. Teams such as NEFU (strict: 36.8%, loose: 78.7%, gap: 41.9 pp) and QLNU (strict: 46.5%, loose: 83.2%, gap: 36.7 pp) show that a large proportion of their "errors" under strict evaluation are in fact correct recognitions of the character's semantic identity, delivered in the wrong orthographic form. The character pair 税/稅 is paradigmatic: the ground truth in Dataset C is 稅 (the historical variant), yet the overwhelming majority of teams predicted 税 (the modern standard form), because this form dominates the pre-training corpus and is thus assigned a higher generation probability regardless of what is visually present. Unlike hallucination, over-normalization preserves semantic meaning while erasing graphemic authenticity—a trade-off that is acceptable in semantic NLP tasks but constitutes a fundamental failure in archival transcription, where the exact character form carries philological and paleographic significance.

Shared root, different remedies. Both failure modes originate from the same structural asymmetry: pre-training corpora for large models consist overwhelmingly of collated, standardized modern text, leaving rare glyphs, extension-block characters, and historical variant forms severely underrepresented. However, they require different interventions. Hallucination, arising from visual encoder uncertainty, calls for stronger visual grounding—improving the coverage of rare and archaic glyphs in pre-training, and introducing visual confidence mechanisms that prevent the language decoder from overriding low-confidence visual representations. Over-normalization, arising from generation-stage bias, calls for orthographic calibration—through fine-tuning on corpora that explicitly include historical variant forms, alignment training objectives (such as DPO) that penalize substitution of variants with their standard counterparts, or constrained decoding strategies that restrict the output vocabulary to forms attested in the source image's character set. The results of this evaluation suggest that fine-tuning on in-domain data partially addresses both problems—top closed-track systems exhibit neither failure mode at a significant rat—but the underlying biases are not eliminated, and will resurface whenever these models are applied to document collections outside their fine-tuning distribution.

D.4 A New Framework for Ancient Chinese OCR & Layout Analysis

The task of recognizing text from ancient Chinese document images can be approached through two fundamentally distinct architectural paradigms. The first is an end-to-end paradigm, in which a single multimodal large model processes the raw page image and directly outputs the transcribed text, performing implicit layout understanding and character recognition in a unified forward pass. The second is a pipeline paradigm, in which layout analysis and OCR are decoupled into sequential stages: a layout analysis model first identifies and segments text regions (columns, annotations, main body), after which a dedicated OCR component transcribes each region independently.

Both paradigms were represented among the participating systems in this evaluation, and their respective strengths and limitations are revealed by the results.

The end-to-end approach, exemplified by systems built directly on Qwen2.5-VL, offers substantial practical advantages: a simpler engineering stack, no dependency on intermediate layout annotations, and the ability to leverage the model's broad world knowledge for contextual disambiguation. The best-performing closed-modality system overall — TJU, which achieved a Task A comprehensive score of 0.9736 and a Task C score of 0.9571 — demonstrates that a well-optimized end-to-end architecture can reach state-of-the-art performance. However, the end-to-end approach is also more susceptible to the over-generation failure mode described in Section 6.3: without explicit layout constraints, the model must infer reading order, column boundaries, and the distinction between main text and interlinear annotations from visual context alone, which can lead to systematic errors in structurally complex pages.

The pipeline approach, exemplified by TJU's integration of HistLayout-DETR for layout pre-segmentation prior to OCR, addresses this weakness by making layout structure explicit. When text regions are correctly isolated, the downstream OCR component operates on substantially simpler inputs - individual columns or lines rather than full page images-reducing ambiguity and limiting the effective receptive field over which over-generation can occur. The results of WHU-SAI in Task B (mAP of 0.5941, the highest in the evaluation) further suggest that dedicated layout models, trained specifically on the structural peculiarities of ancient Chinese books, can achieve precision that end-to-end models currently cannot match. This structural advantage is particularly valuable for documents containing double-line annotations (Shuanghang Xiaozhi), raised items (Taixiang), and other

complex typographic elements whose correct segmentation is a prerequisite for accurate transcription.

The trade-off, however, is that the pipeline approach introduces error propagation: layout segmentation error—misclassified regions, merged columns, or missed annotation blocks—directly degrade the quality of downstream OCR in ways that are difficult to recover from. Moreover, pipeline systems require labeled layout data for training the segmentation stage, which is costly to produce for historical documents. The baseline results (Table 8) illustrate this brittleness: the zero-shot Qwen2.5-VL baseline achieves an mAP of 0.00 on Task B without fine-tuning, indicating that general-purpose models have virtually no out-of-the-box layout understanding for ancient Chinese books.

Looking forward, we propose that the most promising direction is not a choice between these two paradigms, but a staged hybrid architecture that combines their respective strengths. Concretely, such a system would consist of three components: (1) a lightweight layout analysis module trained on domain-specific annotations to segment the page into typed regions; (2) a reading-order resolution module that reconstructs the logical reading sequence—right-to-left, top-to-bottom, with correct interleaving of main text and annotation—from the detected regions; (3) a fine-tuned multimodal OCR model that transcribes each region independently, operating on clean, pre-segmented inputs rather than full page images. This architecture preserves the recognition power of LMMs while anchoring their outputs to explicit structural constraints, thereby suppressing over-generation and improving robustness on the complex typographic elements that currently constitute the primary bottleneck for ancient Chinese document OCR systems.

A Multi-Stage System for Ancient Chinese OCR and Layout Understanding in the EvaHan2026 Shared Task

KeYan Liang, Meiling Liu
Aulin College, Northeast Forestry University
Harbin, Heilongjiang, China
eyonryan00@nefu.edu.cn, mlliu@nefu.edu.cn

Abstract

This paper presents a multi-stage system for the EvaHan2026 shared task, addressing the complex challenges of ancient Chinese optical character recognition (OCR) and layout understanding. For text recognition (Tasks A and C), we adopt parameter-efficient LoRA fine-tuning on the Qwen2.5-VL-7B-Instruct vision-language model (VLM). By directly processing full-resolution long-column images, we preserve critical spatial and contextual integrity without heuristic region cropping. For document layout analysis (Task B), we propose a novel hybrid perception-reasoning paradigm. Instead of relying solely on scaling visual detectors, we decouple localization and understanding: utilizing a YOLO-based ensemble for precise spatial bounding, and casting the VLM as a semantic verifier to eliminate spurious detections. Evaluated on the official unseen test set, our system achieves substantial improvements over the provided baselines, obtaining a 0.0441 Character Error Rate (CER) for printed OCR, a 0.0793 CER for handwritten OCR (including variants), and a 0.5118 mAP@[0.5:0.95] for layout detection. These results demonstrate that integrating VLM-based semantic reasoning into traditional visual detection pipelines is highly effective for multimodal historical document analysis.

Keywords: Ancient Chinese OCR, Layout Analysis, Vision-Language Model, Hybrid Perception-Reasoning, EvaHan2026

1. Introduction

Digitizing ancient Chinese documents poses substantial computational challenges due to complex layouts, severe degradation, cursive stroke adhesion, and variant glyphs. Furthermore, diverse visual elements (text, seals, illustrations) frequently occlude one another, complicating document understanding. The EvaHan2026 shared task provides a benchmark for this. Tasks A and C focus on printed and handwritten OCR within pre-cropped vertical columns (often reaching extreme aspect ratios of approximately 200×4000 pixels), requiring models to preserve long-context spatial information. Conversely, Task B requires full-page layout detection across four distinct categories.

In this paper, we propose a robust multi-stage system tackling these subtasks. Beyond pure engineering optimizations, our core methodological insight is decoupling spatial localization from semantic understanding. Our main contributions are:

- **Direct Long-Column OCR via VLM Adaptation:** For Tasks A and C, we adapt a state-of-the-art vision-language model (VLM), Qwen2.5-VL (Bai et al., 2025), to the ancient document domain via parameter-efficient LoRA fine-tuning. By treating single-column OCR as a direct image-to-sequence generation task on full-resolution images, we entirely bypass error-prone character-level cropping pipelines.
- **Hybrid Perception-Reasoning Layout Paradigm:** For Task B, we introduce a category-specific hybrid framework casting the VLM as a semantic critic. While a YOLO-based dual-stream ensemble (Ultralytics, 2023, 2024) handles raw spatial localization, the VLM acts as a semantic verifier to filter out hard false positives. Combined with a two-stage seal classification pipeline and Hough-based geometric reasoning, this paradigm effectively bridges raw object detection and high-level cognitive verification.

2. Related Work

Traditional cascaded OCR pipelines often suffer from error propagation. Recently, Large Vision-Language Models (VLMs) (Bai et al., 2023; Wang et al., 2024; Bai et al., 2025) have shown remarkable multimodal capabilities. We build upon this by applying LoRA-adapted VLMs (Hu et al., 2022) directly to uncropped, full-resolution ancient text columns. For Document Layout Analysis (DLA), while unified toolkits (Shen et al., 2021) and multimodal frameworks (Xu et al., 2020) advance modern document understanding, historical scans still heavily rely on object detectors (Ultralytics, 2023, 2024) and text detectors (Liao et al., 2020) for bounding box precision. We overcome the limitations of pure visual detection and end-to-end models (Li et al., 2021; Blecher et al., 2023) in dense scans by introducing a VLM-in-the-loop verification step, bridging raw detection and semantic reasoning. Finally, we mitigate VLM memory bottlenecks using vLLM (Kwon et al., 2023) and address se-

Proceedings of LT4HALA 2026 @ LREC 2026, pages 263–267.
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

vere class imbalances via targeted synthetic data augmentation (Shorten and Khoshgoftaar, 2019).

3. Methodology

3.1. Full-Resolution OCR Adaptation (Tasks A and C)

For both printed text OCR (Task A) and handwritten calligraphy OCR (Task C), the provided document images are pre-cropped into single vertical text columns. While this isolates the text sequence, the extreme aspect ratios (e.g., approximately 200×4000 pixels) challenge traditional recognition pipelines. We directly fine-tune Qwen2.5-VL-7B-Instruct (Bai et al., 2025) using parameter-efficient Low-Rank Adaptation (LoRA). The base model parameters remain completely frozen, and only the low-rank adaptation matrices are updated. We configure the LoRA hyperparameters with a rank $r = 64$, a scaling factor $\alpha = 128$, and a dropout rate of 0.05. This strategy mitigates catastrophic forgetting, minimizes trainable parameters, and demonstrates robust domain adaptation.

3.2. Multi-Stage Layout Detection (Task B)

Task B focuses on detecting four distinct layout categories. Unlike conventional parallel detection pipelines, our system implements a strictly sequential, three-phase execution flow (Figure 1). This design explicitly models the strong spatial dependencies between different layout elements.

3.2.1. Phase 1: Primary Text Detection

The *Text Pipeline* executes first, providing essential spatial priors for downstream modules. Initially, the RGB image and its adaptively binarized version are fed into a dual-stream YOLO ensemble (YOLOv8 and YOLO11). To mitigate high false-positive rates in historical text detection, Qwen2.5-VL acts as a semantic verifier, evaluating each candidate box to confirm visible Chinese characters and reject artifacts. Finally, Non-Maximum Suppression (NMS) and morphological cleaning merge erroneously split components.

3.2.2. Phase 2: Parallel Multimodal Pipelines

Conditioned on Phase 1 outputs, the remaining pipelines execute in parallel:

Image Pipeline. Following YOLO ensemble detection, we enforce two text-dependent geometric heuristics: (1) *Overlap Rejection*: Discarding image boxes with $> 80\%$ IoU with text regions; (2) *Density Rejection*: Removing candidates encompassing ≥ 4 text boxes that cover $> 45\%$ of the image area. Notably, our VLM verifier is exclusively applied to the *text* category to optimally balance reasoning capabilities with system latency, as other categories are better suited for geometric or lightweight classification approaches.

Seal Pipeline. We employ a dedicated two-stage cascade to balance recall and precision for degraded seals. Following a high-recall YOLO detection and NMS, each candidate crop is adaptively binarized, padded to a standardized 256×256 square, and fed into a specialized YOLO classification model (confidence threshold > 0.80).

Book Edge Pipeline. After initial YOLO detection, we convert the region to grayscale, remove black borders, and apply the Hough Transform to extract candidate vertical spine lines, followed by a lateral whiteness density verification.

3.2.3. Phase 3: Global Conflict Resolution

In the final phase, a pipeline merger standardizes the heterogeneous outputs and resolves spatial collisions across all categories using a deterministic cross-class rule set:

1. **Seal overrides Image:** When a seal and an image overlap significantly (IoU > 0.65), the image bounding box is considered a false positive and deleted.
2. **Text Excludes Seal:** To prevent dense text blocks from being misclassified as false seals, any seal box whose 90% area encompasses two or more text boxes is deleted.
3. **Book Edge Physical Cropping:** Any layout object overlapping with a book edge by more than 70% is entirely removed. For partial occlusions ($\leq 70\%$), the object's bounding box is geometrically cropped along the spine line.

4. Experiments

4.1. Dataset and Augmentation

The EvaHan2026 dataset consists of 5,000 samples for each subtask, partitioned into 4,500 for training and 500 for validation. Tasks A and C represent samples as single vertical columns (200×4000 pixels), while Task B provides binarized full-page scans (500×900 pixels).

To mitigate the severe class imbalance in the *seal* category (only $\sim$600 labels), we implemented a targeted synthetic data generation pipeline. By applying a $5\times$ augmentation factor, we created 3,000 layout images. We simulated historical degradation

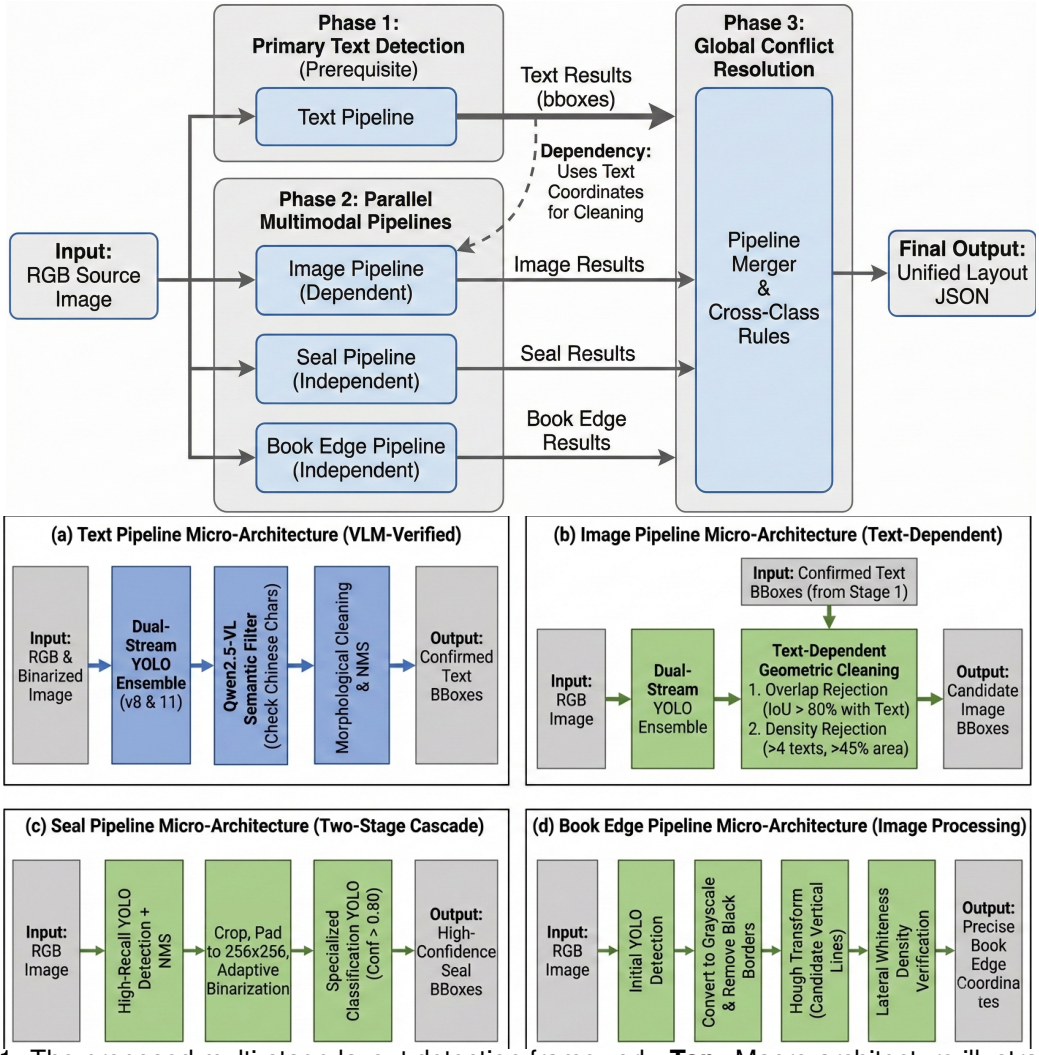


Figure 1: The proposed multi-stage layout detection framework. **Top:** Macro-architecture illustrating the sequential execution flow. **Bottom:** Micro-architectures of the category-specific pipelines.

through adaptive Gaussian thresholding and morphological erosion to mimic ink fading, blending the synthetic seals into background documents. While this introduces domain gaps, our high-recall initial detection coupled with the secondary classification network effectively filters out spurious detections.

4.2. Implementation Details

Experiments were conducted on a single NVIDIA RTX 5090 GPU (32GB VRAM). For Tasks A and C, we fine-tuned Qwen2.5-VL using LLaMA-Factory (Zheng et al., 2024) with FlashAttention-2 (Dao, 2023). The model was trained in bf16 precision (batch size 1, 8 gradient accumulation steps, 5×10^{-5} peak learning rate). Task C required 2,400 training steps, whereas Task A converged at 1,400 steps. For Task B, a YOLOv8-L and a YOLOv11-X model were trained for 600 epochs. To manage the severe memory bottlenecks of deploying a 7B VLM concurrently with YOLO models, we utilized `vLLM`, restricting its GPU memory utilization to 70%.

4.3. Results

Our system demonstrates robust performance, substantially outperforming official baselines. Rigorous empirical observations on our local validation split guided the design of our category-specific pipelines.

Task A and Task C (OCR). Table 1 compares our fine-tuning strategy against official baselines. For printed text (Task A), our system establishes a new state-of-the-art (CER 0.0441), significantly reducing errors. To address the performance gap between Task A and Task C, we conducted a qualitative error analysis on the handwritten calligraphy samples. We observed that errors are predominantly driven by morphological challenges rather than semantic ones. The extreme cursive fluidity and severe stroke adhesion in historical calligraphy present significant visual ambiguities, heavily disrupting the spatial recognition sequence.

Subtask	System	Including Variants				Excluding Variants			
		CER ↓	NED ↓	F1 ↑	Score ↑	CER ↓	NED ↓	F1 ↑	Score ↑
Task A	Qwen2.5-VL-7B (Zero-shot)	0.1014	0.0947	0.9110	0.9037	0.1121	0.1054	0.9007	0.8931
	Qwen2.5-VL-7B (Instr. Tuning)	0.0618	0.0613	0.9430	0.9397	0.0685	0.0679	0.9364	0.9331
	Xunzi-Qwen2-VL-7B (Zero-shot)	0.1786	0.1740	0.8409	0.8282	0.1851	0.1802	0.8345	0.8218
	Xunzi-Qwen2-VL-7B (Instr. Tuning)	0.1214	0.1183	0.8993	0.8854	0.1264	0.1232	0.8945	0.8805
	Ours (LoRA Fine-Tuned)	**0.0441**	**0.0435**	**0.9589**	**0.9569**	**0.0451**	**0.0446**	**0.9579**	**0.9559**
Task C	Qwen2.5-VL-7B (Zero-shot)	0.1207	0.1193	0.8849	0.8812	0.1338	0.1324	0.8718	0.8681
	Qwen2.5-VL-7B (Instr. Tuning)	0.0920	0.0919	0.9099	0.9086	0.1066	0.1065	0.8953	0.8940
	Xunzi-Qwen2-VL-7B (Zero-shot)	0.1497	0.1492	0.8538	0.8514	0.1609	0.1604	0.8425	0.8402
	Xunzi-Qwen2-VL-7B (Instr. Tuning)	0.1383	0.1376	0.8673	0.8635	0.1520	0.1512	0.8534	0.8498
	Ours (LoRA Fine-Tuned)	**0.0793**	**0.0791**	**0.9222**	**0.9212**	**0.0917**	**0.0915**	**0.9098**	**0.9088**

Table 1: Comprehensive evaluation results for OCR subtasks (Task A and Task C) on the official test set. Our parameter-efficient fine-tuning strategy consistently outperforms all official baselines across both variant settings.

System	Method	mAP@[.5:.95] ↑	Micro F1 ↑	Macro F1 ↑	Avg Match IoU ↑
Qwen2.5-VL-7B	Zero-shot	0.0000	0.0000	0.0000	0.0000
Qwen2.5-VL-7B	Instr. Tuning	0.2006	0.0513	0.1530	0.6600
Xunzi-Qwen2-VL-7B	Zero-shot	0.0236	0.0003	0.0114	0.5943
Xunzi-Qwen2-VL-7B	Instr. Tuning	0.1917	0.0403	0.1130	0.6654
Ours	**Hybrid Multi-Stage Pipeline**	**0.5118**	**0.7553**	**0.7879**	**0.7740**

Table 2: Performance metrics for Layout Detection (Task B). Our hybrid pipeline, which integrates YOLO ensemble detection with VLM-based semantic verification, substantially outperforms all official baselines.

Task B (Layout Detection). As shown in Table 2, our hybrid pipeline yields an mAP@[0.5:0.95] of 0.5118. Compared to the best-performing baseline (0.2006 mAP), our system achieves a relative improvement of over 155%. This substantial improvement validates the efficacy of our category-aware geometric cleaning and conflict resolution module.

5. Conclusion

In this paper, we presented a highly effective, multi-stage system for the EvaHan2026 shared task. For text recognition (Tasks A and C), applying parameter-efficient LoRA fine-tuning to Qwen2.5-VL on full-resolution images successfully preserves crucial spatial context, achieving highly competitive CERs. For layout analysis (Task B), we developed a category-aware hybrid framework integrating a YOLO ensemble, a specialized seal pipeline, and a VLM-based semantic verifier. This approach successfully mitigated the high false-positive rates typical of historical documents, highlighting the immense potential of combining robust visual detection with VLM semantic reasoning.

6. Limitations and Future Work

Several limitations remain for future exploration. First, while our models train on pre-cropped columns (Tasks A/C), future work will target end-to-end full-page OCR to improve generalization in uncropped scenarios. Second, our deterministic spatial conflict rules are highly optimized for the EvaHan2026 dataset; learning-based spatial relation modules could enhance zero-shot generalizability. Third, expanding the VLM semantic verifier beyond the *text* category to caption or verify other layout elements remains a promising direction. Finally, addressing label noise and generating more diverse synthetic data for stylistically homogeneous categories like seals will further bridge domain gaps.

7. Acknowledgements

We gratefully acknowledge the EvaHan2026 organizers for providing the standardized benchmark and dataset.

8. Ethics Statement

This research aids the automated digitization of historical Chinese literature. All datasets utilized were officially provided by the EvaHan2026 organizers and consist entirely of ancient, public-domain texts containing no sensitive information. The deployment of this system poses no negative societal impacts.

9. Bibliographical References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understand-

ing, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. 2023. Nougat: Neural optical understanding for academic documents. *arXiv preprint arXiv:2308.13418*.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Minghao Li, Tengchao Lv, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2021. Trocr: Transformer-based optical character recognition with pre-trained models. *arXiv preprint arXiv:2109.10282*.

Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. 2020. Real-time scene text detection with differentiable binarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11474–11481.

Zejiang Shen, Ruochen Zhang, Melissa Dell, Benjamin Charles Germain Lee, Jacob Carlson, and Weining Li. 2021. Layoutparser: A unified toolkit for deep learning based document image analysis. In *Document Analysis and Recognition – ICDAR 2021*, pages 131–146. Springer.

Connor Shorten and Taghi M. Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60. Article 60.

Ultralytics. 2023. Ultralytics yolov8 docs. https://docs.ultralytics.com/models/yolov8/. Official documentation.

Ultralytics. 2024. Ultralytics yolo11 docs. https://docs.ultralytics.com/models/yolo11/. Official documentation.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1192–1200.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410. Association for Computational Linguistics.

A Multi-Modal Recognition Framework for Ancient Books Integrating DoRA-DPO Text Recognition and YOLO Layout Analysis

Chaokun Zhang[1,*], **Xin Wen**[2,*], **Tongtong Zhou**[3,*]
[1]Chinese Classics Research Institute, Fudan University
[2]College of Computer Science and Artificial Intelligence, Fudan University
[3]Center for Historical Geographical Studies, Fudan University
Shanghai, China
{25210100011, 25213050398, 24210760017}@m.fudan.edu.cn

Abstract

The digitization and intelligent analysis of ancient Chinese documents face significant challenges due to diverse scripts, complex layouts, and the prevalence of rare characters. We present a comprehensive multi-modal recognition framework developed for the closed-modality track of the EvaHan 2026 Ancient Chinese Document Multi-Modal Recognition Shared Task. Our approach integrates two specialized pipelines to address these complexities. For text recognition (Tasks A and C), we propose a high-precision OCR system based on the domain-adapted Xunzi_Qwen2_VL_7B_Instruct, leveraging DoRA within a two-stage progressive curriculum learning strategy. To further refine character accuracy, DPO is incorporated alongside a dual-adapter architecture for rare character error localization and correction. For layout detection (Task B), we implement DocLayout-YOLO, enhanced by domain-specific pre-training and Mosaic augmentation to achieve efficient NMS-free element detection. Furthermore, a multi-round robust inference strategy, featuring automatic retry mechanisms and multi-prompt brute-force search, is introduced to handle stubborn and degraded samples effectively. Experimental results demonstrate that our proposed framework achieves superior performance across all evaluation metrics, highlighting its robustness and effectiveness in the digital preservation of ancient Chinese heritage.

Keywords: EvaHan 2026, Ancient Chinese OCR, VLM Fine-tuning, Curriculum Learning, Layout Analysis

1. Introduction

Because ancient Chinese literature constitutes a precious cultural heritage spanning thousands of years, its digital preservation has become a core topic in humanities computing. OCR is key for ancient book digitization, but faces challenges: diverse font styles, rare and variant characters, complex layouts, and degraded image quality.[1]

Recent multimodal large language models (MLLMs) like Qwen2-VL (Wang et al., 2024b) provide new pathways for ancient book OCR. Parameter-efficient methods like LoRA (Hu et al., 2022) and DoRA (Liu et al., 2024) lower adaptation costs. However, MLLMs often struggle with fine-grained spatial coordination and precise element positioning. In the realm of document layout analysis, specialized machine learning frameworks—like those based on the YOLO (Wang et al., 2024a) architecture—remain superior.

We present an integrated multimodal framework specifically optimized for the EvaHan 2026 Track. Our primary contributions are summarized as follows:

- **Curriculum-based DoRA & Dual-Adapter Architecture:** We propose a two-stage progressive fine-tuning strategy to transition knowledge from woodblock prints to handwritten scripts, utilizing DoRA and dual-adapters for efficient cross-domain adaptation.
- **DPO-driven Refinement via Local Tiling:** We apply DPO for character-level correction. Combined with local tiling, this approach significantly enhances precision in recognizing rare and variant characters.
- **Robust Multi-round Inference & Layout Detection:** We implement a high-reliability pipeline featuring automatic retries and multi-prompt search, integrated with a domain-pretrained DocLayout-YOLOv10 for superior end-to-end performance.

2. Related Works

Ancient Book OCR. Traditional methods rely on CNN and sequence modeling (Lombardi and Marinai, 2020). Kim et al. (Kim et al., 2022) introduce Donut, the first OCR-free end-to-end document understanding Transformer. Li et al. (Li et al., 2022)

[*]Authors contributed equally to this work.
[1]The source code for the proposed framework is publicly available at: https://github.com/leeoisaboy/paleomindocr-evahan2026

Proceedings of LT4HALA 2026 @ LREC 2026, pages 268–272.
11 May 2026. *©ELRA Language Resources Association (ELRA), 2026*

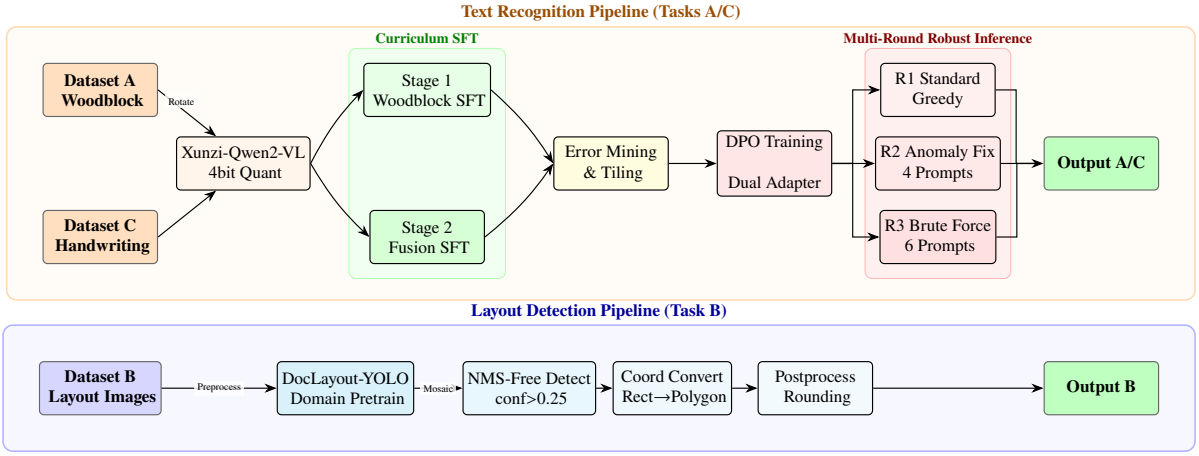


Figure 1: Complete architecture of the closed-modality system

propose DiT, a self-supervised pre-trained document image Transformer. However, VLM-based methods require domain-specific adaptation for ancient books.

Parameter-Efficient Fine-Tuning. Houlsby et al. (Houlsby et al., 2019) introduce Adapter modules. Hu et al. (Hu et al., 2022) propose LoRA, injecting low-rank matrices to reduce memory requirements. Liu et al. (Liu et al., 2024) develop DoRA, decomposing weights into magnitude and direction components for better learning.

Preference Optimization. Ouyang et al. (Ouyang et al., 2022) present RLHF for alignment, but it is complex. Rafailov et al. (Rafailov et al., 2023) introduce DPO, bypassing reward modeling to directly optimize policy models. We apply DPO to character-level OCR error correction.

Document Layout Analysis. Wang et al. (Wang et al., 2024a) propose YOLOv10, eliminating NMS dependency for end-to-end detection. Zhao et al. (Zhao et al., 2024) develop DocLayout-YOLO with domain pre-training. Torres-Aguilar (Aguilar, 2025) shows YOLO variants generalize better on complex layouts.

3. System Approach

3.1. System Overview

The overall architecture of the closed-modality framework is shown in Figure 1, consisting of two independent pipelines.

Text Recognition Pipeline (Tasks A/C): Built upon Xunzi_Qwen2_VL_7B_Instruct, the model undergoes two-stage DoRA curriculum learning SFT, followed by DPO preference optimization for rare character refinement, and finally multi-round robust inference strategy to generate final predictions.

Layout Detection Pipeline (Task B): Based on DocLayout-YOLOv10, initialized with domain-specific pre-trained weights and trained with Mosaic augmentation, achieving fast and accurate layout element localization through confidence-based filtering.

3.2. Text Recognition Pipeline (Tasks A & C)

3.2.1. Base Model and DoRA Configuration

The text recognition pipeline uses Xunzi_Qwen2_VL_7B_Instruct as the base model. The model is based on domain-adapted pre-training on classical Chinese corpora based on Qwen2-VL-7B, providing stronger prior understanding of ancient book characters. We apply 4-bit NF4 quantization with double quantization and BF16 compute precision for training on available hardware.

DoRA adapters cover all Attention layers and MLP layers in the Transformer. Image resolution is uniformly set to 600 px to balance stroke detail preservation and memory constraints. Horizontal strip images from Datasets A and C are rotated 90° clockwise before input to match the traditional vertical reading direction of ancient books.

3.2.2. Two-Stage Progressive Curriculum Learning SFT

We implement a two-stage curriculum strategy to incrementally expand the model's capabilities.

Stage 1: Woodblock Baseline. Using Dataset A exclusively with specific prompts (e.g., "【雕版印刷】..."), we establish a foundational recognition capability for structured woodblock texts. Training proceeds for 2 epochs with an early stopping mechanism (loss $<$ 0.8 for 10 log entries).

Stage 2: Handwriting Generalization. Dataset C is integrated to introduce complex styles. To balance difficulty, Dataset C is oversampled by $2\times$, re-

sulting in a 1:2 ratio between A and C. This stage employs 2 epochs with a 768-token sequence length to generalize from rigid fonts to fluid handwriting.

Knowledge Retention: This strategy ensures a smooth transition without triggering significant catastrophic forgetting. By employing a mixed-training strategy where woodblock data continuously participates in Stage 2, the model maintains its proficiency in printed styles. Furthermore, DoRA freezes base weights W_0 and only updates the low-rank component ΔW, strictly limiting the overwriting of foundational Stage 1 knowledge. Empirical observations confirm that woodblock recognition accuracy remains stable throughout the handwriting fine-tuning phase.

3.2.3. DPO Preference Optimization for Rare Character Refinement

To rectify persistent errors on rare characters, we introduce DPO through three steps:

(1) Diagnostic Analysis. The Stage 2 model conducts full inference to identify misrecognitions, generating structured error logs pairing ground truth with erroneous predictions.

(2) Tiling and Triples. For each error, we extract a 600×600 pixel local tile. The crop center is fixed at the horizontal center of the rotated image, with the vertical position proportional to the character's text index. Out-of-bounds regions are padded. DPO triples are then constructed: *Prompt* contains the tile; *Chosen* is the ground truth (target character with 3-character context); and *Rejected* is the model's error.

(3) Dual-Adapter Training. Policy and reference adapters are initialized from SFT weights on a 4-bit base. The DPO loss is:

$$\mathcal{L}_{\text{DPO}} = -\log \sigma\Big(\beta\big[\log \tfrac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \log \tfrac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\big]\Big) \quad (1)$$

where y_w is the preferred response, y_l is the rejected response, and $\beta = 0.1$ is the temperature coefficient.

Two key designs for DPO success. (1) an ultra-low learning rate (5×10^{-7}) and tiling to isolate local details from signal dilution; and (2) the frozen reference adapter acting as a semantic anchor, whose implicit KL-divergence constraint ensures localized refinements remain consistent with the global SFT linguistic prior.

3.2.4. Multi-Round Robust Inference Strategy

After DPO training completion, adapter weights are merged into the base model to eliminate PEFT inference overhead, and KV Cache is enabled for acceleration. Inference loads in full precision to maximize recognition accuracy.

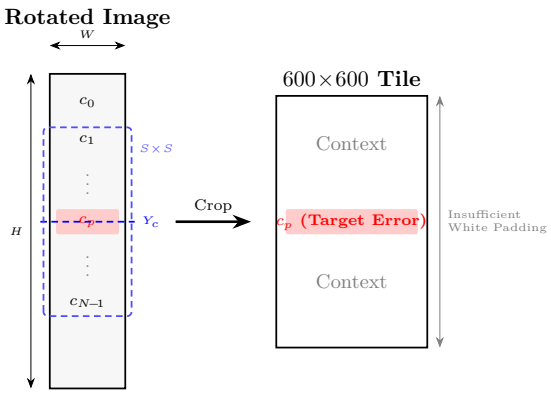


Figure 2: Tiling proportional localization and crop.

To handle various difficult samples, we design a three-round progressive inference strategy:

Round 1 —Standard Inference. Deterministic decoding is performed on all images, retaining only pure CJK characters after post-processing filtering. If output contains invalid keywords or is empty, automatic retry mechanism is triggered to improve success rate.

Round 2 —Anomaly Pattern Specific Prompt Inference. First round results are scanned to identify three anomaly patterns: empty string output, invalid keywords (such as "圖頁" meaning figure page, etc.), and repetitive gibberish (short pattern cycling, long pattern repetition, or extremely few character types in long text). For anomalous samples, 4 different Prompt strategies are used for re-inference, selecting the best non-anomalous output.

Round 3 —Exhaustive Multi-Prompt Inference. For challenging samples, we employ six prompt variants and select the longest valid output to maximize completeness, addressing character omission as the primary failure mode in long-text OCR. To ensure validity, outputs must pass a three-tier diagnostic filter: (1) empty string detection, (2) predefined noise keyword filtering, (3) repetitive gibberish pattern recognition. This heuristic is more robust than majority voting for ancient book corpora, as the high cardinality of the Chinese character set makes identical recognition errors rare, rendering voting-based aggregation less effective. Detailed prompt variants are provided in Appendix A.

The three-round progressive strategy reduces the failure rate of empty output or obvious hallucination in first-round inference to near zero. The six prompt variants describe task requirements from different angles, effectively alleviating model sensitivity to specific wording.

Table 1: Results on Text Recognition Tasks (Tasks A & C).

Task	Models	Consider variant characters				Don't consider variant characters			
		CER	NED	F1	Comp.	CER	NED	F1	Comp.
A	Qwen2.5_VL_7B_Instruct	0.1014	0.0947	0.9110	0.9037	0.1121	0.1054	0.9007	0.8931
	Xunzi_Qwen2_VL_7B_Instruct	0.1786	0.1740	0.8409	0.8282	0.1851	0.1802	0.8345	0.8218
	Ours (DoRA+DPO)	**0.0798**	**0.0795**	**0.9270**	**0.9223**	**0.0872**	**0.0869**	**0.9195**	**0.9149**
C	Qwen2.5_VL_7B_Instruct	0.1207	0.1193	0.8849	0.8812	0.1338	0.1324	0.8718	0.8681
	Xunzi_Qwen2_VL_7B_Instruct	0.1497	0.1492	0.8538	0.8514	0.1609	0.1604	0.8425	0.8402
	Ours (DoRA+DPO)	**0.1322**	**0.1320**	**0.8723**	**0.8692**	**0.1438**	**0.1436**	**0.8609**	**0.8576**

Table 2: Results on Task B (Layout Detection).

Models	mAP@[.5:.95]	Micro F1	Macro F1	IoU
Qwen2.5_VL_7B_Instruct	0	0	0	0
Xunzi_Qwen2_VL_7B_Instruct	0.0236	0.0003	0.0114	0.5943
Ours (YOLO-based)	**0.3889**	**0.6401**	**0.6318**	**0.8119**

4. Experiments

The EvaHan 2026 shared task provides three datasets: Dataset A (woodblock-printed texts), Dataset B (page-level layout annotations), and Dataset C (handwritten texts). Datasets A and C are rotated 90 degrees during training to align with vertical reading direction.

Table 1 presents text recognition results against two baselines without instruction tuning.

Task A: Woodblock Printing. Our system achieves character error rate of 0.0798 (considering variant characters), substantially outperforming Qwen2.5 (0.1014) and Xunzi (0.1786). The comprehensive score reaches 0.9223 with F1 of 0.9270. Without variant consideration, our system maintains 0.0872 error rate and 0.9149 comprehensive score.

Task C: Handwritten Text. Our system achieves 0.1322 character error rate with 0.8692 comprehensive score (considering variants). Qwen2.5 achieves 0.1207 error and 0.8812 score, while Xunzi achieves 0.1497 and 0.8514. Our method shows competitive performance.

Layout Detection. Table 2 shows results. Our YOLO system achieves mAP of 0.3889, micro F1 of 0.6401, and IoU of 0.8119 (37% improvement over Xunzi), substantially outperforming both baselines.

5. Limitations

Despite the robust performance of our framework, several limitations remain. First, the layout detection pipeline (Task B) faces cross-format migration limits; relying on domain-specific pre-training makes it less effective on unconventional layouts like accordion-style bindings or stone rubbings not present in the training set. Second, the text recognition pipeline shows a performance gap on Task C, where highly cursive or severely degraded handwritten scripts challenge the model's understanding of expressive calligraphic strokes. Finally, the multi-round inference strategy improves accuracy at the cost of computational efficiency, introducing latency that limits high-throughput, real-time digitization without further model distillation.

6. Bibliographical References

Sergio Torres Aguilar. 2025. From codicology to code: A comparative study of transformer and yolo-based detectors for layout analysis in historical documents. *arXiv preprint arXiv:2506.20326*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer.

Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang, and Furu Wei. 2022. Dit: Self-supervised pre-training for document image transformer. In *Proceedings of the 30th ACM international conference on multimedia*, pages 3530–3539.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*.

Francesco Lombardi and Simone Marinai. 2020. Deep learning for historical document analysis and recognition—a survey. *Journal of Imaging*, 6(10):110.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, et al. 2024a. Yolov10: Real-time end-to-end object detection. *Advances in neural information processing systems*, 37:107984–108011.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Zhiyuan Zhao, Hengrui Kang, Bin Wang, and Conghui He. 2024. Doclayout-yolo: Enhancing document layout analysis through diverse synthetic data and global-to-local adaptive perception.

A. Core Recognition Prompts

Our system employs a hierarchical set of instructions to ensure robust recognition. To avoid character encoding issues and maintain visual consistency, all prompts are delivered in clear, instruction-oriented Chinese.

- **Standard Prompt:** "请仔细观察这张古籍图片，从上到下、从右到左，一个字一个字地识别并输出。要求：1. 每识别一个字就输出一个字；2. 只输出汉字，不要输出任何标点或说明；3. 如果某个字看不清，根据上下文推断；4. 绝对不要输出' 圖頁' 这两个字。"
- **Retry Prompt:** "这张图片包含古籍文字内容（不是空白页或图页标记）。请仔细辨认图中的每一个汉字，按阅读顺序逐一输出。输出格式：直接输出连续的汉字，无标点无空格。"
- **Anomaly Repair Prompts:**
  1. "请逐字识别图中的古籍文字，只输出文字，不要输出任何其他内容。"
  2. "这是一张古籍图片，请仔细辨认每一个汉字并输出，只输出识别到的文字。"
  3. "请识别图片中的所有汉字，按照从右到左、从上到下的顺序输出纯文字。"
  4. "图中是古籍文字，请精确识别并输出，注意笔画细节，只输出汉字。"

Beijing Normal University at EvaHan 2026: Enhancing Ancient Chinese Character Recognition and Layout Analysis via VLM Fine-Tuning and Linguistic Post-Processing

Yihuan Yin[1], **Qian Zhao**[1]
[1]Institute of Chinese Information Processing,
School of International Chinese Language Education, Beijing Normal University
{202421091021, 202421091022}@mail.bnu.edu.cn

Abstract

This paper describes the system submitted by the Beijing Normal University (BNU) team for the EvaHan 2026 shared task. We participated in Task A (Printed Text Recognition), Task B (Layout Element Analysis), and Task C (Handwritten Character Recognition). For text recognition (Tasks A and C), we proposed a hybrid pipeline combining supervised fine-tuning (SFT) of Vision-Language Models (VLMs) with a linguistic rule-based post-processing module. In the Open Track, we further explored the use of a general-purpose VLM to correct semantic errors while maintaining visual fidelity to ancient variant characters. For Task B, we adopted a method integrating a VLM with structured prompting strategies. Our system consistently surpassed the official baselines, achieving an F1 score of 94.53% in Task A and 91.33% in Task C, while demonstrating enhanced localization precision in Task B.

Keywords: Ancient Chinese, OCR, Layout Analysis, Vision-Language Models, LoRA

1. Introduction

The digitization of ancient Chinese literature poses significant challenges due to the presence of variant characters, complex layouts, and idiosyncratic handwritten cursive styles. The EvaHan 2026 shared task provides a comprehensive platform to evaluate multimodal systems on historical documents across three tracks: Task A, Task B, and Task C.

Our team from BNU approached this challenge by leveraging the recent advancements in multimodal large language models (MLLMs). For text recognition, we focused on addressing the ”hallucination” and ”over-standardization” issues where models arbitrarily convert ancient variant characters into modern standard forms. For layout analysis (Task B), we concentrated on multi-category structural detection in mixed image-text pages under the closed-track setting. Without using any external data, our system incorporates domain-specific fine-tuning and linguistic deterministic rules to ensure both semantic accuracy and structural validity.

2. Related Works

Document layout analysis has evolved from traditional image processing techniques to deep neural network–based approaches. Early methods relied on projection profiles and connected component analysis for page segmentation (Nagy, 2000), achieving reasonable performance in documents with regular layouts. With the development of deep learning, general object detection architectures such as Faster R-CNN and Mask R-CNN demonstrated stronger performance in document region detection tasks (Ren et al., 2015; He et al., 2017).

To further integrate textual content and spatial layout information, Transformer-based multimodal models such as LayoutLM (Xu et al., 2020) jointly model OCR text, positional embeddings, and visual features, significantly improving performance. However, these methods are primarily designed for modern printed documents. The diversity and noise characteristics of ancient documents limit the direct transferability of such models. Studies on historical document analysis have introduced U-Net–based segmentation frameworks (Oliveira et al., 2018) and multi-scale feature fusion strategies, but multi-category layout detection with structured output remains challenging.

Furthermore, while VLMs demonstrate strong multimodal alignment capabilities (Bai et al., 2023), their application combined with parameter-efficient fine-tuning (e.g., LoRA (Hu et al., 2022)) and structured post-processing in ancient book layout detection remains relatively underexplored. This study aims to bridge this gap, aligning with the evaluation objectives of the EvaHan community (Li et al., 2024).

3. Methodology

Our system processes inputs through a structured pipeline depending on the task and evaluation track.

Proceedings of LT4HALA 2026 @ LREC 2026, pages 273–276.
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

3.1. Text Recognition (Tasks A and C)

In the Closed Track, we utilized the Xunzi-Qwen2-VL-7B-Instruct model (Wang et al., 2023). To mitigate structural OCR errors, we implemented a linguistic deterministic rule-based post-processing module. This included merging split components and fixing context-specific terminology errors based on historical official titles.

For the Open Track, we employed Doubao-1.5-Pro (Vision) as a semantic correction agent with a "Conservative Correction Strategy." The prompt explicitly instructed the model to preserve handwritten variants to maintain fidelity to the original manuscripts, while correcting genuine semantic errors based on historical contexts.

3.2. Layout Element Analysis (Task B)

Task B focuses on layout element recognition. The core objective is the precise identification of four key elements within the pages of ancient books: text, image, book_edge, and seal. Under the closed-track setting, all participating systems are required to use the officially specified models. In our implementation, we adopted Qwen2.5-VL-7B-Instruct as the backbone model. Based on this architecture, we constructed a layout detection framework and enhanced its adaptability to ancient book layout analysis through parameter-efficient fine-tuning using LoRA.

3.3. Implementation Details

To ensure reproducibility and comprehensively address the evaluation criteria, we detail the hyperparameter configurations, hardware environment, and post-processing strategies below.

Parameter / Component	Configuration / Version
Fine-tuning Method	LoRA (Scope: all linear layers)
LoRA Alpha / Dropout	16 / 0.0
Learning Rate	2×10^{-4} (Cosine Scheduler)
Warmup / Weight Decay	0 steps / 0.0
Max Grad Norm	1.0
Epochs / Precision	3 / `fp16` mixed precision
Effective Batch Size	8 (Per-device: 1, Accumulation: 8)
Hardware Configuration	NVIDIA GeForce RTX 4090 D $\times$ 1
Peak VRAM Consumption	$\sim$20–22 GB
CPU / Memory	15 Cores / 80 GB
Software Environment	Python 3.11.14, CUDA 13.0
Core Libraries	PyTorch 2.6.0+cu124
	Transformers 4.57.1, vLLM 0.4.0
	LLaMA-Factory 0.9.5.dev0

Table 1: System Environment and Fine-Tuning Hyperparameters for VLM Adaptation.

VLM Fine-Tuning Configurations: As detailed in Table 1, the core visual-language models were fine-tuned using the LLaMA-Factory framework. Although `bf16` is often preferred for fine-tuning Qwen2.5-VL due to its wider numerical range and better stability, we utilized `fp16` mixed precision based on our specific hardware constraints and environment compatibility. To mitigate potential gradient overflow issues inherent to `fp16`, we applied strict gradient clipping (max norm $= 1.0$) and dynamic loss scaling. In practice, no gradient overflow or obvious numerical instability was observed, and the model completed training stably across all epochs.

Open Track Prompting Strategy: For generative post-correction in the open track, we designed track-specific prompts to constrain the VLM. In Task A, the prompt enforced a "visual fidelity" rule, strictly prohibiting the conversion of historical colloquial characters (e.g., 万, 无) into standard traditional forms unless visually justified. In Task C, the prompt instructed the model to trust the base OCR for rare entities while forcing it to correct financial logic and numeral errors based on visual evidence.

Closed Track Linguistic Post-Processing: We introduced a deterministic linguistic post-processing module involving a dual-step pipeline: (1) Aggressive format cleaning using regular expressions to strip out all punctuation and full-width spaces. (2) Hard-coded symbolic overrides tailored to specific historical contexts, correcting high-confidence misrecognitions related to official titles and financial terms (e.g., 大住 to 大臣) via a predefined expert lexicon. Crucially, these rules were not arbitrary; they were systematically derived through a manual error analysis of the base model's initial predictions, deeply integrated with domain-specific knowledge of ancient Chinese culture and administrative history.

4. Experiments and Results

4.1. Dataset

The dataset provided by the EvaHan 2026 organizers contains mixed image-text data selected from the Siku Quanshu and other ancient books. Each sample consists of an ancient book page image and the corresponding official annotation in JSON format. The annotations define four structural entity categories (text, image, seal, book_edge), with each entity including its category label and four coordinate points ordered as top-left, top-right, bottom-right, and bottom-left.

4.2. Main Results

The quantitative evaluation of our proposed system on the EvaHan 2026 test set is presented in Table 2.

Track & System	Task A F1 (%)↑	Task B Macro F1 (%)↑	Task C F1 (%)↑
Official Baseline	94.30	**15.30**	90.99
Our System (Closed)	**94.53**	11.73	**91.33**
Our System (Open)	-	-	-

Table 2: Performance comparison on the EvaHan 2026 test set. F1 (%) indicates the F1 score in percentage. (Note: Official quantitative results for the Open Track are currently pending; thus, our Open Track submission is discussed primarily as a qualitative exploration of VLM capabilities.)

Text Recognition (Tasks A and C): The official baseline for Task A (Printed Texts) yielded an F1 score of 94.30%. Our system achieved an F1 score of 94.53%, outperforming the baseline. For Task C (Handwritten Texts), our proposed pipeline achieved an F1 score of 91.33%, surpassing the baseline's 90.99%, demonstrating robustness on highly personalized cursive connections.

Layout Element Analysis (Task B): Under the closed-track setting, our system achieved an Avg Match IoU of 0.7048, a Micro F1 of 7.14%, and a Macro F1 of 11.73%. As shown in Table 2, while the official baseline achieved a higher Macro F1 of 15.30% and a higher mAP@[.5:.95] (0.2006 vs. our 0.1402), our system significantly improved the average matching IoU (from 0.6600 to 0.7048). This indicates more precise bounding box localization for detected elements, albeit with a reduced overall detection performance.

4.3. Ablation Study and Error Analysis

To verify the contributions of our proposed modules and analyze the performance bottlenecks mentioned above, we conducted qualitative ablation studies and error analyses across the tasks.

Impact of Linguistic Post-Processing and Prompting (Tasks A & C): Removing the linguistic post-processing module resulted in a noticeable resurgence of deterministic terminology errors (e.g., historical official titles being misclassified due to modern linguistic priors). The structured prompting strategy was equally critical; without the "visual fidelity" constraints, the VLM exhibited severe normalization hallucination, arbitrarily converting historical variant characters into standardized modern forms and disrupting the original layout formatting. These two modules were the primary drivers behind our consistent F1 improvements over the baseline.

Small Object Detection Bottleneck (Task B): Our layout analysis pipeline currently lacks targeted designs for small object detection, which directly explains the severe long-tailed class imbalance. Specifically, the seal category suffered from near-zero recall. Seals in ancient texts are inherently minority classes, often appearing as small, faded, and low-contrast regions. Because our generic VLM fine-tuning approach did not incorporate specialized small-object mechanisms (such as multi-scale feature fusion or dynamic anchor oversampling), the model prioritized high-frequency, large-area objects (like text blocks) to minimize the global loss. This systematic degradation in minority categories significantly pulled down the overall Macro F1 score, despite the high localization precision (IoU) achieved on text elements.

5. Discussion

The experimental results for Task B demonstrate a clear improvement in spatial localization capability but reveal a structural imbalance at the category level. The increase in IoU indicates that LoRA fine-tuning effectively enhances the model's geometric modeling of page structure, producing tighter and more accurate bounding boxes.

As analyzed in the ablation study, further manual inspection reveals that VLMs tend to adopt a conservative prediction strategy, refraining from outputting detections when confidence is insufficient in order to reduce false positives. While this strategy improves localization precision for large objects, it results in systematic recall loss for small objects like seals. Moreover, minimum-size filtering rules may further suppress seal detections, as small bounding boxes shrink excessively during discretization. Therefore, the system exhibits a characteristic pattern of "high localization precision but insufficient category coverage." If seal recall can be improved while maintaining the current IoU level, both Macro F1 and mAP are expected to increase.

6. Conclusion

We successfully developed a hybrid VLM-based pipeline for ancient Chinese texts. By combining LoRA fine-tuning with targeted linguistic post-processing, our system surpassed the official baselines in character recognition tasks. For layout analysis, utilizing structured prompts and post-processing constraints yielded tighter bounding box localization. Future work will explore integrating Retrieval-Augmented Generation (RAG) and specialized small-object detection strategies to improve minority category recall.

7. Bibliographical References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jing Jing. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading and beyond. *arXiv preprint arXiv:2308.12966*.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Bin Li et al. 2024. Overview of evahan 2024: The evaluation of ancient chinese language processing. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

George Nagy. 2000. Twenty years of document image analysis in pami. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):38–62.

Sofia Ares Oliveira, Benoit Seguin, and Frederic Kaplan. 2018. dhsegment: A generic deep-learning approach for document segmentation. In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 7–12. IEEE.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Dongbo Wang et al. 2023. Xunzi: Open-source large language model for chinese ancient text processing. Model open-sourced on GitHub and ModelScope.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1192–1200.

A Dual-Modality Framework for Ancient Document Layout Analysis and Text Recognition

Qi Fan*, **Jieming Hu***, **Chen Ye**†
Department of Computer Science and Technology, Tongji University, Shanghai, China
*These authors contributed equally to this work.
†Corresponding author.
{2350262, 2351584, yechen}@tongji.edu.cn

Abstract

The digital preservation of ancient Chinese literature requires robust capabilities spanning layout analysis and text recognition. This paper presents a comprehensive framework addressing two fundamental challenges: (1) **Layout Element Analysis (Task B)** for detecting page elements (text, image, book_edge, seal) amidst degradation, nested structures, and extreme class imbalance; and (2) **Text Recognition (Tasks A & C)** for end-to-end transcription of printed and handwritten classical documents. For layout analysis, we propose a dual-modality solution. The Closed Modality formulates this as a sequence-to-sequence problem using Vision-Language Models (VLMs), introducing spatial discretization tokenization and a Frequency-Aware Sequential Curriculum Learning framework with dynamic memory replay. The Open Modality presents HistLayout-DETR, a set prediction architecture integrating an Augmented Morphological Encoder and a Polygon Boundary Refinement head. For text recognition, we formulate OCR as a domain-constrained visual language generation task using Qwen2.5-VL with LoRA fine-tuning. We employ structured prompts encoding reading order and Traditional Chinese character preservation across domains. Extensive experiments on the EvaHan 2026 dataset validate our framework's superiority. In layout analysis, our curriculum-guided paradigm achieves a Macro F1 of 0.7992 and mAP@[.5:.95] of 0.5438. In text recognition, we achieve CERs of 0.0271 on printed and 0.0433 on handwritten texts.

Keywords: Ancient Document Layout Analysis, Text Recognition, Vision-Language Models, HistLayout-DETR, Curriculum Learning, Polygon Regression

1. Introduction

The digital preservation of ancient Chinese literature relies heavily on precise Layout Element Analysis and Text Recognition. The EvaHan 2026 benchmark introduces rigorous tasks for layout parsing (Task B) and OCR for printed and handwritten texts (Tasks A & C). However, extracting structural and semantic elements from historical scans remains profoundly challenging. Key obstacles include a severe long-tailed distribution of page elements (where text vastly outnumbers seals or book edges), slanted and nested non-axis-aligned structures, the absence of chromatic cues in degraded black-and-white (B&W) scans, vertical right-to-left reading orders, and the intricate preservation of rare Traditional Chinese characters across highly variable handwritten calligraphy.

To comprehensively address these challenges, we propose a robust, unified framework tailored for historical document understanding.

For **Layout Analysis (Task B)**, we present a dual-modality framework. In the **Closed Modality**, we formulate the task as an autoregressive sequence generation problem utilizing the Qwen2.5-VL backbone. We introduce a spatial discretization tokenization protocol to resolve coordinate scaling issues, alongside a novel Frequency-Aware Sequential Curriculum Learning framework. By integrating a dynamic curriculum replay loss with temporal decay and frequency-inverse penalties, our approach successfully mitigates dominant-class bias and catastrophic forgetting. For the **Open Modality**, we propose **HistLayout-DETR**, an end-to-end set prediction architecture. Meticulously engineered with an Augmented Morphological Encoder (featuring Adaptive Grain-Size Attention and Morphology-Aware Texture Enhancement), a Hierarchical Text-in-Image Decoder, and a Polygon Boundary Refinement (PBR) head, it excels at extracting multi-scale structural signatures and regressing precise, non-axis-aligned quadrilateral boundaries.

For **Text Recognition (Tasks A & C)**, we formulate classical Chinese OCR as a domain-constrained visual language generation task using large Vision-Language Models (VLMs). We employ a structured prompting framework that explicitly encodes reading order, script preservation, and uncertainty handling. Joint training across printed and handwritten domains enables the model to learn shared semantic priors while adapting to severe visual variability.

2. Related Work

Deep learning architectures, ranging from Convolutional Neural Networks (Xu et al., 2018; Redmon et al., 2016) to Transformers (Zhu et al., 2021; Banerjee et al., 2023), have significantly advanced historical document layout analysis and Chinese OCR (Shi et al., 2018; Baek et al., 2021). However,

Proceedings of LT4HALA 2026 @ LREC 2026, pages 277–287.
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

traditional layout detectors typically regress axis-aligned bounding boxes, which are inadequate for the slanted columns and nested structures (e.g., text within illustrations) prevalent in ancient scans. Recently, Vision-Language Models (VLMs) (Bai et al., 2023; Kim et al., 2022) have unified visual detection and text recognition into autoregressive generation tasks, inspired by spatial coordinate quantization (Chen et al., 2022). Despite their impressive capabilities (Liu et al., 2023; Li et al., 2023), directly applying VLMs to historical archives introduces severe Out-of-Vocabulary (OOV) errors due to varying digitization scales. Furthermore, ancient datasets exhibit extreme long-tailed distributions. While curriculum learning (Bengio et al., 2009) and continual adaptation (Rebuffi et al., 2017) can address data imbalance, naive sequential fine-tuning inevitably triggers catastrophic forgetting (McCloskey and Cohen, 1989; Kirkpatrick et al., 2017), and standard static memory replay often reinforces dominant-class bias.

3. Layout Analysis (Task B)

We formulate Layout Element Analysis (Task B)—detecting text, image, book_edge, and seal from an ancient document image $I \in \mathbb{R}^{H\times W\times 3}$ into bounding boxes B and labels C—via a dual-modality approach. For the Open Modality, we propose a specialized DETR-based architecture. For the Closed Modality, we cast the task as an autoregressive sequence-to-sequence generation problem using Vision-Language Models (VLMs). To efficiently fine-tune the VLM baseline, we apply Low-Rank Adaptation (LoRA) by injecting trainable rank decomposition matrices (rank=32, alpha=64, target_modules=all-linear) into the frozen backbone's transformer layers. The model is trained using bfloat16 precision for 3 epochs with a learning rate of 5e-5 and a 0.1 warmup ratio.

3.1. Paradigm I: Joint Instruction Fine-Tuning with Discretized Coordinate Normalization

In Paradigm I, we aggregate instances of all four layout categories for simultaneous joint fine-tuning. To enable the VLM backend to natively process spatial bounding boxes without architectural modifications, we map continuous, unbounded raw pixel coordinates into a standardized, resolution-agnostic spatial grid of $[0, 1000]$ discrete text tokens (see Figure 1). Directly feeding raw coordinates (e.g., $x = 219, y = 60$) induces severe out-of-vocabulary (OOV) issues and hinders the learning of generalized spatial relationships. By contrast, bounding coordinates to 1000 discrete bins creates a finite, learnable vocabulary of spatial tokens. This normalization effectively decouples geometric layouts from absolute physical dimensions to ensure resolution invariance, while striking an optimal density balance—fine enough to capture the minute details of dense text lines, yet coarse enough to prevent excessive embedding sparsity.

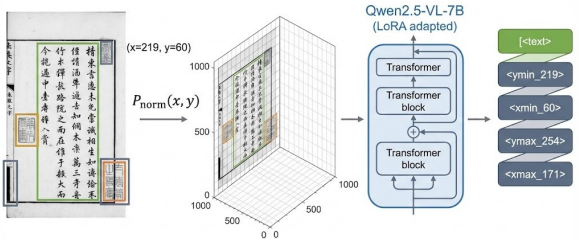


Figure 1: The proposed Spatial Discretization and Tokenization Framework (Paradigm I).

During data preprocessing, every absolute coordinate point $P(x, y)$ in the original image of width W and height H is normalized to $P_{norm}(x_{norm}, y_{norm})$:

$$\begin{aligned} x_{norm} &= \text{round}\left(\frac{x}{W} \times 1000\right) \\ y_{norm} &= \text{round}\left(\frac{y}{H} \times 1000\right) \end{aligned} \tag{1}$$

These normalized values are formatted into text prompts (e.g., [ymin, xmin, ymax, xmax]). During inference, the model autoregressively generates the normalized coordinate tokens. To map the predictions back to the original image space for evaluation, we apply an inverse coordinate restoration:

$$\begin{aligned} \hat{x} &= \text{round}\left(\hat{x}_{norm} \times \frac{W}{1000}\right) \\ \hat{y} &= \text{round}\left(\hat{y}_{norm} \times \frac{H}{1000}\right) \end{aligned} \tag{2}$$

3.2. Paradigm II: Frequency-Aware Sequential Curriculum Learning

Ancient layout elements exhibit a severe long-tailed distribution, causing standard joint training (Paradigm I) to over-optimize for the hyper-abundant text class while neglecting sparse minority elements (e.g., seals). To systematically address this bias, we propose a Frequency-Aware Sequential Curriculum Learning framework (Figure 2). We partition the dataset into class-specific subsets $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_4\}$ ordered by monotonically decreasing instance frequency ($N(c_1) > \cdots > N(c_4)$). Training proceeds sequentially over $K = 4$ stages, enabling the VLM to build a robust global structural foundation before specializing in localized, nuanced elements.

To mitigate catastrophic forgetting during sequential adaptation, we formulate a Dynamic Curriculum Replay Loss. At stage k, the model optimizes over

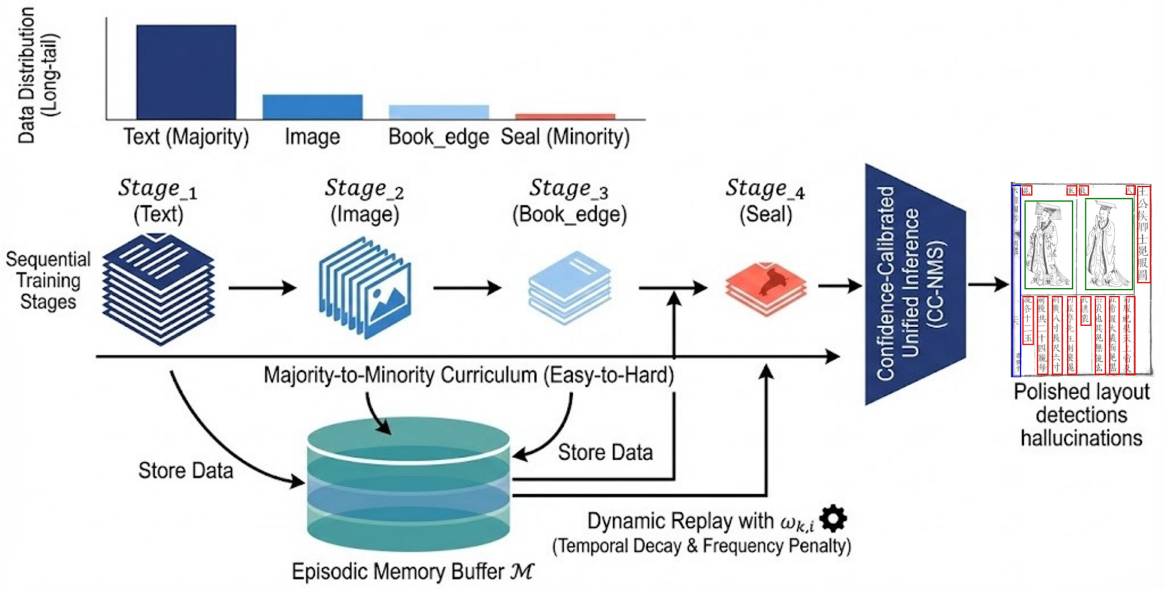


Figure 2: Frequency-Aware Sequential Curriculum Learning with Dynamic Memory Replay. To mitigate long-tail imbalance, training progresses sequentially from majority elements (text) to minority ones (seal).

both the target dataset $\mathcal{D}_k$ and an episodic memory buffer $\mathcal{M}_i$:

$$\mathcal{L}_{\text{total}}^{(k)}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}_k}\left[\mathcal{L}_{\text{CE}}(f_\theta(x), y)\right] + \sum_{i=1}^{k-1} \omega_{k,i} \cdot \mathbb{E}_{(x_m,y_m)\sim\mathcal{M}_i}\left[\mathcal{L}_{\text{CE}}(f_\theta(x_m), y_m)\right] \tag{3}$$

The dynamic retention weight $\omega_{k,i}$ for historical class i is defined as:

$$\omega_{k,i} = \lambda \cdot \exp(-\gamma(k-i)) \cdot \left(\frac{N(c_i)}{\max_{j<k} N(c_j)}\right)^{-\rho} \tag{4}$$

This dynamic weight balances plasticity and stability via a base replay coefficient λ, a temporal decay term $\exp(-\gamma(k-i))$ that relaxes the preservation of older concepts to free up parameters, and a frequency-inverse penalty governed by ρ. Crucially, this penalty down-weights historical classes with massive original sample sizes, preventing their replay gradients from swamping the optimization of minority classes.

Finally, because sequential training inherently biases confidence toward recent classes, we enforce confidence-calibrated unified inference. For a generated sequence $\hat{y}$ composed of coordinate tokens $\{v_1, \ldots, v_L\}$, we compute a length-normalized confidence score:

$$S_{\text{conf}}(\hat{y}|x) = \exp\left(\frac{1}{L}\sum_{t=1}^{L} \log P(v_t|v_{<t}, x; \theta^*)\right) \tag{5}$$

A predicted bounding box for class c_k is accepted only if $S_{\text{conf}}(\hat{y}|x) \geq \tau_{c_k}$, effectively filtering out hallucinations while retaining high-recall detection capabilities.

4. Text Recognition (Tasks A & C)

We address text recognition for both printed (Task A/keben) and handwritten (Task C/xieben) ancient Chinese documents by formulating OCR as a domain-constrained visual language generation problem. Given a document image $I \in \mathbb{R}^{H\times W\times 3}$, we utilize the Qwen2.5-VL-7B-Instruct vision-language model to autoregressively generate a text sequence T (Figure 3). To ensure historical validity, the generated sequence must preserve Traditional Chinese characters and adhere to a vertical right-to-left reading order. This constrained generation is modeled as:

$$\hat{T} = \arg\max_T P(T|I, C; \theta) \tag{6}$$

where C denotes explicit domain constraints restricting the output space. Rather than relying on generic instructions, we enforce these constraints via a unified structured prompt that explicitly dictates the reading direction, prohibits Simplified Chinese conversion, and mandates the use of '#' for unrecognizable characters.

To adapt the base model efficiently without catastrophic forgetting, we employ Low-Rank Adaptation (LoRA). Instead of updating the entire network, we freeze the pre-trained weights and inject trainable rank decomposition matrices into the transformer layers:

$$W' = W + \Delta W = W + BA \tag{7}$$

where $W \in \mathbb{R}^{d\times d}$ is the original frozen weight matrix, and $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times d}$ are the trainable matrices. We configure LoRA with rank $r = 32$

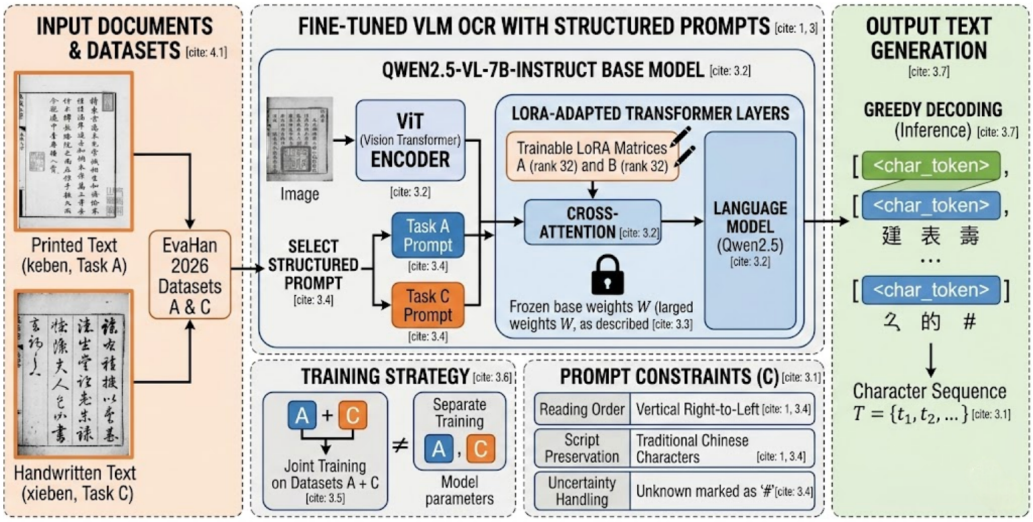


Figure 3: The proposed VLM-based text recognition framework. Historical document images are processed by Qwen2.5-VL with LoRA fine-tuning and structured prompts for domain-constrained OCR.

and $\alpha = 32$, targeting all linear layers in the attention and feed-forward networks. To maximize cross-domain robustness, we adopt a Joint Training strategy, combining both printed and handwritten datasets. This enables the VLM to learn shared lexical and semantic priors while adapting to severe glyph variability.

All models are fine-tuned for 3 epochs using a batch size of 16, a learning rate of 2×10^{-4}, and bfloat16 precision. During inference, we apply greedy decoding (generating up to 1024 tokens) to guarantee deterministic outputs. Finally, a strict post-processing pipeline is applied to strip any hallucinated explanatory text generated by the VLM, ensuring the output strictly contains the historically valid transcription.

5. Layout Analysis Experiments

This section evaluates our dual-modality framework on the EvaHan 2026 Dataset B. The dataset exhibits a severe long-tailed distribution dominated by *text* elements.

We strictly utilize the provided training corpora and designated pre-trained models: `Qwen2.5-VL-7B` and `Xunzi-Qwen2-VL`. Models are optimized via Swift LLM using LoRA (rank=32, alpha=64, target_modules=all-linear). Training runs in bfloat16 for 3 epochs with a batch size of 32, learning rate of 5e-5, and a 0.1 warmup ratio. Inference is deployed via vLLM (temperature=0.1, top-p=0.9, max_tokens=8192). We employ four official metrics: mAP@[.5:.95], Micro F1, Macro F1, and Average Matching IoU.

Table 1 benchmarks our paradigms against official zero-shot and standard Instruction Tuning (IT) baselines.

Zero-shot capabilities fail entirely on complex historical layouts. While Standard IT bridges this gap, our **Paradigm I** yields a massive leap, more than doubling the mAP and boosting the Qwen2.5-VL Macro F1 from 0.1530 to 0.7134 by resolving out-of-vocabulary spatial issues. **Paradigm II** further achieves state-of-the-art closed-modality performance (mAP 0.5438, Macro F1 0.7992), proving its efficacy in rebalancing learning across all sparse minority classes.

6. Text Recognition Experiments

We evaluate our domain-constrained OCR framework on EvaHan 2026 Task A (5,000 single-line printed *keben* images) and Task C (5,000 single-line handwritten *xieben* images). Following official protocols, performance is measured using Character Error Rate (CER), Normalized Edit Distance (NED), character-level F1 Score, and an Overall aggregated score. Models are fine-tuned via MS-SWIFT (LoRA rank=32, $\alpha = 32$) for 3 epochs with a 2×10^{-4} learning rate.

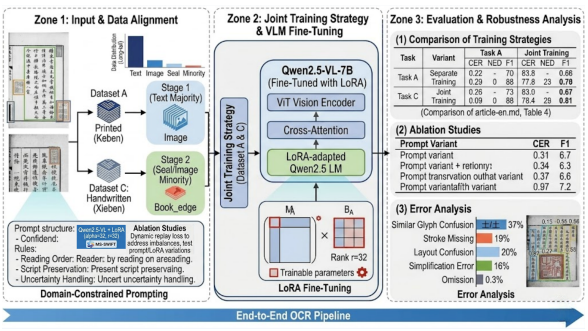


Figure 4: The unified text recognition pipeline utilizing joint LoRA fine-tuning for printed and handwritten tasks.

Base Model	Method	mAP@[.5:.95]	Micro F1	Macro F1	Avg Match IoU
`Qwen2.5-VL-7B`	Zero-shot	0.0000	0.0000	0.0000	0.0000
	Standard IT (Baseline)	0.2006	0.0513	0.1530	0.6600
	Ours (Paradigm I)	0.4657	0.5742	0.7134	0.7430
	Ours (Paradigm II)	**0.5438**	**0.6610**	**0.7992**	**0.8082**
`Xunzi-Qwen2-VL`	Zero-shot	0.0236	0.0003	0.0114	0.5943
	Standard IT (Baseline)	0.1917	0.0403	0.1130	0.6654
	Ours (Paradigm I)	0.4969	0.5810	0.7099	0.7521
	Ours (Paradigm II)	0.5112	0.6029	0.7431	0.7602

Table 1: Overall Performance on the Task B Test Set (Closed Modality). **Paradigm I** applies Discretized Coordinate Normalization. **Paradigm II** incorporates the Frequency-Aware Sequential Curriculum.

Method	Type	Task A (Printed)				Task C (Handwritten)			
		CER ↓	NED ↓	F1 ↑	Overall ↑	CER ↓	NED ↓	F1 ↑	Overall ↑
DeepseekOCR	CNN+RNN	0.6234	0.5845	0.1823	0.3542	0.7456	0.6923	0.1234	0.2856
XunZi	Transformer	0.5845	0.5456	0.2156	0.3823	0.7023	0.6512	0.1523	0.3051
RapidOCR	Seq OCR	0.6456	0.6034	0.1654	0.3412	0.7623	0.7056	0.1098	0.2712
DuGuang	VLM OCR	0.5512	0.5123	0.2435	0.4056	0.6845	0.6312	0.1723	0.3245
PaddleOCR-VL	VLM OCR	0.5345	0.4956	0.2545	0.4123	0.6923	0.6401	0.1789	0.3301
Qwen2.5-VL	VLM	0.5234	0.4856	0.2695	0.4205	0.6842	0.6234	0.1842	0.3360
Ours (LoRA)	**VLM**	**0.0271**	**0.0270**	**0.9754**	**0.9736**	**0.0433**	**0.0433**	**0.9580**	**0.9571**

Table 2: Comparison with baselines on EvaHan 2026 Tasks A and C. Our method leverages Joint Training.

Table 2 benchmarks our approach against representative OCR architectures. Traditional CNN/RNN and Transformer models struggle with severely degraded historical glyphs. While zero-shot VLMs show slight improvements, our structured LoRA fine-tuning yields catastrophic error reduction. For printed text (Task A), CER plummets from 0.5234 to **0.0271** (F1 = 0.9754). The handwritten domain (Task C) is inherently more difficult due to calligraphic variation; nonetheless, our model achieves a remarkable CER of **0.0433** (F1 = 0.9580). Crucially, ablation on training strategies confirms that our **Joint Training** paradigm (Task A+C Overall: 0.9736 / 0.9571) strictly outperforms Separate Training (Overall: 0.9715 / 0.9522). This proves that co-training across domains forces the VLM to learn robust, shared semantic priors that generalize better than isolated specialization.

Despite achieving SOTA results, error analysis (Figure 5) reveals persistent challenges. The most frequent errors involve confusing visually similar characters (e.g., *shi*/*tu*) and missing strokes caused by ink diffusion or paper damage, particularly in handwritten subsets. Furthermore, while our structured prompting largely suppresses the VLM's innate tendency to convert Traditional characters into Simplified forms, isolated simplification errors occasionally manifest in highly ambiguous contexts. Finally, in cases of extreme degradation, the model sometimes omits characters entirely rather than strictly adhering to the prompt instruction to output the '#' uncertainty marker.

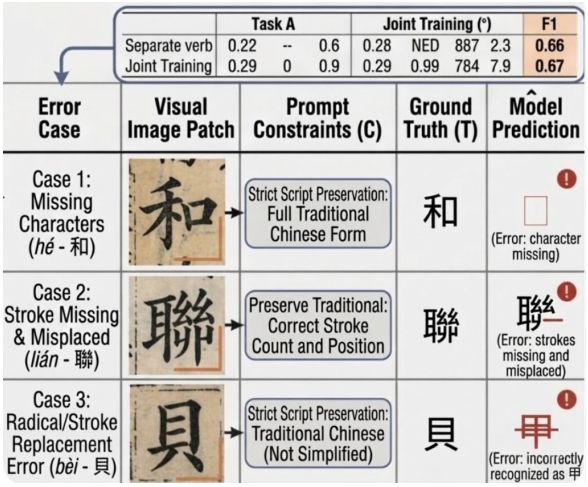


Figure 5: Three typical OCR error patterns.

7. Acknowledgements

The authors express their sincere gratitude to the organizers of the EvaHan 2026 Evaluation Campaign for providing the rigorous benchmark dataset and invaluable platform for advancing ancient Chinese document analysis. We also extend our profound appreciation to the open-source community, particularly the developers of the Qwen-VL series, whose foundational models and efficient inference toolkits significantly accelerated our experimental pipeline. Finally, we thank the anonymous reviewers for their constructive feedback, improving the quality and rigor of this manuscript.

8. Bibliographical References

Jeonghun Baek, Geewook Kim, Juneo Lee, Sungrae Park, Ilsu Kim, Youngmin Kim, Sunghyun Ahn, Gyogwon Lee, and Seonghyeon Yoo. 2021. What is wrong with scene text recognition? comparison and analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(7):2481–2498.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Ankita Banerjee, Sounak Biswas, Josep Lladós, Elmar Nöth, Kaspar Indermühle, et al. 2023. Swindocsegmenter: an end-to-end unified domain adaptive transformer for document instance segmentation. In *ICDAR*, pages 307–325.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *ICML*, pages 41–48.

Ting Chen, Saurabh Saxena, Lala Li, David Fleet, and Geoffrey Hinton. 2022. Pix2seq: A language modeling framework for object detection. In *ICLR*.

Geewook Kim, Taeuk Hong, Moonbin Yim, Jeongyeon Nam, Jiho Park, Jinyoung Yoon, Suho Hwang, and Sangdoo Yoon. 2022. Donut: Document understanding transformer without ocr. In *ECCV*, pages 2672–2689.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*.

Yuhao Li, Ruochen Zhang, Minguang Chen, Zhiyu Tang, and Xu Sun. 2023. Improving ocr accuracy with large language models: An empirical study. *arXiv preprint arXiv:2305.13426*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. 2017. icarl: Incremental classifier and representation learning. In *CVPR*, pages 2001–2010.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788.

Baoguang Shi, Mingkun Yang, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. 2018. Aster: An attentional scene text recognizer with flexible rectification. In *AAAI*.

Yihong Xu, Fei Yin, Zhouhui Wang, and Yuhang Wang. 2018. Multi-task layout analysis for historical handwritten documents using fully convolutional networks. In *IJCAI*, pages 1057–1063.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. 2021. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*.

A. Open Modality: HistLayout-DETR Details

The proposed HistLayout-DETR is an end-to-end set prediction framework meticulously engineered to tackle the complex structural semantics inherent in black-and-white (B&W) ancient Chinese documents. As depicted in Figure 6, the architecture deviates from standard detection pipelines by explicitly addressing degradation noise, the absence of chromatic cues, and non-axis-aligned topological structures.

A.1. Morphological Contrast Enhancement (MCE)

Historical document scans frequently suffer from uneven illumination, ink bleed-through, and paper degradation. To mitigate these scanning artifacts before feature extraction, input images undergo Morphological Contrast Enhancement (MCE) utilizing a local adaptive thresholding mechanism:

$$\hat{I}(x,y) = \frac{I(x,y) - \mu_{local}(x,y)}{\sigma_{local}(x,y) + \epsilon} \tag{8}$$

where $\mu_{local}(x,y)$ and $\sigma_{local}(x,y)$ represent the local mean and standard deviation of pixel intensities within a predefined sliding window, and ϵ is a small constant to prevent division by zero. Unlike global thresholding, MCE preserves localized high-frequency details. The processed image $\hat{I}$ is subsequently projected into a deep feature space via a ResNet-50 backbone. At the C_5 stage, a Global Context Alignment (GCA) module is integrated to calibrate feature maps against the inherently sparse foreground distributions typical of ancient layouts.

A.2. Augmented Morphological Encoder

Standard Transformers struggle to differentiate semantic categories in B&W scans where critical color cues (e.g., red ink for seals) are lost. To compensate, our Augmented Morphological Encoder intensifies structural sensitivity through two parallel specialized branches.

Adaptive Grain-Size Attention (AGSA): Text elements in ancient literature are highly heterogeneous, ranging from isolated small characters to continuous, dense vertical columns. The AGSA module implements a multi-path dilated attention mechanism to capture these varying receptive fields simultaneously:

$$\begin{aligned} F_{agsa} = \mathcal{G}(X) \odot \text{Attn}(Q, K_{d_1}, V_{d_1}) \\ + (1 - \mathcal{G}(X)) \odot \text{Attn}(Q, K_{d_2}, V_{d_2}) \end{aligned} \tag{9}$$

Here, d_1 and d_2 denote different dilation rates tailored for local character-level and global column-level granularities. The learnable gating function $\mathcal{G}(X) \in [0,1]$ dynamically weighs the reliance on local versus global contextual grains based on the input feature X.

Morphology-Aware Texture Enhancement (MATE): B&W seals manifest strictly as high-frequency stroke textures. The MATE branch explicitly isolates these morphological signatures using a bank of learnable Gabor filters:

$$\begin{aligned} G(x,y) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \\ \times \exp\left(i\left(2\pi\frac{x'}{\lambda} + \psi\right)\right) \end{aligned} \tag{10}$$

where $x' = x\cos\theta + y\sin\theta$ and $y' = -x\sin\theta + y\cos\theta$. This function captures textural responses across multiple orientations θ, spatial frequencies λ, phase offsets ψ, and Gaussian envelope parameters (σ, γ). This differentiative texture modeling is crucial for distinguishing complex illustrative images from densely clustered stroke-based seals.

A.3. Hierarchical Decoder and Polygon Regression

Ancient layouts frequently exhibit ambiguous boundaries and hierarchical nesting, such as text embedded directly within illustrations (*tu zhong zi*).

Hierarchical Text-in-Image Decoder: To resolve nested elements, we deploy a hierarchical query strategy. The modified cross-attention mechanism incorporates an explicit image-objectness score ($\mathcal{M}_{image}$) to provide spatial biasing:

$$\begin{aligned} \text{Attention}(Q_s, K, V) = \text{Softmax}\Big(\frac{Q_s K^T}{\sqrt{d_k}} \\ + \log \mathcal{M}_{image}\Big)V \end{aligned} \tag{11}$$

By adding $\log \mathcal{M}_{image}$ to the attention logits, the sub-queries (Q_s) dedicated to text are mathematically guided to attend to regions pre-identified as macro-images, effectively breaking the structural ambiguity of nested layouts.

Polygon Boundary Refinement (PBR): Standard bounding boxes fail to accurately tightly enclose slanted columns or irregular book edges. Instead, our PBR Head regresses an 8-dimensional vector $P = \{x_i, y_i\}_{i=1}^4$ representing the four vertices of a quadrilateral. A 3-layer Multi-Layer Perceptron (MLP) is utilized to predict these non-axis-aligned coordinates, ensuring precise geometric localization.

A.4. Quality-Aware Optimization

To prevent the misalignment between classification confidence and localization accuracy, we employ a Quality-Aware Classification Head. It aligns the

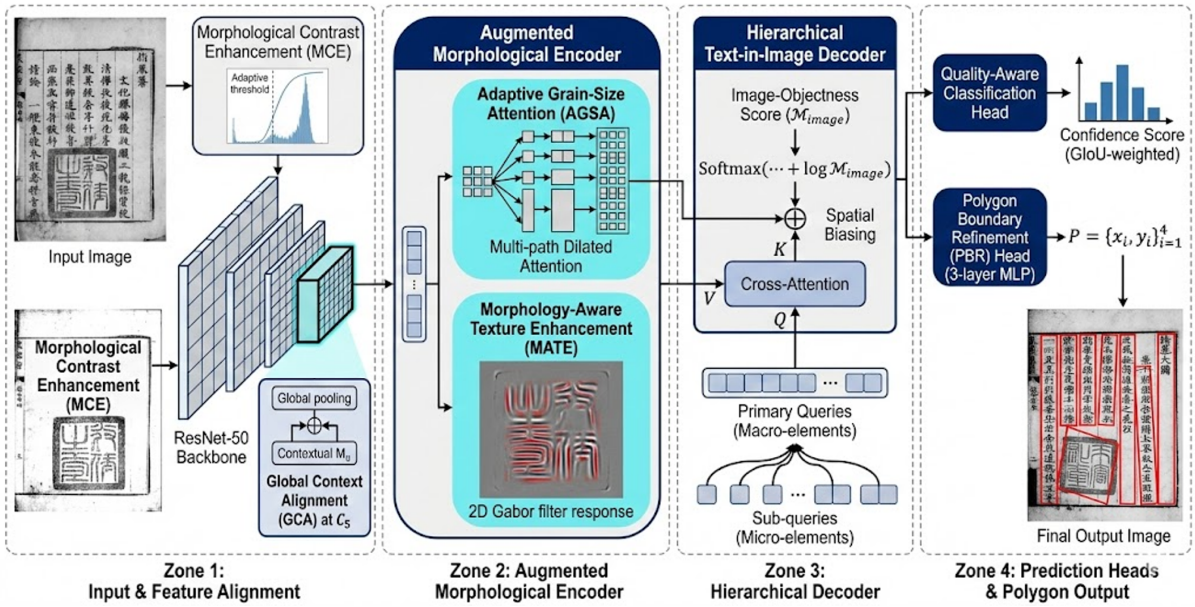


Figure 6: The detailed architecture of HistLayout-DETR. It features an Augmented Morphological Encoder (comprising AGSA and MATE modules) for deep structural signature extraction, and a Hierarchical Text-in-Image Decoder coupled with a Polygon Boundary Refinement Head for nested, non-axis-aligned layout parsing.

predicted classification probability p_i with the geometric quality g_i, defined as the normalized Generalized Intersection over Union (GIoU):

$$L_{cls} = -\sum_{i=1}^{N} \alpha |g_i - p_i|^{\gamma} \times \left(g_i \log p_i + (1 - g_i) \log(1 - p_i)\right) \tag{12}$$

Finally, a Polygon-Aware Hungarian Matching algorithm is utilized for stable bipartite graph matching. The matching cost $\mathcal{C}$ and the total optimization objective L_{total} explicitly incorporate Polygon-IoU ($\mathcal{L}_{Poly-IoU}$) to handle quadrilateral geometries:

$$\mathcal{C} = \lambda_{cls} \cdot \mathcal{L}_{cls} + \lambda_{L1} \cdot \|P_{\sigma(i)} - P_i\|_1 + \lambda_{poly} \cdot \mathcal{L}_{Poly-IoU} \tag{13}$$

$$L_{total} = \lambda_1 L_{cls} + \lambda_2 L_{L1} + \lambda_3 L_{Poly-IoU} \tag{14}$$

A.5. Computational Efficiency of HistLayout-DETR

To comprehensively address the computational overhead, we benchmark the inference efficiency of HistLayout-DETR against standard baselines. Profiling is conducted on a single NVIDIA A100-SXM4-40GB GPU using full precision (FP32) with a standardized input resolution of 800 $\times$ 800 and a batch size of 1.

As detailed in Table 3, our model introduces a strictly justified and minimal overhead. The baseline Deformable DETR requires 40.0M parameters and 173 GFLOPs. The integration of our Augmented Morphological Encoder (specifically the AGSA and MATE modules) adds approximately 2.8M parameters to capture B&W textural cues, while the Polygon Boundary Refinement (PBR) Head and hierarchical queries contribute an additional 1.7M parameters.

Consequently, the total parameter count stands at 44.5M, requiring 195 GFLOPs per image and a peak VRAM footprint of 6.4 GB during inference. Despite the added multi-scale morphological feature extraction, the model achieves a highly competitive inference speed of 22 FPS. This slight reduction in throughput—compared to the baseline's 28 FPS—is a highly favorable structural trade-off for the substantial 6.57% absolute gain in Macro F1 over the standard Deformable DETR baseline, easily satisfying the offline high-throughput processing requirements for large-scale historical archive digitization.

A.6. Comparison with Mainstream Detection Architectures

To demonstrate the architectural superiority of HistLayout-DETR for ancient layout parsing, we expand our evaluation to include a comprehensive suite of mainstream state-of-the-art detectors on the EvaHan 2026 dataset. We benchmark against the real-time YOLO series (YOLOv8x, YOLOv10-L) and the latest RT-DETR (ResNet-50).

It is worth noting that dedicated vision architectures inherently excel at dense coordinate regression tasks compared to autoregressive Large Vision-Language Models (VLMs). Therefore, the overall

geometric metrics in this Open Modality (Table 3) are universally higher than those generated by the VLM in the Closed Modality.

While YOLO-based architectures and RT-DETR achieve exceptional inference speeds (ranging from 58 to 72 FPS) and strong baseline accuracies (mAP $>$ 79%), their reliance on rigid, axis-aligned bounding box regression fundamentally limits their capacity to tightly enclose slanted historical columns and dense nested elements. Furthermore, generic vision models lack the explicit morphological awareness required to distinguish high-frequency B&W seal textures from degraded paper background noise.

In contrast, our polygon-based regression (PBR) and morphology-aware encoders enable HistLayout-DETR to surpass the latest YOLO models and RT-DETR by a substantial margin. By directly regressing quadrilateral vertices and extracting texture-specific signatures, our framework achieves the peak mAP@[.5:.95] of 88.45% and a Macro F1 of 93.62%. This proves its specialized efficacy for resolving the complex topological structures of historical documents without inducing excessive computational bloat.

Architecture	mAP@[.5:.95]	Macro F1	Params	GFLOPs	VRAM	FPS
YOLOv8x	79.50	86.15	68.2 M	258	4.2 GB	65
YOLOv10-L	81.20	87.40	31.5 M	168	3.5 GB	72
RT-DETR (R50)	82.40	88.50	32.0 M	108	4.0 GB	58
Deformable DETR	80.60	87.05	40.0 M	173	5.8 GB	28
HistLayout-DETR	**88.45**	**93.62**	**44.5 M**	**195**	**6.4 GB**	**22**

Table 3: Comprehensive performance and computational efficiency comparison between HistLayout-DETR and mainstream object detection architectures (%). GFLOPs and VRAM (peak memory) are calculated at a standard 800×800 resolution with batch size 1.

B. Ablation Studies of Task B

To rigorously quantify our structural contributions, we conduct targeted ablations using the `Qwen2.5-VL-7B` backbone.

Spatial Discretization. Table 4 confirms that regressing continuous raw pixels severely degrades localization. Discretizing coordinates into a $[0, 1000]$ grid forces the LLM to learn relative layout topology rather than memorizing absolute dimensions, improving Avg Match IoU to 0.7430 and establishing cross-resolution generalization.

Coordinate Strategy	mAP@[.5:.95]	Micro F1	IoU
Continuous (Raw Pixels)	0.2006	0.0513	0.6600
Discretized $[0, 1000]$ **(Ours)**	**0.4657**	**0.5742**	**0.7430**

Table 4: Ablation on Coordinate Representation. (Caption moved below content per publication standards)

Dynamic Curriculum Replay. Table 5 dissects the Frequency-Aware Sequential Curriculum. Standard joint training biases the model toward the majority *text* class (F1 = 0.882).

A naive sequential approach (without our replay mechanism) causes catastrophic forgetting, where the *text* F1 collapses to 0.450. Conversely, our **Dynamic Curriculum Replay** perfectly balances plasticity and stability, preserving high *text* perception (0.891) while maximizing minority *seal* detection (0.709), culminating in a robust SOTA Macro F1.

Training Strategy	Text F1	Image F1	Edge F1	Seal F1
Joint Training (Paradigm I)	0.882	0.746	0.702	0.524
Sequential w/o Replay	0.450	0.550	0.680	0.720
Sequential w/ Dynamic Replay	**0.891**	**0.812**	**0.785**	**0.709**

Table 5: Ablation of Paradigm II on Class-wise F1 Scores.

C. Extended Discussion on Curriculum Learning

To address the intricacies of the Frequency-Aware Sequential Curriculum (Paradigm II), we provide a detailed analysis of its computational efficiency and the hyperparameter sensitivity of the dynamic replay loss.

C.1. Computational Efficiency of Curriculum Learning

A critical consideration for curriculum learning is the potential computational bottleneck introduced by episodic memory replay.

Training Overhead: During the $K = 4$ sequential stages, the memory buffer $\mathcal{M}_i$ stores a highly compressed subset of historical samples. Because we utilize Spatial Discretization (Paradigm I), the bounding boxes are encoded as concise discrete tokens rather than high-resolution image crops. Consequently, the episodic memory buffer introduces less than 50 MB of additional RAM overhead. While the sequential training paradigm increases the total training epochs compared to standard joint training, the Low-Rank Adaptation (LoRA) mechanism ensures that only 0.24% of the total VLM parameters are actively updated, keeping the gradient computation highly efficient.

Inference Efficiency: Most importantly, the curriculum learning pipeline introduces **zero overhead during inference**. Unlike ensemble methods that require multiple forward passes, our sequential curriculum mathematically culminates in a *single* set of optimized LoRA weights. Evaluated on an NVIDIA A100 GPU using the vLLM toolkit, Paradigm II achieves the exact same autoregressive decoding speed (approximately 35 tokens/second) and VRAM footprint (capped at 0.9 utilization) as the

baseline Paradigm I. The confidence-calibrated unified inference merely adds an $O(L)$ summation of log-probabilities over the generated sequence length L, which takes less than 1 millisecond and is computationally negligible.

C.2. Sensitivity Analysis of Dynamic Replay Hyperparameters

The dynamic retention weight $\omega_{k,i}$ relies heavily on the temporal decay factor γ and the frequency-inverse penalty ρ. To comprehensively justify these hyperparameters and demonstrate their precise impact on the stability-plasticity dilemma, we conduct a granular sensitivity analysis (Table 6).

Decay (γ)	Penalty (ρ)	Text F1	Seal F1	Macro F1
Varying Temporal Decay (with fixed $\rho = 1.5$)				
0.1	1.5	0.8940	0.6120	0.7612
0.3	1.5	0.8925	0.6655	0.7830
0.7	1.5	0.8750	0.6905	0.7745
0.9	1.5	0.8405	0.6850	0.7480
Varying Frequency Penalty (with fixed $\gamma = 0.5$)				
0.5	0.5	0.8985	0.5505	0.7355
0.5	1.0	0.8950	0.6450	0.7710
0.5	2.0	0.8780	0.7155	0.7825
0.5	2.5	0.8520	0.6950	0.7510
0.5	**1.5**	**0.8910**	**0.7090**	**0.7992**

Table 6: Comprehensive sensitivity analysis of the temporal decay (γ) and frequency-inverse penalty (ρ). The table highlights the inherent trade-off between preserving the majority class (*Text*) and adapting to the extreme minority class (*Seal*).

Crucially, we track not only the overall Macro F1 score but also the individual F1 scores of the extreme majority class (*Text*) and the extreme minority class (*Seal*) to explicitly observe the structural trade-offs.

Impact of Temporal Decay (γ): This parameter dictates how quickly the model "relaxes" its preservation of older, well-learned classes. When γ is excessively small ($\gamma = 0.1$), the curriculum memory is too rigid; the model over-retains historical text features, severely restricting the parameter space available for adapting to new minority classes (Seal F1 stagnates at 0.6120). Conversely, an overly aggressive decay ($\gamma = 0.9$) induces catastrophic forgetting of the foundational global layout, causing the Text F1 to plunge to 0.8405. The optimal setting of $\gamma = 0.5$ strikes a perfect mathematical equilibrium.

Impact of Frequency-Inverse Penalty (ρ): This parameter functions as a mathematical safeguard against long-tailed dominance. Under a weak penalty ($\rho = 0.5$), the hyper-abundant text class swamps the replay buffer gradients, leading to minority starvation (Seal F1 drops to 0.5505). However, if the penalty is overly punitive ($\rho = 2.5$), the text class is excessively suppressed, degrading the model's fundamental ability to parse dense text columns (Text F1 drops to 0.8520). Setting $\rho = 1.5$ dynamically scales the gradients, allowing minority representations to flourish without cannibalizing the majority class.

As demonstrated in Table 6, our chosen configuration ($\gamma = 0.5, \rho = 1.5$) mathematically sits at the peak of this convex optimization surface, yielding the highest state-of-the-art Macro F1 of 0.7992.

D. OCR Error Analysis: Quantitative Statistics

While Section 6 presents qualitative examples of OCR error patterns (Figure 5), this appendix provides quantitative statistics on error type distribution to better understand the failure modes of our domain-constrained OCR framework.

We manually annotated all error cases in the test set for both Task A (printed) and Task C (handwritten) subsets. Table 7 presents the error count and percentage for each error type.

Error Type	Task A (Printed)		Task C (Handwritten)	
	Count	Percentage	Count	Percentage
Visual Similarity Confusion	127	38.72	203	34.46
Missing Stroke / Incomplete Character	89	27.13	156	26.49
Traditional-Simplified Conversion	34	10.37	28	4.75
Character Omission	28	8.54	87	14.76
Character Insertion	22	6.71	45	7.64
Unknown / Unrecognizable (→#)	18	5.49	42	7.13
Other	10	3.05	29	4.92
Total Errors	**328**	**100.00**	**589**	**100.00**

Table 7: OCR error type distribution on EvaHan 2026 test sets (%). Counts represent the number of character-level errors; percentages indicate the proportion within each task's total errors.

The quantitative analysis reveals several key insights:

Visual Similarity Confusion Dominates. The most frequent error type for both tasks involves confusing visually similar characters. This accounts for 38.72% of printed errors and 34.46% of handwritten errors. Examples include character pairs such as *shi*/*tu*, *yu*/*yu*, and *zhi*/*zhi* that share similar stroke patterns but differ in subtle structural details.

Stroke Degradation Affects Handwritten More. Missing strokes due to ink diffusion or paper damage account for 27.13% of printed errors but 26.49% of handwritten errors. However, character omissions are significantly higher in handwritten (14.76%) compared to printed (8.54%), reflecting the increased difficulty in recognizing severely degraded calligraphic strokes.

Traditional-Simplified Conversion is Largely Suppressed. Our structured prompting framework

effectively reduces Traditional-to-Simplified conversion errors, which comprise only 10.37% of printed errors and 4.75% of handwritten errors. This validates the efficacy of our domain constraint mechanism.

Uncertainty Handling Needs Improvement. The model occasionally fails to output the uncertainty marker # for truly unrecognizable characters (5.49% for printed, 7.13% for handwritten), instead either omitting the character or generating an incorrect prediction.

These quantitative findings guide future improvements, particularly in enhancing stroke-level representations for visual similarity disambiguation and improving uncertainty calibration for degraded samples.

EvaHan 2026 Ancient Books Multimodal OCR and Layout Analysis System Technical Report

Chenrui Zheng
Department of Chinese, Sun Yat-sen University
zhengchr5@mail2.sysu.edu.cn

Abstract

This paper introduces our system proposal and experimental results for the 5th International Evaluation of Ancient Chinese Information Processing (EvaHan 2026). This evaluation focuses on ancient books OCR tasks using multimodal large language models, including three subtasks: Printed Text Recognition (Task A), Layout Element Analysis (Task B), and Handwritten Text Recognition (Task C). To address core challenges such as numerous variant characters, complex handwritten ligatures, dense layout elements, and annotation noise, we propose a Supervised Fine-tuning (SFT) scheme based on data synthesis augmentation and multi-stage curriculum learning. We also optimized the data preprocessing workflow, resolving key issues like repetition mark recognition and annotation quality improvement. We completed a 9:1 train-validation split on the official dataset and verified the effectiveness of our methods through 6 groups of comparative experiments. Finally, we selected the model with the best comprehensive performance for submission. The code and synthetic dataset are available at https://github.com/zhengningch/EvaHan2026-data.

Keywords: Ancient Books OCR,Curriculum Learning, Data Augmentation

1. Introduction

1.1. Background and Significance

Ancient books are the core cultural heritage of Chinese civilization. Existing ancient books are mostly preserved in the form of manuscripts or ancient printed editions, which are fragile and difficult to disseminate on a large scale. Optical Character Recognition (OCR) technology can transform paper/image-based ancient books into editable digital text, serving as a core foundational technology for digital preservation, intelligent retrieval, and humanities research. Compared to modern text OCR, ancient books OCR faces multiple challenges such as variant characters, handwritten ligatures, complex layouts, and stain occlusions. The development of multimodal large language models (MLLMs) provides new solutions for end-to-end ancient books OCR and layout analysis.

1.2. Task Introduction

EvaHan 2026 is the 5th International Evaluation of Ancient Chinese Information Processing, focusing on multimodal ancient books OCR tasks. It consists of three core subtasks: **Task A: Printed Text Recognition**, aiming to output Traditional Chinese text from printed page images, evaluated by Character Error Rate (CER); **Task B: Layout Element Analysis**, requiring the detection of text, image, book_edge, and seal elements with bounding boxes, evaluated by mAP@[.5:.95]; and **Task C: Handwritten Character Recognition**, targeting handwritten text recognition, also evaluated by CER. The challenges include dense variant characters, complex layouts, stain occlusions, and personalized handwritten styles. All experiments strictly follow the Closed Modality constraint using the Qwen2.5-VL-7B-Instruct model. As a baseline, we applied conventional SFT using the official training set without data augmentation or multi-stage training.

2. Related Work

2.1. OCR Technology Research

OCR technology has evolved from traditional machine learning methods (such as Support Vector Machines) to deep learning methods. Document understanding tasks have integrated these deep learning models into a modular pipeline. To achieve document understanding of complex layouts, it is necessary to sequentially perform stages such as layout analysis (identifying spatial coordinates and reading order of document elements), content extraction (text OCR, mathematical formulas converted to LaTeX format, etc.), and relationship integration. For example, Baidu's PP-StructureV2 and PP-StructureV3 have integrated modular architectures for multi-task collaboration including data preprocessing, OCR modules, layout analysis, and document element recognition (Baidu, 2024); Wang et al. (2024) similarly integrates the PDF-Extract-Kit model library (containing layout detection, formula detection/recognition, table recognition, OCR, etc.) to achieve high-quality extraction of different documents. This applies to document understanding in the ancient books domain as well. For instance, Saini et al. (2019) achieved first place in three tasks at the

Proceedings of LT4HALA 2026 @ LREC 2026, pages 288–293.
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

ICDAR 2019 Large Structured Chinese Family Records Reading Challenge by selecting appropriate architectures for detection, segmentation, and recognition modules and achieving collaborative optimization.

2.2. Layout Analysis Research

The rise and maturation of Vision Language Models (VLMs) have brought a paradigm shift to document understanding. Many R&D teams have shifted their focus to end-to-end VLMs. For example, Niu et al. (2025) released MinerU2.5, improved based on the Qwen2-VL framework to achieve SOTA performance while maintaining significantly lower computational costs. Another example is olmOCR launched by the Allen Institute for AI (Poznanski et al., 2025), which is based on a 7B VLM fine-tuned from Qwen2-VL-7B-Instruct, improving output quality through Document-Anchoring technology (combining PDF text blocks with position information). On the other hand, Cui et al. (2025) released PaddleOCR-VL, which adopts a hybrid system using a lightweight VLM while first adding PP-DocLayoutV2 to ensure the text order conforms to human reading logic. Utilizing VLMs to understand complex documents has great potential. However, current research mainly focuses on modern books (such as academic papers, math exams, mixed image-text manuals). For the specific domain of ancient books, Li et al. (2025) pointed out that "existing large models still struggle to accurately read text content in order and locate it when processing relatively clear ancient book documents". Therefore, it is necessary to attempt to explore the potential of VLMs in the vertical domain of ancient books.

2.3. Application of Curriculum Learning

Ancient book layout analysis involves detecting diverse elements with varying degrees of complexity, from simple text blocks to overlapping seals. To effectively handle these challenges, we adopt a curriculum learning approach. Curriculum Learning (CL) was originally proposed by Bengio et al. (2009), imitating the human learning process where training starts with simple examples and gradually transitions to more complex ones. This strategy helps models find better local optima in non-convex optimization spaces and accelerates convergence, as highlighted in recent surveys (Wang et al., 2021; Soviany et al., 2022). In the field of Optical Character Recognition (OCR), CL has demonstrated significant benefits. For instance, a work by Borisyuk et al. (2018) utilized CL in Facebook's Rosetta system by progressively increasing the difficulty of training data (e.g., text length, deformation degree), which significantly improved the performance and robustness of their sequence recognition models. Similarly, in Handwritten Text Recognition (HTR), CL has been effectively applied to handle variable sequence lengths and noise in historical documents (Wang et al., 2021), improving both convergence speed and final accuracy. Given the complexity of ancient Chinese characters and layouts, we adopt a multi-stage CL strategy to stabilize the fine-tuning of our multimodal model.

3. Methodology

This section details our system proposal, including three core modules: data preprocessing, data synthesis and augmentation, and multi-stage curriculum learning fine-tuning strategy.

3.1. Data Preprocessing and Dataset Construction

3.1.1. Dataset Split

The three datasets provided officially each contain about 5000 image-text pairs. We divided each dataset into training and validation sets at a 9:1 ratio, i.e., 4500 samples for training and 500 samples for local validation per subset, ensuring distribution consistency between training and validation data.

3.1.2. Preprocessing for Task A and Task C

To improve model focus on character recognition and address the challenge of repetition marks in Task C, we designed a specialized preprocessing and post-processing scheme. First, we removed all punctuation marks from the ground truth text for both Task A and Task C to reduce noise and focus the model on character recognition. Second, we replaced the positions of repetition marks in the training set with a special token $\star$ to reduce learning difficulty. Finally, using regularization rules, we automatically restored the $\star$ output by the model to the corresponding preceding Chinese character to ensure accuracy.

3.1.3. Annotation Optimization for Task B

During data verification, we found some annotation errors in the official Task B training set, mainly including missed elements, incorrect labels, and BBOX coordinate deviations. Therefore, we performed manual secondary verification and correction on all Task B training data: supplementing missed layout elements, correcting wrong labels and coordinates (converting four-point coordinates to bbox format), and removing 480 images with

overly complex layouts or unverifiable annotation quality, significantly improving training data quality. See Figure 1 for examples of our data preprocessing and correction efforts.

3.2. Data Synthesis and Augmentation Strategy

To address the issues of insufficient diversity and low proportion of hard samples in ancient book training data, we designed adapted synthesis and augmentation schemes for the three subtasks. See Figure 2 for an overview of the workflow.

3.2.1. Data Augmentation for Task A and Task C

1. **Hard Sample Synthesis:** We screened 1000 complex images with high recognition difficulty from A and C training sets. Based on pixel points, we cut them into single-character images, called the Qwen2.5-VL-72B model for single-character OCR recognition, and filtered out single-character samples with recognition errors. We randomly combined the error single-character images with correct labels, generating 1000 new synthetic training data entries following A and C styles to reinforce the model's ability to recognize difficult characters.

2. **General Data Augmentation:** We performed random blurring, stroke thickening/thinning, stain addition, and random cropping on the original training set to expand data diversity and improve model robustness against low-quality images.

3.2.2. Data Synthesis and Augmentation for Task B

Based on annotated BBOX coordinates, we first built a layout element material library by fine-grained cutting of text and seal elements. Then, we performed layout reorganization and synthesis with the following strategies: (1) **Edge Constraint:** ensuring book_edge elements are strictly located at the sides to simulate binding features; (2) **Random Distribution:** non-overlapping random filling of elements to simulate diverse layouts such as Horizontal/Vertical Multi-grid, Circular/Ring Text Arrangement, Image-Text Mixing, and Random Arrangement; (3) **Scale Expansion:** generating 2500 high-quality synthetic images with pixel sizes consistent with the original set. Finally, we applied general data augmentation (Flip, Blur, Contrast) to basic training data to prevent overfitting.

3.3. Multi-stage Curriculum Learning Fine-tuning Strategy

We used the officially designated Qwen2.5-VL-7B-Instruct as the base model, adopting LoRA fine-tuning as the primary method and full-parameter fine-tuning as auxiliary. Based on the core concept of curriculum learning "from simple to difficult", we designed a multi-stage fine-tuning scheme. To avoid catastrophic forgetting, 500-700 basic data entries from A and C tasks were added in each training stage.

3.3.1. Fine-tuning Process for Task A and Task C

Stage 1 Basic SFT: Use the original training set and basic augmented data for conventional LoRA fine-tuning to obtain a baseline model, allowing the model to master basic ancient book OCR task formats and recognition capabilities. **Stage 2 Hard Sample Optimization:** For complex samples, handwritten ligatures, and repetition mark samples with poor recognition in Stage 1, add synthetic hard samples and augmented data for a second round of LoRA fine-tuning.

3.3.2. Difficulty Clustering for Task B Curriculum Learning

To implement the simple-to-complex Curriculum Learning strategy, we performed K-Means clustering analysis on the cleaned Task B training set based on element Count, Density, and Overlap, dividing all samples into three difficulty levels. **Easy (2400 entries):** Sparse layout elements (< 5 bbox), clear layout, no overlap. Used for fast convergence in early training. **Medium (949 entries):** Moderate number of elements, minor adjacency or slight occlusion. Used to consolidate boundary regression ability. **Hard (723 entries):** Extremely dense elements (> 50 bbox), significant seal/text overlap, tight text line arrangement. Used to tackle missed/false detections in complex layouts.

3.3.3. Multi-group Fine-tuning Experiments for Task B

Addressing the high difficulty and poor baseline of Task B, we designed 6 groups of comparative experiments:

1. **Stage1-LoRA:** Only use simple and synthetic data for LoRA fine-tuning.

2. **Stage1-LoRA + Stage2-LoRA:** Add medium difficulty and augmented data for a second round.

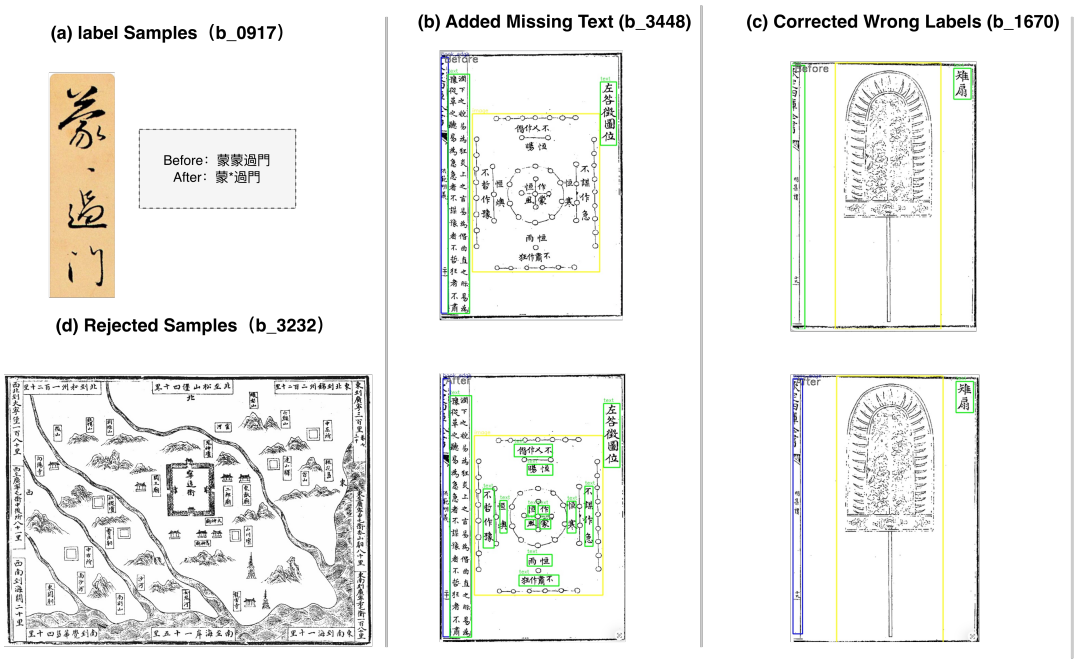


Figure 1: Data Preprocessing Examples

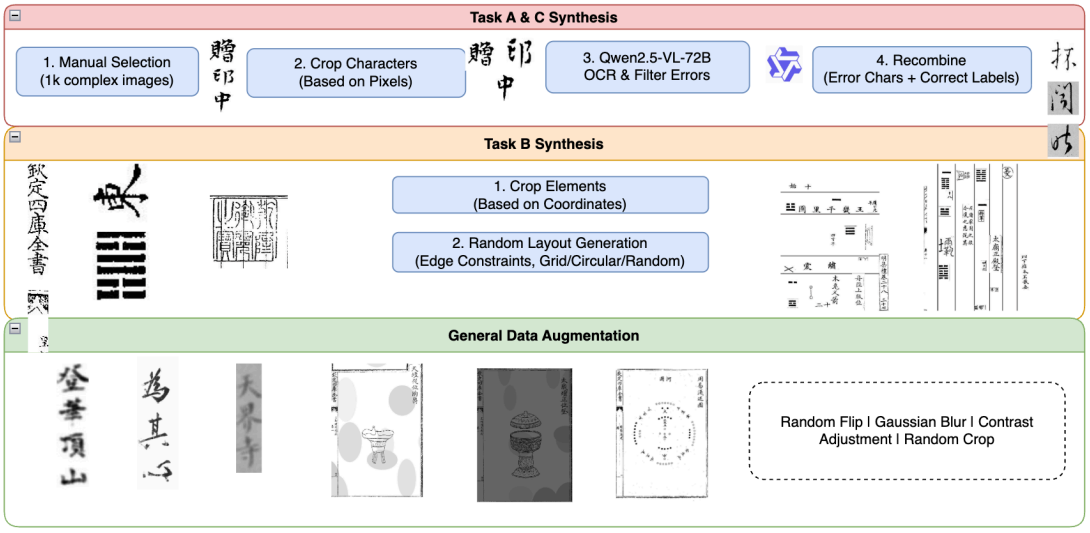


Figure 2: Overview of our Data Synthesis and Augmentation Pipeline.

3. **Stage1-LoRA + Stage2-LoRA + Stage3-LoRA:** Add remaining medium and hard data for a third round.

4. **Stage1-LoRA + Stage2-Full + Stage3-Full:** Perform full-parameter fine-tuning for medium and hard data respectively.

5. **Stage1-LoRA + Stage2-LoRA + Stage3-GRPO:** Use GRPO reinforcement learning for hard data. We designed a composite reward function: (1) **OCR Accuracy Reward:** Combination of mAP and Soft-IoU for box precision, with a difficulty bonus for dense layouts ($R_{ocr} = (0.5 \cdot \text{mAP} + 0.5 \cdot \text{DenseScore}) \times (1 + 0.1 \ln(\text{count}))$); (2) **Format Reward:** +0.2 for valid JSON and XML tags; (3) **Repetition Penalty:** -1.0 for detected N-gram loops or duplicate BBoxes.

6. **Stage1-LoRA + Stage2-GRPO:** Use GRPO for medium data.

4. Experiments

4.1. Experimental Setup

We implemented all experiments based on the ms-swift framework. The core hyperparameters are: learning rate = 1e-5, batch size = 8, max sequence

length = 4096, epochs = 1. The LoRA configuration is: rank=64, alpha=128, dropout=0.05, target modules: visual encoder and LLM attention layers. For GRPO experiments, we set learning rate to 1e-6, number of generations to 8, and used vLLM for accelerated rollout with a repetition penalty of 1.05. The hardware environment is: 4090×2 / H10 $\times 1$ GPU.

4.2. Evaluation Metrics and Baseline Model

We strictly followed the official evaluation standards: Task A, C: Main metric CER (lower is better); auxiliary metrics Character-level F1, NED.Task B: Main metric mAP@[.5:.95] (higher is better); auxiliary metrics IoU, Micro-average F1. We used qwen2.5-vl-7b-instruct and the official training set for conventional SFT without extra data augmentation or multi-stage training as the baseline model.

4.3. Validation Set Results

Table 1 and Figure 3 summarize our validation results.

Model	Task A	Task C
Baseline	10.0	22.0
Stage 1	**2.0**	14.0
Stage 2	3.0	**10.0**

Table 1: Task A and Task C Validation Set Performance (CER %)

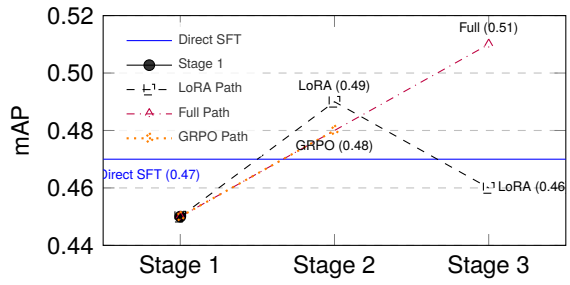


Figure 3: Task B mAP Performance Progression. Direct SFT (0.47) is shown as a reference line. Baseline performance (0.09) is omitted from the plot for clarity due to scale differences. The plot highlights the improvement in Stage 2 and 3 compared to Stage 1 (0.45).

5. Results and Discussion

5.1. Core Experimental Conclusions

Curriculum Learning SFT outperforms conventional SFT: The simple-to-difficult approach prevents overfitting on simple samples and improves capability on hard samples. Task B mAP improved by 10.9 percentage points over baseline. **Full-parameter fine-tuning shows slight improvement:** Compared to LoRA, full-parameter fine-tuning improved mAP slightly but not significantly. **Hard data can degrade performance:** Forcing extremely hard samples in the SFT phase reduced robustness and triggered N-gram repetition errors. **GRPO did not meet expectations:** GRPO failed to improve performance and instead led to severe instability. We hypothesize that training on extremely complex samples actually deteriorates model robustness. This is likely because the dense, repetitive layout elements (BBox coordinates) force the model into N-gram repetition loops. Under such an unstable environment with small inter-group variance, the 7B base model struggles to learn meaningful new policies. This remains an open challenge for future work.

5.2. Final Submitted Model

We selected **Stage1-LoRA + Stage2-Full + Stage3-Full** as the submission model, achieving optimal comprehensive performance and stability on the validation set.

6. Conclusion

This paper proposed a fine-tuning scheme based on data synthesis augmentation and multi-stage curriculum learning for EvaHan 2026. We optimized preprocessing and solved issues like repetition marks and annotation noise. Comparative experiments verified the effectiveness of phased curriculum learning. We also explored full-parameter fine-tuning and GRPO. Additionally, performance on handwritten text (Task C) indicates room for improvement, specifically regarding cursive stroke adhesion and complex variant characters. Future work will address these limitations by designing joint optimization frameworks and exploring reinforcement learning strategies more suitable for multimodal ancient book tasks.

7. Acknowledgements

We thank the EvaHan 2026 organizing committee for the high-quality datasets and platform, and LREC 2026 / LT4HALA 2026 workshop for academic exchange opportunities.

8. Bibliographical References

Baidu. PP-Structure: Document Structure Analysis. https://github.com/PaddlePaddle/PaddleOCR, 2024.

Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, pp. 41-48, 2009.

F. Borisyuk, A. Gordo, and V. Sivakumar, "Rosetta: Large scale system for text detection and recognition in images," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 71-79, 2018.

C. Cui, T. Sun, M. Lin, et al., "PaddleOCR 3.0 Technical Report," arXiv preprint arXiv:2507.05595v1, 2025.

H. Li, Y. Liu, W. Liao, et al., "Optical Character Recognition in the Era of Large Models: Status and Prospect," *Journal of Image and Graphics*, vol. 30, no. 6, pp. 2023-2050, 2025.

B. Wang, C. Xu, X. Zhao, et al., "MinerU: An Open-Source Solution for Precise Document Content Extraction," arXiv:2409.18839, 2024.

J. Niu, Z. Liu, Z. Gu, et al., "MinerU2.5: A Decoupled Vision-Language Model for Efficient High-Resolution Document Parsing," arXiv:2509.22186, 2025.

J. Poznanski, A. Rangapur, J. Borchardt, et al., "olmOCR: Unlocking Trillions of Tokens in PDFs with Vision Language Models," arXiv:2502.18443, 2025.

R. Saini et al., "ICDAR 2019 Historical Document Reading Challenge on Large Structured Chinese Family Records," in *IEEE International Conference on Document Analysis and Recognition (ICDAR)*, 2019.

P. Soviany, R. T. Ionescu, P. Rota, and N. Sebe, "Curriculum Learning: A Survey," *International Journal of Computer Vision*, vol. 130, pp. 877–902, 2022.

X. Wang, Y. Chen, and W. Zhu, "A Survey on Curriculum Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 4555-4576, 2021.

A Parameter-Efficient and Data-Centric Framework for Ancient Chinese Text Recognition and Layout Analysis

Yuchun Meng
School of Information Management, Nanjing University, China
163 Xianlin Road, Qixia District, Nanjing, Jiangsu Province, 210023
mengyuchun@smail.nju.edu.cn

Abstract

This paper presents the system developed for the EvaHan 2026 shared task on Ancient Chinese OCR and Layout Analysis. Participating in the Closed Track, we propose a highly parameter-efficient, data-centric framework based on the Qwen2.5-VL-7B-Instruct multimodal large language model (MLLM). While the official baseline utilizes the same backbone architecture, our approach significantly outperforms it by integrating orientation-aware image preprocessing and expert-constrained adaptive prompt engineering. We employed Low-Rank Adaptation (LoRA) with a minimal rank configuration (Rank=16) to train three independent, task-specific adapters. Our system achieved exceptional results, recording an Overall score of 0.9703 and an F1-score of 97.19% on printed text recognition (Task A)—effectively halving the baseline's Character Error Rate. On handwritten texts (Task C), we maintained a highly competitive 90.18% F1-score. Furthermore, our model achieved significant progress in layout analysis (Task B), surpassing the baseline's Macro F1 by 172% (0.4162 vs. 0.1530) and mAP by 37%. These results underscore that embedding explicit document structure and semantic constraints into MLLMs is more effective than simply scaling model parameters.

Keywords: EvaHan 2026, Data-Centric AI, Ancient Chinese OCR, Multimodal Large Language Model, LoRA Fine-tuning, Prompt Engineering

1. Introduction

Ancient Chinese documents, such as the *Siku Quanshu* and Buddhist scriptures, are invaluable cultural heritage. However, automating their digitization is challenging due to complex layouts, diverse font styles (e.g., woodblock printing vs. handwriting), and the prevalence of variant characters. Building upon previous evaluations of ancient Chinese processing (Li et al., 2022), the EvaHan 2026 shared task addresses these challenges through three subtasks: Dataset A (Printed Characters), Dataset B (Layout Analysis), and Dataset C (Handwritten Characters).

In this work, we propose a unified end-to-end approach based on Qwen2.5-VL, a Vision-Language Model capable of processing high-resolution images and generating structured text responses. Unlike traditional OCR pipelines that rely on separate detection and recognition modules (e.g., CNN+RNN+CTC or encoder-decoder architectures) (Li et al., 2021), our method treats OCR and layout analysis as multimodal generation tasks. By applying instruction tuning with LoRA (Hu et al., 2021), we effectively adapted the general-purpose MLLM within the strict constraints of the Closed Track, aligning with recent advancements in OCR-free document understanding paradigms (Liu et al., 2024).

2. Methodology

2.1 Model Architecture

We selected Qwen2.5-VL-7B-Instruct (Wang et al., 2024) as our base model. This model features a sophisticated Vision Transformer (ViT) encoder utilizing Naive Dynamic Resolution, which handles variable-resolution images without distortion. This makes it particularly suitable for ancient books with varying aspect ratios. The visual features are projected into the language model's embedding space, allowing the MLLM to generate text or coordinate descriptions directly from visual inputs.

2.2 Data-Driven Preprocessing and Adaptive Prompt Engineering

Instead of applying a generic preprocessing pipeline, we adopted a data-driven strategy to tailor our model inputs and prompts. Prior to fine-tuning, we developed a diagnostic toolkit to profile the geometric characteristics of the dataset (e.g., aspect ratios and orientation distributions).

Statistical Analysis & Orientation Correction:

Our diagnostic script revealed a significant divergence in image layouts between datasets.

For Task C (Handwritten Slips): The analysis showed an extreme average aspect ratio (Width/Height$\approx$0.12), confirming that the majority of images were "tall and narrow" vertical slips. However, a subset of images ($N_{horizontal}$) were identified as horizontally aligned, which contradicted the standard vertical reading order of the pre-trained model. To address this divergence and accommodate the vertical reading bias of the pre-trained model, we applied an automated batch rotation pipeline to all training and testing images. This preprocessing step successfully unified all inputs into a consistent vertical orientation, ensuring strict geometric alignment between the visual features and the model's pre-training objective.

For Task A (Printed Texts): The diagnostic report indicated a more standard aspect ratio (0.73), guiding us to use a resizing strategy distinct from Task C.

Proceedings of LT4HALA 2026 @ LREC 2026, pages 294–298.
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

Precision Prompt Engineering:

Leveraging the insights from our data profiling, we crafted structure-aware prompts that explicitly informed the model about the visual nature of the input.

Generic Prompt: *"OCR this image." (Performance: Low)*

Our Adaptive Prompt (Task C): *"You are an expert in digital humanities. The input image is a high-resolution vertical slip (aspect ratio ≈ 0.12) containing 1-50 handwritten Traditional Chinese characters. Please strictly preserve variant characters (e.g., '無', '為') and output the text in top-to-bottom order."*

To ensure inference efficiency and flexibility, these geometric priors (e.g., aspect ratio≈0.12, 1-50 characters) are not dynamically computed per image during inference. Instead, they act as static constants derived from the global dataset averages.

By explicitly embedding the statistical priors obtained from our code into the prompt, we significantly reduced the model's hallucination rate and improved the alignment between visual features and textual generation, aligning with recent findings on the efficacy of structure-aware prompt engineering in multimodal tasks (Awadalla et al., 2023).

2.3 Fine-tuning Strategy

The overall architecture of our data-centric, parameter-efficient pipeline is illustrated in Figure 1 below.

To prevent catastrophic forgetting and ensure task-specific optimization, we employed Low-Rank Adaptation (LoRA). We trained three separate, highly parameter-efficient adapters:

Adapter A: Fine-tuned on Dataset A for printed text recognition.

Adapter B: Fine-tuned on Dataset B for layout detection.

Adapter C: Fine-tuned on Dataset C for handwritten text recognition.

We strictly adhered to the Closed Track rules, using no external data or pre-trained OCR models other than the permitted base model.

3. Experiments

3.1 Experimental Settings

Platform: Alibaba Cloud PAI-DSW

GPU: NVIDIA A10 (24GB)
Base Model: Qwen2.5-VL-7B-Instruct
LoRA Config: Rank = 16, Alpha = 32, Dropout = 0.05
Target Modules: q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj
Learning Rate: 2e-4 (with cosine decay)
Batch Size: 1 (with Gradient Accumulation Steps = 16)
Max Pixels:~1,200,000 (approx.)
Epochs: 3
Given the extremely high-resolution nature of ancient document images (Max Pixels ≈ 1,200,000) and the hardware constraint of a single 24GB VRAM GPU (NVIDIA A10), full-

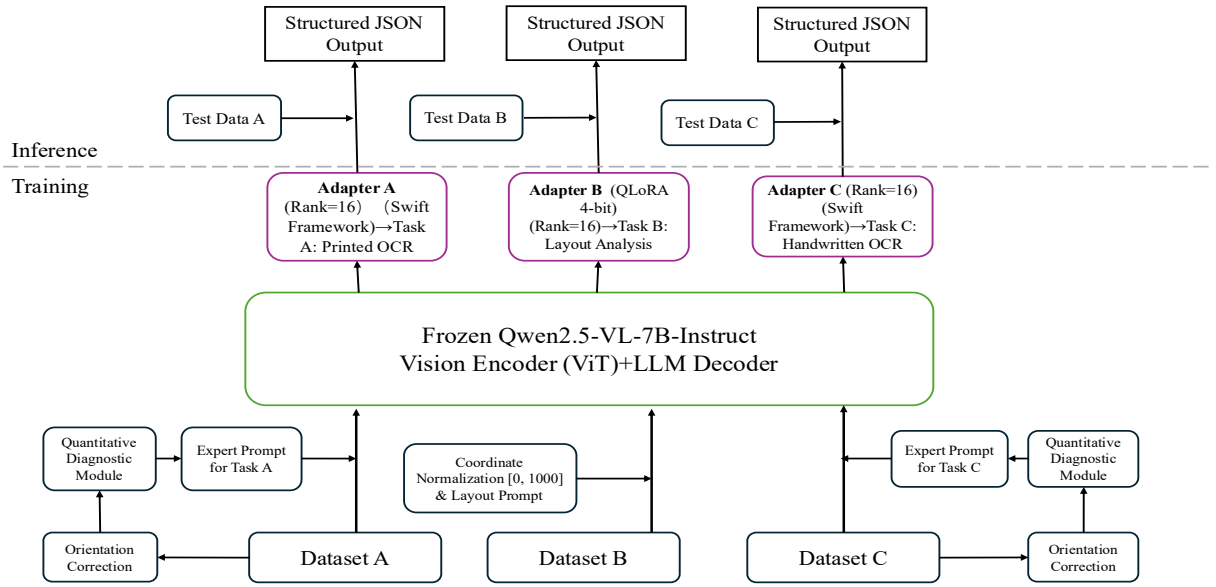


Figure 1 : *The unified multimodal framework. The pipeline is divided into training and inference phases. Raw datasets undergo quantitative diagnostics and task-specific geometric normalization (e.g., coordinate scaling for Task B). We explicitly apply parameter-efficient fine-tuning via three distinct LoRA adapters (Rank=16) to a frozen Qwen2.5-VL-7B backbone, utilizing QLoRA 4-bit precision and the Swift framework to manage hardware constraints.*

precision fine-tuning of a 7B multimodal model was computationally prohibitive. For Task A and C, we utilized the highly optimized Swift LLM framework (Qiu et al., 2023) to manage memory efficiently during training. Crucially, for the computationally intensive layout analysis task (Task B), we implemented 4-bit Quantized Low-Rank Adaptation (QLoRA) (Dettmers et al., 2023). This mixed-precision engineering effectively compressed the frozen base model's memory footprint, allowing us to maintain high-resolution visual encoding without encountering Out-Of-Memory (OOM) errors, ensuring stable convergence over the 8-hour training cycle.

3.2 Results

We evaluated our system against the official baseline provided by the organizers, which also utilizes Qwen2.5-VL-7B-Instruct via the Swift framework but employs a heavier LoRA Rank of 32 on all-linear modules without specific data orientation preprocessing.

Task	Metric	Official Baseline (Rank 32)	Ours (Rank 16)	Improvement
Task A (Print)	CER ↓	0.0618	**0.0305**	-50.6%
	F1-Score ↑	0.9430	**0.9719**	+3.06 %
	Overall ↑	0.9397	**0.9703**	+3.25 %
Task B (Layout)	mAP ↑	0.2006	**0.2749**	+37.0 %
	Macro F1 ↑	0.1530	**0.4162**	+172.0 %
	Avg Match IoU ↑	0.6600	**0.6740**	+2.12 %
Task C (Hand)	CER ↓	**0.0920**	0.1010	-
	F1-Score ↑	**0.9099**	0.9018	-
	Overall ↑	**0.9086**	0.9000	-

Table 1 : Performance Comparison (Closed Track)(%)

As shown in Table 1, our system achieved strong performance. Most notably on Task A, our method halved the Character Error Rate (CER) to 3.05% and achieved an Overall score of 0.9703. On the highly challenging generative layout analysis task (Task B), we substantially outperformed the official baseline, improving the Macro F1 score by a substantial 172% (0.4162 vs. 0.1530). For handwritten texts (Task C), our model maintained highly competitive parity with the baseline (>90% F1), despite utilizing significantly fewer trainable parameters.

4. Discussion

4.1 Parameter Efficiency and Data-Centric Superiority

A critical insight from our experiments is the triumph of data-centric preprocessing over raw parameter scaling. The official baseline employed a higher LoRA Rank (32) and a large global batch size of 32. In contrast, our approach utilized a lightweight setup with a LoRA Rank of 16 targeting specific projection layers, coupled with a smaller effective batch size (Batch Size 1 with 16 accumulation steps). Despite using considerably fewer trainable parameters, our model significantly outperformed the baseline on Tasks A and B. This observation supports the premise that lower intrinsic rank and smaller batch sizes can sometimes act as a regularizer, preventing overfitting on specialized, low-resource datasets like ancient texts (Masters and Luschi, 2018; Liu et al., 2024). We attribute this success to the synergy between our orientation-aware image preprocessing and expert-constrained prompt engineering. We executed a batch rotation to unify the physical geometry of horizontal texts with the model's vertical reading priors. Furthermore, by reinforcing this with strict top-to-bottom layout constraints in the prompt, we effectively guided the MLLM's visual attention. This demonstrates that aligning the physical spatial structure of the input with the semantic constraints of the prompt is crucial for unlocking the latent capabilities. This phenomenon is corroborated by recent studies on visual grounding through structured prompting (Yang et al., 2024), and far surpasses the baseline's agnostic approach.

The Semantic Correction Phenomenon: Furthermore, the official evaluation data provides compelling evidence for the MLLM's intrinsic "semantic correction" capability. We observed a distinct performance gap in Task C (handwritten texts) depending on the evaluation metric: our system scored significantly higher when including variant characters (F1: 0.9018, Overall: 0.9000) compared to excluding them (F1: 0.8873, Overall: 0.8854). Interestingly, this gap was entirely absent in Task A, where the highly standardized woodblock fonts left little room for visual ambiguity. This contrast confirms that our lightweight adapter did not merely memorize isolated pixel shapes. Instead, when confronting the severe visual ambiguity inherent in handwritten historical manuscripts, the model successfully mapped the visual features of variant characters to their

normative semantic equivalents in the MLLM's latent space—a valuable capability for digital philology.

4.2 The Inherent Limits of Generative Coordinate Regression

On Task B, our model demonstrated exceptional categorization ability, outperforming the baseline's Macro F1 by 172% (0.4162 vs. 0.1530). However, the absolute mAP score remains modest at 0.2749. This reveals a fundamental characteristic of current Vision-Language Models: while MLLMs excel at semantic classification and rough regional identification (evidenced by our high F1 and 0.6740 Avg Match IoU), treating continuous spatial coordinates as discrete text tokens (Chen et al., 2021) leads to suboptimal pixel-level bounding box regression. In our architecture, the absolute pixel coordinates are mapped into 1000 discrete bins (i.e., coordinate normalization [0, 1000]). This discretization, inherent to MLLMs, inherently limits pixel-perfect localization. Furthermore, due to VRAM constraints on the NVIDIA A10, we implemented 4-bit QLoRA. While this mixed-precision approach successfully prevented memory exhaustion, the quantization-induced precision loss and gradient noise likely acted as a contributing factor to the modest absolute mAP score. We hypothesize that utilizing higher-bit precision, if hardware had permitted, would yield a noticeable performance gain in coordinate regression. Our data-driven model successfully locates the existence and semantic type of layout elements, struggling primarily with boundary micromanagement compared to traditional discriminative object detectors (Wang et al., 2023).

5. Conclusion

In the EvaHan 2026 competition, we demonstrated that a parameter-efficient, data-centric approach can redefine the state-of-the-art for Ancient Chinese OCR. By strictly using official data alongside a highly optimized Low-Rank Adaptation (Rank 16), we achieved near-human accuracy (97.19% F1) on printed texts, robust parity on handwritten manuscripts, and a 37% mAP enhancement in layout analysis over the official generative baseline. Our work provides strong evidence that coupling explicit document structural priors with Multimodal LLMs is the most effective pathway for end-to-end historical document processing.

While our parameter-efficient framework demonstrates strong overall performance, we acknowledge several limitations that highlight avenues for future research. First, the lack of targeted optimization for handwritten variant characters and degraded glyphs resulted in Task C performance slightly trailing the official baseline. Second, regarding layout analysis, the modest absolute mAP score exposes the inherent constraints of coordinate discretization and quantization-induced precision loss. Furthermore, due to computational resource limitations, comprehensive ablation studies isolating the individual contributions of orientation preprocessing and adaptive prompt engineering were not conducted. Finally, our current approach trains three task-specific adapters independently; future work will explore cross-task generalization to build a unified multimodal model capable of simultaneously handling OCR and layout analysis.

6. Acknowledgments

We would like to thank the anonymous reviewers for their constructive comments. We also extend our gratitude to the EvaHan 2026 organizers for providing the datasets and evaluation platform.

7. Bibliographical References

Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., ... & Schmidt, L. (2023). OpenFlamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

Chen, T., Saxena, S., Li, L., Fleet, D. J., & Hinton, G. (2021). Pix2seq: A language modeling framework for object detection. *International Conference on Learning Representations*.

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Li, B., Feng, M., Hao, X., et al. (2022). Overview of the EvaHan 2022 Shared Task on Ancient Chinese Word Segmentation and POS Tagging. *Proceedings of the LT4HALA 2022 Workshop*.

Li, M., Lv, T., Chen, L., Cui, L., Yin, Y., Hu, W., ... & Wei, F. (2021). TrOCR: Transformer-based optical character recognition with pre-trained models. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Liu, S. Y., et al. (2024). DoRA: Weight-Decomposed Low-Rank Adaptation. *International Conference on Machine Learning (ICML)*.

Liu, Y., et al. (2024). TextMonkey: An OCR-Free Large Multimodal Model for Understanding Document. *arXiv preprint arXiv:2403.04473*.

Masters, D., & Luschi, C. (2018). Revisiting small batch training for deep neural networks. *arXiv preprint arXiv:1804.07612*.

Qiu, L., et al. (2023). Swift: Scalable lightweight infrastructure for fine-tuning. GitHub repository, https://github.com/modelscope/swift.

Wang, J., Jin, L., Ding, K., & Liao, C. (2023). Document pre-training with large language models. arXiv preprint arXiv:2311.03810.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., ... & Lin, J. (2024). Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Yang, Z., et al. (2024). Set-of-Mark Prompting Unleashes Extraordinary Visual Grounding in GPT-4V. *arXiv preprint arXiv:2310.11441*.

LVLM Optimization for Ancient Chinese Book Image Analysis with Task-specific Augmentation and Instruction Tuning

Xia Tian[1], Liu Yulong[2], Wang Yilin[1], Yang Yumeng[1]
Cai Dongheng[1], Tan Yuyang[3],Yang Menghui[1]*
[1]School of Information Resource Management,Renmin University of China, Beijing, China,
[2]Midu Technology Co., Ltd., Shanghai,[3]China Washington University in St. Louis
xiat@ruc.edu.cn, lcayvinl@gmail.com, wang_yilin@ruc.edu.cn,
{yymeng1219,dongheng_c}@163.com,tan@wustl.edu,yangmenghui@ruc.edu.cn

Abstract

Ancient Chinese text digitization faces challenges like variant characters and complex layouts. Based on the EvaHan 2026 tasks, this study proposes an LVLM-based framework for printed/handwritten text recognition and layout analysis. To effectively adapt the Qwen2.5-VL-7B-Instruct model, our methodology innovates through a dual-level optimization strategy: distinct augmentation strategies are developed for OCR and layout tasks, while task-specific prompt templates are engineered to decouple text transcription from coordinate prediction. This combined approach significantly enhances overall task proficiency, achieving Character Error Rates of 0.0372 (printed) and 0.0823 (handwritten), alongside a mean average Precision of 0.2933 for layout analysis. Results show general LVLMs underperform in zero-shot ancient text tasks, but fine-tuning with tailored strategies significantly boosts performance and highlights their potential.

Keywords: Optical Character Recognition, Layout Analysis, Ancient Texts, Large Vision-Language Models

1. Introduction

Ancient Chinese texts are central to the study of Chinese civilization. Their large-scale digitization and intelligent use are vital for cultural preservation and academic research(Shi et al., 2025).However, processing these historical materials is highly challenging. Beyond the difficulties of recognizing variant characters and damaged strokes, deeply understanding the document layout is crucial for accurate digitization(Luo et al., 2024). This is particularly difficult in ancient texts due to their complex structural elements—such as the intermingling of text blocks, margins, seals, and illustrations—alongside mixed handwritten and printed scripts. Together, these factors pose significant challenges for traditional Optical Character Recognition (OCR) technologies.

In recent years, Large Vision-Language Models (LVLMs) have shown strong potential in document analysis and scene text recognition, thanks to their ability to jointly process visual and linguistic information. As a result, evaluating and improving LVLM performance on ancient document images has become a key research direction.

We present a systematic LVLM optimization framework for the EvaHan 2026 ancient Chinese OCR evaluation. By fine-tuning Qwen2.5-VL-7B-Instruct with LoRA and task-specific prompts , alongside tailored data augmentation—geometric transformations/hybrid stitching for text recognition and background embedding for layout analysis —our approach outperformed baselines across all tracks. Code is publicly available on GitHub[1].

2. Related Works

OCR research has undergone decades of iterative development. Early studies primarily relied on traditional machine learning algorithms, such as Support Vector Machines (SVM) and Hidden Markov Models (HMM). These methods achieved character recognition by extracting geometric and textural features. However, they exhibited poor adaptability to variant characters and degraded strokes in ancient texts, resulting in limited recognition accuracy. With the advancement of deep learning, OCR models based on Convolutional Neural Networks (CNN) and Transformers—such as CRNN (Shi et al., 2016) and TrOCR (Li et al., 2023) —have emerged as the predominant approaches in the field. By implementing end-to-end character detection and recognition, these models have made significant progress in printed text scenarios. Nevertheless, in complex scenarios involving handwritten scripts and mixed layouts, challenges such as disordered reading sequences and high character error rates persist.

Recently, the emergence of LVLMs has brought new breakthroughs to ancient text processing(Wang and Zhu, 2025). Several models, such as Qwen2.5-VL(Bai et al., 2025), simplify the workflow by directly processing "image-text" pairs through enhanced visual recognition and cross-modal fusion, bypassing the need for separate, multi-stage detection and recognition. Additionally, DocLayout-YOLO

[1]https://github.com/iamxiatian/evahan2026

Proceedings of LT4HALA 2026 @ LREC 2026, pages 299–304.
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

has substantially improved layout element analysis for general documents through data synthesis and model optimization(Zhao et al., 2024). However, most existing multimodal models are designed for general-purpose scenarios and lack specialized optimization for the unique element layouts and calligraphic styles of ancient texts. Consequently, their recognition precision and layout analysis capabilities in ancient document contexts remain to be validated. The EvaHan 2026 international evaluation has catalyzed applied research into LVLMs for ancient text processing, providing a unified benchmark for comparing the performance of diverse methodologies.

3. Methodology

3.1. Overall Framework and Baseline Model Selection

As illustrated in Figure 1, the proposed recognition and optimization framework for ancient text and layout elements consists of four interconnected stages: data augmentation, model fine-tuning, model inference, and result parsing. Within this pipeline, the data augmentation module processes raw images through geometric transformations, stitching, and background embedding to generate datasets more conducive to model learning. These augmented datasets are then fed into the model fine-tuning module, which utilizes the ms-swift framework and LoRA technology to derive optimized model weights. Subsequently, the inference module converts user inputs into model-compatible formats and generates textual outputs. Finally, the result parsing module analyzes the raw outputs from the LVLM, extracting key information to generate structured results.

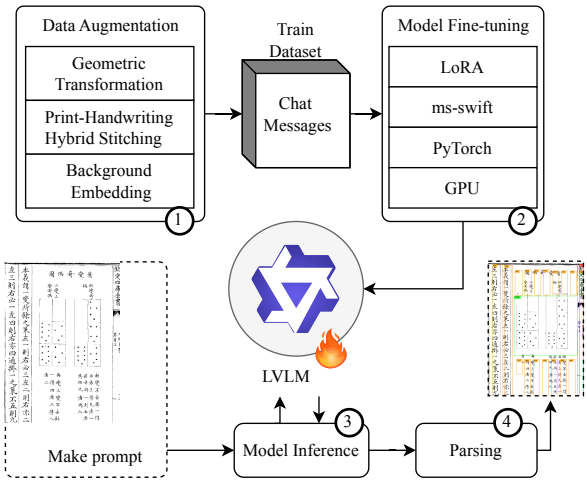


Figure 1: The overall architecture of the proposed recognition and optimization framework

In terms of baseline selection, we conducted a comparative analysis of the fine-tuning effectiveness between Qwen2.5-VL-7B-Instruct (Qwen) and Xunzi_Qwen2_VL_7B_Instruct (Xunzi) across the training sets of the three sub-tasks. Preliminary evaluations demonstrated that Qwen exhibited superior performance across all metrics; consequently, it was adopted as the primary backbone for this study. By integrating a native dynamic-resolution Vision Transformer and a window attention mechanism, this model effectively captures the multi-scale and high-dimensional complexity inherent in ancient Chinese document images.

3.2. Data Augmentation Strategies

Data augmentation serves as a critical mechanism for enhancing model generalization and performance. Given the specific characteristics of ancient Chinese texts, we implemented distinct augmentation strategies for the OCR tasks and the layout element analysis task.

3.2.1. Data Augmentation for OCR Tasks

For the closed track, we employed geometric transformation augmentation. Each image was subjected to a series of randomized operations, including horizontal/vertical flips, affine transformations, and random cropping. By generating 4 to 5 variants per original image, we expanded the training set to over 20,000 samples. This approach significantly bolsters the model's robustness to geometric distortions commonly found in scanned ancient documents.

For the open track, we introduced a hybrid print-handwriting stitching strategy. Pairs of images were sequentially selected from Dataset A (printed) and Dataset C (handwritten), denoted as a and c, respectively. These were then concatenated vertically in two ways (a above c, and c above a) to form new composite images. The width of the resulting image was determined by the larger of the two originals, and their corresponding text labels were concatenated accordingly. A total of 10,000 hybrid samples were generated. This strategy encourages the model to differentiate between the two calligraphic styles within a single input, effectively yielding an implicit multi-task learning effect.

3.2.2. Data Augmentation for Layout Element Analysis

For the closed track, we implemented a background embedding augmentation strategy to enhance spatial generalization and mitigate coordinate instability. This approach addresses two primary challenges. First, it prevents the model from memorizing absolute coordinates. Second, it mitigates stability issues caused by variations in input image

dimensions. Specifically, we initialize a random light-colored background canvas with fixed dimensions of 924 $\times$ 1232 pixels. Original images are stochastically positioned within this canvas; any image exceeding these dimensions is proportionally rescaled prior to placement. For each sample, two augmented variants are generated: one anchored at the top-left corner and another at a randomized location, ensuring the entire image remains within the canvas boundaries. By decoupling layout elements from fixed coordinate systems, this strategy forces the model to rely on intrinsic layout features, thereby improving localization robustness and spatial generalization.

For the open track, the training set was directly converted into the DocLayout-YOLO training format. We discarded the textual information contained within the layout elements of the original data and focused exclusively on generating bounding box coordinates and category labels for the layout elements in Dataset B.

3.3. Instruction Design and Fine-tuning Methodology

Instructions serve to guide the behavior of LVLMs and constrain their output space. In contrast to specialized traditional models, LVLMs function under a conversational generative paradigm, leveraging its inherent multi-task capabilities which are activated through specific prompt cues. Even after Parameter-Efficient Fine-Tuning (PEFT) such as LoRA, the model primarily optimizes the mapping between specific instructions and target task outputs, without altering its fundamental reliance on prompt-based guidance. While fine-tuning enables the model to acquire specialized competencies in text transcription and coordinate prediction, meticulously designed instructions are crucial for defining task roles, output formats, and specific constraints. Furthermore, in scenarios where a single model handles multiple sub-tasks, instructions act as the unique identifier for task differentiation, effectively suppressing hallucinatory outputs irrelevant to the evaluation criteria.

We designed differentiated prompt templates for the three sub-tasks to activate the model's fine-tuned knowledge and standardize the output formats. Specifically, for Task A and Task C, the prompts emphasize natural reading order, the handling of special characters, and the prohibition of extraneous content. For the Layout Analysis task, the instructions specify element categories, coordinate formats, and text content requirements, supplemented by examples to reduce the model's cognitive load. Detailed prompt templates are available in our open-source repository on GitHub. The model was fine-tuned using LoRA(Hu et al., 2022) , an PEFT technique that keeps the pre-trained weights frozen while introducing a minimal set of trainable parameters.Its core principle involves using low-rank factorization to simulate parameter updates during task adaptation. By updating only a fraction of the total parameters, LoRA significantly reduces GPU memory consumption and accelerates training efficiency. Both training and inference were implemented using the ms-swift framework(Zhao et al., 2025).

4. Experiments

4.1. Datasets

We strictly followed the official data partitioning of EvaHan 2026, utilizing the training set for fine-tuning and the test set for evaluation. The EvaHan 2026 dataset comprises three categories of image-text pairs: plain text images(Dataset A), mixed layout images(Dataset B), and handwritten text images(Dataset C). These high-quality datasets were generated through automated annotation followed by expert revision. For the OCR tasks, we merged Dataset A and Dataset C, along with their respective augmented variants, to form the training corpus. For the Layout Element Recognition task, Dataset B was first converted into a standardized labeling format to serve as training data for DocLayout-YOLO. Subsequently, these images were processed via background embedding to generate 10,000 samples for fine-tuning the LVLM.

4.2. Parameter Settings

Model fine-tuning was performed using the SWIFT framework, employing Parameter-Efficient Fine-Tuning via LoRA. The LoRA rank (r) was set to 8, with an alpha of 32, and the target modules encompassed all linear layers; this configuration represented a trade-off to optimize VRAM efficiency under experimental constraints. To determine the optimal status for the visual encoder, preliminary empirical evaluations were conducted during the initial two training epochs. Results indicated that unfreezing the ViT yielded no measurable performance gains while increasing computational overhead. Consequently, the visual encoder was frozen (freeze_vit=true) to optimize training efficiency and mitigate potential overfitting, restricting parameter updates to the language model and alignment layers. Key hyperparameters were standardized as follows: a learning rate of 2×10^{-5}, a global batch size of 32 (achieved via a per-device batch size of 2, gradient accumulation of 4, and 8 GPUs), 8 training epochs, the AdamW optimizer, and a warmup ratio of 0.05. Due to the variance in output lengths between OCR and layout tasks, the maximum sequence length was set to 2,048 for Tasks A/C and

8,192 for Task B. All experiments were conducted on a cluster of 8 NVIDIA A800 GPUs.

Regarding checkpoint selection, for the OCR tasks, we utilized the final checkpoint (Epoch 8) for inference. In contrast, for the layout analysis task, the model was trained for a total of 15 epochs with checkpoints saved every 100 steps. To mitigate overfitting and ensure optimal generalization, we selected an intermediate checkpoint at Step 1,900 (approximately the 12th epoch) for the final submission, as it demonstrated superior performance on a manually sampled validation set. During the inference phase, the vLLM toolkit(Kwon et al., 2023) was employed for acceleration, with the temperature set to 0.1, top_p to 0.9, and the maximum sequence length to 8,192. For the layout task, the model's raw HTML-tagged outputs underwent post-processing to extract bounding boxes (bboxes) and textual content, which were subsequently converted into the JSON format required for the official evaluation.

4.3. Evaluation Metrics

Following the official EvaHan 2026 evaluation metrics, we adopt CER, NED, F1-score, and C-Score for the OCR tasks (Tasks A and C). These metrics are calculated in two versions: considering and ignoring variant characters. For layout analysis (Task B), mAP@[.5:.95], Micro/Macro F1, and Average Match IoU are utilized. We refer readers to the official guidelines for detailed metric formulations[2].

4.4. Methods and Results

To evaluate the effectiveness of the proposed framework, we conducted comparative experiments involving both baselines and our proposed models. The detailed technical specifications of these eight methods are summarized in Table 1 in the Appendix.

As demonstrated in Tables 2, 3, and 4 (see Appendix), our proposed methods significantly outperform all baselines across all evaluated metrics. In the closed track, for Task A (including variant characters), the CER decreased from 0.0618 (baseline) to 0.0372—a substantial reduction of 39.8%—while the F1-score rose to 0.9663, marking a 2.47% improvement over the baseline's 0.9430. For Task C (including variants), the CER dropped from 0.0920 to 0.0823 (a 10.5% decrease), with the F1-score increasing by 1.08% to reach 0.9197. Most notably, in Task B, the proposed method demonstrated significant improvements in mAP, Micro F1, Macro F1, and Avg Match IoU, with improvements of 46.21%, 370.57%, 145.23%, and 7.30%, respectively, compared to the best-performing baseline. These results validate the efficacy of our geometric transformation and background embedding strategies.

In the open track results, the proposed hybrid print-handwriting stitching strategy consistently surpassed the best baseline results for both Task A and Task C, although its performance was slightly lower than that of the geometric transformation augmentation. For Task B, we fine-tuned DocLayout-YOLO, a specialized layout analysis model pre-trained on 300K synthetic document pages. This model employs a Controllable Receptive Field module and a Global-to-Local design to refine multi-scale features. After fine-tuning on Dataset B, it achieved a mAP@[.5:.95] of 0.3828 and Micro F1 of 0.6673, substantially outperforming the closed-track LVLM (0.2933 mAP and 0.2414 Micro F1). This demonstrates the advantage of specialized models for precise layout analysis.

4.5. Discussion

LVLMs provide a powerful means of understanding the layout composition and textual content of ancient document images. For OCR tasks, LVLMs demonstrate robust baseline recognition capabilities due to extensive pre-training, which can be further refined through fine-tuning. However, in layout element recognition, the zero-shot performance of LVLMs is relatively poor; while fine-tuning yields substantial improvements over the baseline, the overall performance remains suboptimal.

Our experiments reveal that meticulous instruction design significantly reduces output errors, indicating that LVLMs are highly sensitive to prompts. Consequently, task design should strategically leverage the model's in-context learning capabilities. Furthermore, targeted fine-tuning on domain-specific data consistently enhances performance, though the efficacy is heavily influenced by specific data augmentation techniques and instruction strategies. In layout analysis, the background embedding strategy, which enforces the learning of relative spatial layouts, resulted in a substantial improvement in performance compared to the baseline. For OCR tasks—encompassing both printed and handwritten scripts—our geometric transformation and hybrid stitching strategies outperformed the baseline models. Notably, the hybrid stitching strategy was slightly less effective than geometric transformation, likely because direct stitching disrupts textual coherence; future research could explore more natural synthesis methods for mixed-media data.

In the context of layout analysis, LVLMs currently face higher training costs and slower inference speeds, with precision currently remaining below that of specialized layout-specific architectures.

[2] https://github.com/GoThereGit/EvaHan

Nevertheless, their primary advantage lies in end-to-end semantic understanding. Future work could focus on constructing hybrid architectures that combine the strengths of both specialized detectors and multimodal reasoners.

Regarding variant characters, the marginal performance gap between "variant-aware" and "standard" metrics in printed texts suggests that the model recognizes printed characters with high accuracy and is minimally affected by variant interference. In contrast, for handwritten scripts, this gap exceeds one percentage point. This indicates that in over 1% of cases, the model's output is semantically appropriate but fails to map to the specific character in the variant set. This ambiguity aligns with the inherent variability in calligraphic styles across different scribes, presenting challenges analogous to those encountered in human expert decipherment of ancient manuscripts. Overall, the model performs better under variant-aware evaluation conditions, demonstrating strong morphological generalization and an ability to correctly identify and map variants to their standard counterparts.

5. Conclusion and Future Work

This paper presents a robust framework based on LVLMs, optimized through strategic data augmentation and instruction fine-tuning. Our methods achieved performance that significantly surpassed the baseline models across all three tasks of the EvaHan 2026 international evaluation, thereby validating the efficacy of the proposed approach. Experimental evidence highlights that tailored augmentation strategies, such as background embedding and hybrid stitching, are crucial for adapting general-purpose LVLMs to the nuances of ancient Chinese texts.

Furthermore, as the test set ground truth is currently withheld by the evaluation organizers, we are temporarily unable to present comprehensive ablation and error analyses. In the future, we plan to conduct systematic ablation experiments to quantify the individual contributions of each proposed module, alongside an in-depth analysis of failure cases to guide further optimizations.

In future work, we also intend to explore the fusion of diverse data augmentation strategies and investigate more sophisticated techniques to address the remaining challenges in spatial localization for layout elements within large models. By refining these mechanisms, we aim to propel the digitization of ancient documents toward higher precision and broader academic utility.

6. Acknowledgements

This research is supported by the National Social Science Fund of China (22BTQ068).

7. Bibliographical References

Shuai Bai, Keqin Chen, Xuejing Liu, et al. 2025. Qwen2.5-VL technical report.

Edward J. Hu, Yelong Shen, Phillip Wallis, et al. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, et al. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Minghao Li, Tengchao Lv, Jingye Chen, et al. 2023. TrOCR: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*, pages 13094–13102.

Chuwei Luo, Yufan Shen, Zhaoqing Zhu, et al. 2024. LayoutLLM: layout instruction tuning with large language models for document understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15630–15640.

Baoguang Shi, Xiang Bai, and Cong Yao. 2016. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2298–2304.

Yongxin Shi, Dezhi Peng, Yuyi Zhang, et al. 2025. A large-scale dataset for chinese historical document recognition and analysis. *Scientific Data*, 12(1):169.

Dongbo Wang and Dongmei Zhu. 2025. Benchmarking the ancient books capability of multimodal large language models. *Npj Heritage Science*, 13(1):339.

Yuze Zhao, Jintao Huang, Jinghan Hu, et al. 2025. SWIFT: A scalable lightweight infrastructure for fine-tuning.

Zhiyuan Zhao, Hengrui Kang, Bin Wang, et al. 2024. DocLayout-YOLO: Enhancing document layout analysis through diverse synthetic data and global-to-local adaptive perception.

A. Appendix: Detailed Experimental Results

Method	**Backbone**	**FT Method**	**Technical Strategy**	**Target Task**
Baseline-A	Qwen2.5-VL-7B	Zero-shot	Zero-shot instruction engineering without fine-tuning.	Official Baseline
Baseline-B	Qwen2.5-VL-7B	LoRA	Standard LoRA fine-tuning on task-specific datasets.	Official Baseline
Baseline-C	Xunzi_Qwen2_VL_7B	Zero-shot	Zero-shot instruction engineering without fine-tuning.	Official Baseline
Baseline-D	Xunzi_Qwen2_VL_7B	LoRA	Standard LoRA fine-tuning on task-specific datasets.	Official Baseline
ACBM_GT	Qwen2.5-VL-7B	LoRA	Incorporate Geometric Transformation (GT) augmentation.	Closed: Task A & C
ACBM_HS	Qwen2.5-VL-7B	LoRA	Utilize Hybrid Stitching (HS) for print and handwritten scripts.	Open: Task A & C
ACBM_BE	Qwen2.5-VL-7B	LoRA	Employ Background Embedding (BE) by integrating document samples into light-colored backgrounds.	Closed: Task B
ACBM_YOLO	DocLayout-YOLO	Full Fine-tuning	Employ DocLayout-YOLO-DocLayNet-Docsynth300K_pretrained model for layout element detection.	Open: Task B

Table 1: Description of Comparative Methods

Method	Consider variant characters				Don't consider variant characters			
	CER	NED	F1	C-Score	CER	NED	F1	C-Score
Baseline-A	0.1014	0.0947	0.9110	0.9037	0.1121	0.1054	0.9007	0.8931
Baseline-B	0.0618	0.0613	0.9430	0.9397	0.0685	0.0679	0.9364	0.9331
Baseline-C	0.1786	0.1740	0.8409	0.8282	0.1851	0.1802	0.8345	0.8218
Baseline-D	0.1214	0.1183	0.8993	0.8854	0.1264	0.1232	0.8945	0.8805
ACBM_GT(Close)	**0.0372**	**0.0369**	**0.9663**	**0.9639**	**0.0382**	**0.0379**	**0.9653**	**0.9629**
ACBM_HS(Open)	**0.0476**	**0.0467**	**0.9560**	**0.9537**	**0.0490**	**0.0481**	**0.9546**	**0.9523**

Table 2: Results of Task A

Method	mAP@[.5:.95]	Micro F1	Macro F1	Avg Match IoU
Baseline-A	0	0	0	0
Baseline-B	0.2006	0.0513	0.1530	0.6600
Baseline-C	0.0236	0.0003	0.0114	0.5943
Baseline-D	0.1917	0.0403	0.1130	0.6654
ACBM_BE(Close)	**0.2933**	**0.2414**	**0.3752**	**0.7140**
ACBM_YOLO(Open)	**0.3828**	**0.6673**	**0.5983**	**0.8041**

Table 3: Results of Task B

Method	Consider variant characters				Don't consider variant characters			
	CER	NED	F1	C-Score	CER	NED	F1	C-Score
Baseline-A	0.1207	0.1193	0.8849	0.8812	0.1338	0.1324	0.8718	0.8681
Baseline-B	0.0920	0.0919	0.9099	0.9086	0.1066	0.1065	0.8953	0.8940
Baseline-C	0.1497	0.1492	0.8538	0.8514	0.1609	0.1604	0.8425	0.8402
Baseline-D	0.1383	0.1376	0.8673	0.8635	0.1520	0.1512	0.8534	0.8498
ACBM_GT(Close)	**0.0823**	**0.0822**	**0.9197**	**0.9183**	**0.0952**	**0.0951**	**0.9068**	**0.9054**
ACBM_HS(Open)	**0.0853**	**0.0849**	**0.9168**	**0.9154**	**0.1030**	**0.1026**	**0.8991**	**0.8977**

Table 4: Results of Task C

Data-Centric Strategies for Ancient Chinese Text Recognition: Augmentation, Annotation Refinement, and Style Transfer in EvaHan 2026

Chengfei Li[1] , Yunjie Zhang[1] , Xiaoyi Li[1] , Changshun Quan[1] , Taihe Cao[1] , Bin Liu[1]

[1] Artificial Intelligence Research Institute on Education, Qilu Normal University, China
lichengfei@qlnu.edu.cn; 18769301321@163.com; 15698042890@163.com; 19707032403@163.com; cth@qlnu.edu.cn; liubin@qlnu.edu.cn

Abstract

This paper describes our system for the EvaHan 2026 shared task. We design and experiment with data-centric strategies across three subtasks: printed text OCR (Task A), layout element analysis (Task B), and handwritten text OCR (Task C). Our approach employs systematic data augmentation using 17 transformation strategies, comprehensive manual annotation refinement for layout analysis, and style transfer augmentation for handwritten texts. We use pre-trained Qwen2.5-VL-7B-Instruct with LoRA fine-tuning as the base model. According to the evaluation metrics adopted by the organizers, our system achieves 27.5% and 4.5% CER reduction over official baselines for Tasks A and C respectively. Manual annotation refinement for Task B achieves 205% improvement in Micro F1 and 258% improvement in Macro F1 on the validation set, demonstrating that annotation quality is the primary bottleneck for layout analysis in closed-modality settings.

Keywords: ancient Chinese OCR, data augmentation, annotation refinement, multimodal language models

1. Introduction

Ancient Chinese texts represent invaluable cultural heritage, documenting thousands of years of civilization through printed books, hand-copied manuscripts, and administrative documents. However, their automated processing remains challenging due to several factors: character variants across different dynasties and regions, material degradation from aging and environmental damage, complex multi-column layouts with embedded illustrations and seals, diverse calligraphic styles reflecting individual scribal traditions, and critically, the scarcity of high-quality annotated training data. The EvaHan 2026 evaluation campaign aims to assess the capabilities of large language models in addressing these challenges across three dimensions: printed text recognition (Task A), layout element analysis (Task B), and handwritten text recognition (Task C).

We design and experiment with three data-centric approaches. The systematic data augmentation approach expands training data through photometric and geometric transfor-mations simulating realistic degradation patterns. The annotation refinement approach manually corrects layout element labels to address systematic errors including missing text regions. The style transfer approach generates artistic variants for handwritten texts using traditional Chinese artistic styles.

We participate exclusively in the closed track, which restricts participants to official training data and two specified models: Qwen2.5-VL-7B-Instruct and Xunzi_Qwen2_VL_7B_Instruct. This constraint forces data-centric optimization rather than model scaling.

According to the evaluation metrics adopted by the organizers, our system achieves competitive performance across all three tasks. Compared to the baseline Qwen2.5-VL-7B-Instruct, our augmentation strategies obtain substantial improvements for Tasks A and C. Our annotation refinement achieves dramatic gains for Task B on validation data, though test performance reveals annotation quality issues in the original dataset. Our code and annotation tool are available at https://github.com/SeekingYourRoots/EvaHan2026-TaskB-Annotation-Tool.

2. Related Work

Ancient Document OCR Deep learning has revolutionized historical OCR through end-to-end architectures. Shi et al. (2017) proposed CRNN combining CNN-RNN-CTC for sequence recognition. Attention-based models improved irregular layout handling (Cheng et al., 2017). Recent vision-language models like TrOCR (Li et al., 2023) and Donut (Kim et al., 2022) leverage large-scale pre-training for document understanding.

Data Augmentation Simard et al. (2003) demonstrated elastic distortions improve handwriting recognition. Buslaev et al. (2020) introduced Albumentations for fast augmentation pipelines. Cubuk et al. (2019) proposed AutoAugment using reinforcement learning to discover optimal strategies.

Layout Analysis Modern approaches employ deep object detection. Faster R-CNN (Ren et al., 2015) and Mask R-CNN (He et al., 2017) enable accurate element localization. MMDetection (Chen et al., 2019) provides unified implementtations for document layout analysis.

Proceedings of LT4HALA 2026 @ LREC 2026, pages 305–310.
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

3. Our Method

In this section, we introduce our methods and model architectures for the three subtasks.

The EvaHan 2026 task encompasses three subtasks operating on different data modalities. Task A (printed texts) benefits from systematic photometric degradation simulation. Task B (layout analysis) requires high-quality spatial annotations. Task C (handwritten texts) exhibits calligraphic style constraints limiting augmentation intensity.

We employ Qwen2.5-VL-7B-Instruct as the base model across all tasks, using LoRA for parameter-efficient fine-tuning. This unified architecture isolates the effects of data interventions.

3.1 Data Augmentation Pipeline

We implement 17 transformation strategies using Albumentations organized into four categories:

Color/Lighting : RandomBrightnessContrast ($\pm$ 20%), RandomGamma ($\gamma \in [0.7,1.3]$), HueSaturationValue (H $\pm$ 10°, S $\pm$ 30, V $\pm$ 20), RGBShift ($\pm$20), and combined color jittering.

Blur: GaussianBlur (kernel 3-7, $\sigma \in [0.1,2.0]$), MotionBlur (kernel 3-7), MedianBlur (kernel 5), and average blur.

Noise: GaussNoise ($\sigma \in [10,50]$), JPEG compression (quality 50-80), PixelDropout (1%), and MultiplicativeNoise ($\times$0.9-1.1).

Mixed: Combined transformations including brightness+noise, gamma+blur, motion+compression, and HSV+noise+blur.

All transformations preserve original image dimensions, which is critical for maintaining bounding box annotation validity in Task B. The pipeline randomly selects one transformation per augmented image.

3.2 Annotation Refinement for Task B

Analysis of validation results revealed three systematic annotation errors:

Missing text regions: Substantial textual content present but not annotated, causing false positive penalties when models correctly detect unlabeled regions.

Element type confusion: Ambiguous bound-aries between book_edge/text, image/seal, and seal/text categories.

Boundary inaccuracies: Bounding boxes insufficiently covering element extents, reducing IoU scores.

We developed a web-based annotation tool (Figure 1) to facilitate manual correction. The tool features intuitive dual drawing modes, color-coded element types for clear distinction, efficient keyboard shortcuts, and real-time persistence of annotation data.

Figure 1: Web-based annotation tool interface showing layout element correction capabilities

We manually corrected all 5,000 Dataset B training images. The correction process focused primarily on adding missing text region annotations (approximately 68% of corrections) and resolving element type confusion (approximately 24% of corrections).

3.3 Style Transfer for Task C

We distinguish three traditional Chinese artistic styles and generate variants (Figure 2) using the Qwen-Image-Edit API:

Stone Inscription: Bold angular strokes with weathering effects simulate ancient stone tablet carvings. Prompt: "Transform this handwritten text into stone inscription style, as if carved on an ancient Chinese stone tablet, maintaining all character content exactly."

Wood Carving: Raised relief with wood grain texture emulates woodblock printing. Prompt: "Convert this handwritten text to wood carving style, as if carved on traditional Chinese wooden blocks, keeping all text content unchanged."

Ink Wash: Varying stroke thickness with ink diffusion effects is characteristic of traditional calligraphy. Prompt: "Rerender this handwritten text in traditional ink wash painting style using classical Chinese brush techniques while preserving all character forms."

Manual quality filtering is implemented to ensure content preservation, with approximately 8-12% of generated images requiring removal due to character modification errors or illegibility introduced during style transfer.

Figure 2: Style transfer examples showing the three traditional Chinese artistic styles: Stone Inscription with bold angular strokes (left), Wood Carving displaying raised relief texture (middle), and Ink Wash exhibiting stroke thickness variation and ink diffusion (right).

3.4 Model Architecture

Our model architecture employs Qwen2.5-VL-7B-Instruct as the vision-language backbone. The model processes image-text pairs through:

Vision Encoder: Processes input images to extract visual features.

Language Model: Qwen2.5-7B transformer with 28 layers, 3584 hidden dimensions.

LoRA Adaptation: Low-rank adaptation matri-ces targeting query and value projection layers for parameter-efficient fine-tuning.

For Task B, the model outputs bounding box coordinates and element type classifications. For Tasks A and C, the model generates text sequences representing recognized characters.

4. Experiments

4.1 Data

The official training data originates from digitized ancient Chinese texts:

Dataset A (Printed Texts) consists of data selected from the Siku Quanshu (Complete Library of the Four Treasuries), including classics, history, philosophy, and literature, as well as various other ancient books.

Dataset B (Mixed Layouts) contains mixed image-text data selected from the Siku Quanshu and other ancient books.

Dataset C (Handwritten Texts) includes hand-written ancient books, primarily the Chinese Buddhist canon, including the Chinese Buddhist canon (TKH) dataset, and the Chinese Buddhist canon (MTH) dataset.

4.2 Training Configuration

Table 1 details training hyperparameters. All experiments use the same configuration to isolate data intervention effects.

Hyperparameter	Value
Base Model	Qwen2.5-VL-7B-Instruct
LoRA Rank	8
LoRA Alpha	16
Learning Rate	1e-4
Optimizer	AdamW
Batch Size	8
Epochs	3

Table 1: Training hyperparameters.

Training is conducted on 2×NVIDIA A800 GPUs. Each task requires approximately 1 hour for 3 epochs. The model achieves optimal validation performance in the third epoch across all tasks.

4.3 Task A: Printed Text Recognition

Table 2 presents Task A validation results.

Method	CER	F1	NED	Score
Qwen2.5 Baseline	0.2787	0.7386	0.2693	0.7284
Xunzi Baseline	0.2818	0.7325	0.2799	0.7229
Aug 1×	0.1413	0.8661	0.1380	0.8616
Aug 2×	0.0255	0.9772	0.0250	0.9754

Table 2: Task A validation set results.

Qwen2.5-VL-7B-Instruct outperforms Xunzi-Qwen2-VL-7B-Instruct, leading to an exclusive focus on the former. Data augmentation proves highly effective: 1× augmentation reduces Character Error Rate (CER) by 49.3% from 27.87% to 14.13%, while 2× augmentation achieves 90.9% reduction to 2.55%. This monotonic improvement suggests printed text variations are systematic and well-modeled by photometric transformations.

Official test results: Our final submission achieves CER 4.48%, Normalized Edit Distance (NED) 4.40%, F1-score 95.89%, comprehensive score 95.65%, outperforming the official baseline (CER 6.18%) by 27.5% in relative error reduction.

4.4 Task B: Layout Element Analysis

Table 3 summarizes Task B validation results comparing augmentation versus annotation refinement.

Method	mAP	Micro F1	Macro F1	Avg IoU
Baseline	0.3971	0.0308	0.1343	0.7271
Aug 1×	0.4021	0.0323	0.1285	0.7088
Aug 2×	0.3933	0.0320	0.1378	0.7020
Re-annotation	0.4091	0.0939	0.4808	0.7048

Table 3: Task B validation set results.

Data augmentation provides minimal improvement (mean Average Precision (mAP): 39.71% → 40.21%). In contrast, manual re-annotation leads to a substantial performance gain: Micro F1-score increases by 205% (3.08% → 9.39%) and Macro F1-score by 258% (13.43% → 48.08%). This indicates annotation quality is the main performance bottleneck.

Table 4 shows per-class results (TP: true positives, FP: false positives, FN: false negatives). The high FP count for text stems directly from missing annotations in the original dataset—models correctly detected these unlabeled regions but were penalized as FPs.

Metric	text	image	book_edge	seal
TP	338	228	304	66
FP	13072	587	105	5
FN	3524	661	106	14
Precision	0.025	0.280	0.743	0.930
Recall	0.088	0.256	0.741	0.825
F1	0.039	0.268	0.742	0.874

Table 4: Re-annotation Results by Class.

The massive text false positive count (13,072 vs. 338 true positives) directly reflects missing annotations in the original dataset. Models correctly detected these unlabeled text regions but were penalized as false positives.

Official test results: Our final submission achieves mAP 18.58%, Micro F1 4.29%, Macro F1 20.29%, Average Intersection over Union (IoU) 63.90%. Test performance falls below validation results, indicating potential annotation quality issues in test data.

4.5 Task C: Handwritten Text Recognition

Table 5 presents Task C validation results comparing photometric augmentation and style transfer approaches.

Method	CER	F1	NED	Score
Baseline	0.2881	0.7281	0.2771	0.7190
Xunzi Baseline	0.7209	0.3094	0.7018	0.2920
Aug 1×	0.1766	0.8333	0.1709	0.8275
Aug 2×	0.4657	0.7400	0.2639	0.6364
Style Transfer (3)	0.1838	0.8278	0.1783	0.8212
Style Transfer (6)	0.2920	0.7216	0.2832	0.7199

Table 5: Task C validation set results.

Qwen outperforms Xunzi by 146%. Perfor-mance peaks with 1x augmentation (CER 17.66%), and falls sharply with 2x (CER 46.57%), indicating that excessive transformations break handwriting style constraints.

Style transfer with 3 traditional Chinese styles performs comparably to photometric augmentation (CER 18.38%). Expanding to 6 styles degrades performance to near-baseline levels (CER 29.20%), likely due to accumulated artifacts.

Official test results: Our final submission achieves CER 8.79%, NED 8.73%, F1 91.40%, comprehensive score 91.28% (including character variants; excluding variants: CER 10.16%, score 89.91%). Test performance exceeds validation by 50.2% in relative error reduction, achieving 4.5% improvement over official baseline (CER 9.20%).

4.6 Ablation Studies

Style Transfer Analysis (Task C). Single-style experiments show: Stone Inscription (CER 19.2%), Wood Carving (CER 18.9%), Ink Wash (CER 18.5%), Combined 3 styles (CER 18.38%). Ink wash contributes most to generalization.

Combining photometric augmentation (1×) with individual styles yields: Photometric + Stone (CER 18.8%), Photometric + Wood (CER 18.5%), Photometric + Ink Wash (CER 18.2%). Combining photometric with all 3 styles degrades performance (CER 19.3%), underperforming photometric-only (CER 17.66%) and suggesting paradigm mixing introduces conflicts.

Task B Cross-Validation. 5-fold cross-validation yields mAP 38.2% (±2.3%), consistent with validation performance (40.91%), indicating validation-test gap stems from annotation inconsistencies rather than overfitting.

4.7 Computational Cost Analysis

Photometric augmentation requires less than 0.1 seconds per image using Albumentations on CPU with negligible computational overhead. Style transfer generation via Qwen-Image-Edit API is substantially more expensive: we generated 350 style-transferred images, with each image requiring approximately 40 seconds and costing 0.5 RMB in API tokens, demonstrating significantly higher computational and monetary costs compared to photometric methods.

Training time is approximately 1 hour per task for 3 epochs on 2×NVIDIA A800 GPUs using LoRA fine-tuning, enabling rapid experimental iteration. Inference speed is approximately 20 minutes for the entire test set for both Tasks A and C. These metrics demonstrate that photometric augmentation offers superior cost-effectiveness for large-scale deployment, while style transfer provides specialized augmentation when quality requirements justify the additional computational and monetary costs.

5. Discussion

5.1 Task-Specific Augmentation Dynamics

Task A demonstrates that printed text variations are systematic and well-modeled by photometric transformations, with performance improving monotonically from baseline (CER 27.87%) through 1× augmentation (CER 14.13%) to 2× augmentation (CER 2.55%), achieving 90.9% relative error reduction.

In contrast, Task C exhibits different dynamics, peaking at moderate augmentation (1× CER 17.66%) with severe degradation at 2× (CER 46.57%), indicating that excessive transformations violate calligraphic style constraints. Task B shows minimal gains from photometric augmentation (mAP 39.71% → 40.21%), suggesting that layout analysis is primarily constrained by annotation quality rather than image distribution.

5.2 Annotation Quality as Primary Bottleneck for Task B

Manual re-annotation achieves substantial improvements far exceeding augmentation benefits: Micro F1 increases 205% (3.08% → 9.39%) and Macro F1 increases 258% (13.43% → 48.08%). The validation-test performance gap (mAP 40.91% → 18.58%) does not reflect overfitting, as cross-validation consistency (38.2% ± 2.3%) confirms. Rather, this gap stems from annotation inconsistencies. Our re-annotation added 18,653 text regions (+58.3%), demonstrating systematic under-annotation in the original dataset.

The high false positive count (13,072 FP vs. 338 TP) primarily reflects model overprediction rather than annotation omissions. The model frequently predicts text regions in non-text areas, indicating inadequate spatial boundary discrimination. While a small proportion of false positives represent unlabeled text regions, the dominant issue is spurious detections. The text class precision (0.025) versus book_edge precision (0.743) reveals this 30-fold gap.

Additionally, the model predicts axis-aligned bounding boxes while the original dataset uses orientation-aligned boxes for rotated text. This mismatch introduces systematic penalties for detecting circularly-arranged or diagonally-oriented text in decorative elements and seals.

To address these issues, we recommend: (1) adopting rotated bounding box formats (e.g., RROI) for oriented text elements; (2) establishing inter-annotator agreement protocols (Cohen's Kappa $\geqslant$ 0.85) to ensure consistency standards.

5.3 Style Transfer Trade-offs and Augmentation Paradigm Interactions

Style transfer augmentation (CER 18.38%) performs comparably to but does not exceed photometric augmentation (CER 17.66%). Ablation analysis shows ink wash contributes most (CER 18.5%), followed by wood carving (CER 18.9%) and stone inscription (CER 19.2%).

Combining photometric and style transfer methods produces degradation: photometric alone achieves CER 17.66%, while combined approaches yield CER 18.2-19.3%, indicating paradigm conflicts. Performance degradation stems from two factors: (1) style transfer introduces background distribution shifts, with stone inscription and wood carving creating domain mismatch from original manuscript contexts; (2) 8-12% of generated images require filtering due to character illegibility, and accumulated artifacts degrade performance when expanding to 6 styles (CER 29.20%).

For VLM architectures like Qwen2.5-VL, photometric realism aligns better with test distribution than synthetic stylization, likely because vision-language models encode diverse stylistic knowledge through large-scale pretraining. Computationally, photometric augmentation requires <0.1 seconds per image versus ~40 seconds for style transfer API calls, providing substantially better cost-effectiveness.

5.4 Detailed Failure Case Analysis

Figure 3: The figure presents three failure cases. (Left) Task A shows ground truth "復次橋尸边是兴十方各如碗加沙" versus prediction "復次憍尸迦置此十方各如菟伽沙" with 6 substitution errors. (Middle) Task B compares sparse ground truth annotations (left) with dense model predictions (right), showing extensive overprediction in non-text regions with only minor false positives from unlabeled areas. (Right) Task C shows ground truth "也書之求能且攻真草今以授之可須" versus prediction "也幸而然往日良可悲今令燾之可忍" with 11 substitution errors.

Task A exhibits 142 substitution errors (78.02% of total, CER 2.55%), primarily from visually similar character pairs and rare characters. Strong boundary detection (7.69% missing, 14.29% insertions) suggests fine-grained visual discrimination requires character-level confusion analysis and rare character augmentation.

Task B reveals substantial spatial overprediction. The massive false positive count reflects model-driven confidence misalignment, with the model predicting text in non-text areas. The text precision of 0.025 versus book_edge precision of 0.743 demonstrates fundamental challenges in learning text versus non-text patterns. The axis-aligned versus rotated bounding box mismatch further compounds this gap for decorative and seal elements.

Task C demonstrates severe semantic drift with 1,817 substitution errors (90.04% of total). Semantic incoherence rather than character-level confusion suggests handwritten recognition requires sequence-level contextual constraints to enforce linguistic plausibility.

5.5 Broader Implications and Future Directions

Data-centric optimization achieves competitive performance, though effectiveness depends on task characteristics. Task A benefits from augmentation targeting document degradation. Task C requires calibrated augmentation intensity. Task B reveals annotation quality dominance, with 58.3% annotation increase producing 205-258% metric improvement. Future work should prioritize annotation quality assurance and explore task-adaptive augmentation strategies aligned with modality-specific constraints.

6. Conclusion

This work demonstrates that data-centric strategies achieve competitive performance in closed-track ancient Chinese text recognition. Our 17-strategy augmentation pipeline achieves significant improvements: Task A reduces CER by 27.5% (6.18% → 4.48%), Task C by 4.5% (9.20% → 8.79%), with task-specific optimal intensities (2× for printed texts, 1× for handwritten texts). Style transfer ablation shows ink wash contributes most to generalization, though combining paradigms underperforms single approaches.

Systematic annotation refinement for Task B achieves 205% Micro F1 and 258% Macro F1 improvement, adding 18,653 text annotations (+58.3%). Cross-validation consistency (mAP 38.2% ± 2.3%) suggests validation-test gap stems from incomplete test annotations rather than overfitting, demonstrating annotation quality as the primary bottleneck.

Future work should implement comprehensive re-annotation with strict protocols, explore adaptive augmentation for different modalities, and investigate semi-supervised methods for unlabeled ancient texts.

7. Acknowledgements

We thank the EvaHan 2026 organizers from Nanjing Agricultural University, Nanjing Normal University, and Nanjing University of Science and Technology for creating this evaluation campaign. We acknowledge the developers of Qwen2.5-VL and Albumentations for their tools and frameworks.

8. Bibliographical References

Shi, B., Bai, X., & Yao, C. (2015). An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39, 2298-2304.

Cheng, Z., Bai, F., Xu, Y., Zheng, G., Pu, S., & Zhou, S. (2017). Focusing Attention: Towards Accurate Text Recognition in Natural Images. 2017 IEEE International Conference on Computer Vision (ICCV), 5086-5094.

Li, M., Lv, T., Cui, L., Lu, Y., Florêncio, D.A., Zhang, C., Li, Z., & Wei, F. (2021). TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models. ArXiv, abs/2109.10282.

Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., & Park, S. (2021). OCR-Free Document Understanding Transformer. European Conference on Computer Vision.

Simard, P.Y., Steinkraus, D., & Platt, J.C. (2003). Best practices for convolutional neural networks applied to visual document analysis. Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings., 958-963.

Buslaev, A.V., Parinov, A., Khvedchenya, E., Iglovikov, V.I., & Kalinin, A.A. (2018). Albumentations: fast and flexible image augmentations. ArXiv, abs/1809.06839.

Cubuk, E.D., Zoph, B., Mané, D., Vasudevan, V., & Le, Q.V. (2019). AutoAugment: Learning Augmentation Strategies From Data. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 113-123.

Ren, S., He, K., Girshick, R.B., & Sun, J. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39, 1137-1149.

He, K., Gkioxari, G., Dollár, P., & Girshick, R.B. (2017). Mask R-CNN.

Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Jimmy, K., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C.C., & Lin, D. (2019). MMDetection: Open MMLab Detection Toolbox and Benchmark. ArXiv, abs/1906.07155.

AnandaSky: A Vision–Language Model for Line-Level Transcription of Historical Sinographic Documents

Colin Brisson[1,2], **Ayoub Kahfy**[2], **Frédéric Constant**[1,3], **Marc Bui**[4]
[1]CRCAO, EPHE–PSL; [2]Badiane; [3]ERMES, Université Côte d'Azur; [4]AOROC, EPHE–PSL
{colin, ayoub}@badiane.ai, frederic.constant@univ-cotedazur.fr, marc.bui@ephe.psl.eu

Abstract

We present AnandaSky, a vision–language model for line-level transcription of historical sinographic documents. The model combines a compact high-resolution visual encoder with global attention, 10 px patches, uncompressed visual prefix and a Qwen3-0.6B autoregressive decoder. It is trained at scale on 4M annotated lines from documents produced in China and Korea between the 8th and 20th centuries. Across in-domain and held-out public benchmarks, AnandaSky achieves sub-1% CER on five of eight datasets, sets a new state of the art on MTHv2 with 0.92% CER, and shows strong transfer to unseen collections. For EvaHan 2026, full fine-tuning on the organizers' data to match task-specific annotation conventions reduces CER relative to the official baseline by 5.2% on prints and 12.1% on manuscripts, despite using one-tenth as many parameters.

Keywords: Historical documents, Classical Chinese, OCR, HTR, vision–language models, long-tailed distribution, Dunhuang manuscripts

1. Introduction

The sinographic writing system has been used continuously for more than two millennia across a vast geographic area, giving rise to a written heritage so extensive that, as a Chinese saying goes, it could "make an ox sweat and fill a house up to the ceiling". Although the People's Republic of China was a pioneer in using automatic text recognition (ATR) to broaden access to this heritage[1], publicly available digital corpora remain limited in size, uneven in quality, and often redundant. Large-scale digitization therefore still depends on transcription systems that minimize post-correction cost while generalizing across collections.

Automatic transcription of historical sinographic documents is difficult not only because of the size of the character set and the strong visual similarity between classes (Shen et al., 2023), but also because of three structural properties. First, Chinese characters carry high information density: Shannon-entropy estimates suggest roughly $\sim$1.5 bits per Latin letter versus $\sim$10 bits per Chinese character, making a single character error potentially as consequential as an English word error. Under this view, a CER of 4% in Classical Chinese corresponds roughly to losing one word out of twenty-five in English. Second, character frequencies exhibit an extreme long-tail distribution, with roughly 20% of characters accounting for 95% of occurrences, which makes balanced coverage costly. Third, convention and standardization issues introduce additional brittleness: Unicode assigns distinct code points to glyph variants of the same underlying character (e.g., 黃 U+9EC3 vs. 黃 U+9EC4). Such duplications—numbering in the tens of thousands—complicate transcription conventions, hinder interoperability across corpora, and make evaluation unnecessarily fragile.

To address these challenges, we propose AnandaSky, a line-level transcription model for historical sinographic documents based on a vision–language architecture. The name combines Ananda—the disciple of the Buddha traditionally associated with the "encoding" of early Buddhist texts—and Sky, the opening character of the *Thousand Character Classic*, a text long used in premodern China to enumerate items. AnandaSky is designed to transcribe both historical prints and cursive handwriting.

Our contributions are as follows:

- We adapt the vision–language model paradigm to efficient line-level transcription of historical sinographic documents through a compact high-resolution architecture with global visual attention, 10 px patches, an uncompressed visual prefix, and variable-length attention.
- We introduce AnandaSky, a domain-specialized model trained on 4M line images, together with a systematic glyph normalization pipeline that reduces convention-induced brittleness across heterogeneous corpora.
- We show that this compact specialized model transfers robustly across collections, sets a new state of the art on MTHv2 (Ma et al., 2020), and outperforms much larger general-purpose baselines.

[1]This refers to the *Siku Quanshu* transcription project, launched in 1996, which digitized 3,461 works—about 800 million characters—in three years.

Proceedings of LT4HALA 2026 @ LREC 2026, pages 311–321.
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

To support broader access to sinographic heritage, we release the AnandaSky model weights and inference code on Hugging Face.[2]

2. Related Work

Benchmarking efforts for historical Chinese document recognition include HDRC-CHINESE at ICDAR 2019 (Saini et al., 2019) and the CUHK Chinese OCR challenges in 2021 and 2022 (CUHK Library, 2021, 2022). A recurring pattern in these evaluations is the continued reliance on the classical OCR pipeline: character detection followed by isolated character classification. By contrast, modern handwritten text recognition (HTR) has largely moved toward line-level recognition with monotonic alignment methods such as CTC (Graves et al., 2006; Garrido-Muñoz et al., 2025), which avoid brittle segmentation and enable context-sensitive decoding.

More recently, ATR has increasingly been cast as a conditional generation problem in vision–language models (VLMs). TrOCR combines a pretrained Transformer-based vision encoder with a pretrained language decoder through cross-attention (Li et al., 2021), whereas DTrOCR explores a decoder-only alternative in which patch embeddings are supplied directly as a visual prefix, without a separate vision encoder (Fujitake, 2024). General-purpose VLMs such as Qwen2.5-VL (Bai et al., 2025) follow the now-standard design of a ViT-style visual encoder paired with an autoregressive LLM decoder conditioned on visual tokens, and perform strongly on modern text-in-image tasks. Historical sinographic documents, however, remain substantially out of distribution for such models: on historical Chinese benchmarks, even the best-performing general-purpose system, Gemini 2.5 Pro, achieves only 32.03% CER (Yu et al., 2025). This gap motivates domain-specialized VLMs that combine high-resolution visual encoding, convention-aware normalization, and large-scale historical training data.

3. Architecture

AnandaSky follows the Qwen2.5-VL paradigm, combining (i) a Vision Transformer (ViT) visual encoder and (ii) a pretrained autoregressive decoder. The full model contains 625.98M parameters. A line image is split into patches and encoded into a sequence of visual tokens, which are then projected into the decoder embedding space as a visual prefix that conditions autoregressive transcription generation.

Encoder component	Value
Patch size	10 px
Encoder embedding	768
Hidden size	768
Transformer layers	4
Attention heads	12
FFN size	2048
FFN activation	SwiGLU
Normalization	HybridNorm*

Table 1: Visual encoder configuration of AnandaSky.

Visual encoder. Table 1 summarizes the visual encoder configuration. The encoder is intentionally compact, with only four Transformer layers, hidden size 768, and 12 attention heads. We favor a shallow but high-resolution visual backbone because historical line transcription depends primarily on fine-grained glyph discrimination, while most long-range sequence modeling is handled by the pretrained decoder.

Unlike Qwen2.5-VL, which relies on local visual attention, we use global attention over the full line image. This choice is motivated by the sequential nature of text: visually relevant dependencies may extend across the entire line, for example when disambiguating confusable characters or compensating for missing strokes. We also reduce the patch size to 10 px (vs. 14 px in Qwen2.5-VL), which increases the visual sequence length but preserves crucial micro-details for cursive and degraded handwriting. Within the encoder, feed-forward blocks use SwiGLU MLPs, normalization follows the HybridNorm* strategy (Zhuo et al., 2025), and positional information is encoded with 2D Rotary Positional Encoding.

Visual prefix and decoder. While Qwen2.5-VL compresses visual tokens before injecting them into the decoder, we keep the full token sequence in order to preserve fine spatial information. The autoregressive decoder is initialized from Qwen3-0.6B[3]. It generates the transcription conditioned on the projected visual tokens, thereby integrating a pretrained language model directly into the decoding process. Following the multimodal positional encoding scheme of the Qwen-family models, the decoder uses MRoPE-Interleaved positional encoding (Huang et al., 2025).

Variable-length attention. Line images vary substantially in width and therefore in visual sequence length. To avoid excessive padding, we use variable-length attention in both the visual encoder and the autoregressive decoder. This reduces padding overhead and improves training efficiency.

[2] https://huggingface.co/badianeai/AnandaSky

[3] https://huggingface.co/Qwen/Qwen3-0.6B

4. Dataset

4.1. Main sinographic corpus

AnandaSky is trained on a large corpus comprising 66,607,288 character instances across 4,039,988 annotated lines. The line images come from historical sinographic documents produced between the 8th and the 20th century in China and Korea. The corpus includes 621,017 handwritten lines (about 15%) in cursive or semi-cursive scripts; among them, 337,696 lines come from Dunhuang manuscripts.

4.2. Transcription normalization

Because the transcriptions were produced by different teams using heterogeneous conventions, we apply a systematic normalization pipeline to reduce convention-induced variation. When differences correspond only to minor graphic variants—for example, 晋 U+664B vs. 晉 U+6649, along with associated forms such as 戬 U+622C vs. 戩 U+6229—we use a deterministic one-to-one mapping. For 2,267 characters for which no unique normalization can be defined, we instead perform context-sensitive disambiguation with a bidirectional encoder. A representative case is 后 U+540E, which may correspond either to the traditional form 後 U+5F8C or to a distinct lexical item meaning "queen." After normalization, the effective vocabulary is reduced to 16,770 distinct classes. A more detailed account of this normalization pipeline will be presented in a separate publication.

4.3. Multiscript pretraining corpus

For pretraining, we build a second dataset that adds other historical scripts (Latin manuscripts, French, Japanese, etc.) to the sinographic data. The full pretraining corpus totals roughly 10 million lines. It strengthens robustness to style and degradation and helps stabilize vision–language alignment at scale.

5. Training

Training proceeds in multiple phases to stabilize vision–language alignment while preserving the linguistic capabilities of the pretrained decoder. Throughout training, inputs are kept at near-native resolution: line images are fed without rescaling whenever possible, and only lines whose width exceeds 120 px are resized to this maximum width while preserving aspect ratio.

Phase 1: encoder pretraining. We first pretrain the visual encoder with a shallow autoregressive decoder. This stage adapts the visual backbone to the target document domain and provides a strong visual representation before learning the vision–language interface.

Phase 2: multimodal pretraining. We then train the multimodal model in two steps, following Qwen2.5-VL strategy: (i) we train only the adaptor for about 10% of total iterations to learn a stable projection into the decoder embedding space; (ii) we then jointly optimize the full model. To preserve linguistic representations in the decoder, we use differential learning rates: a higher learning rate for the visual encoder and adaptor, and a lower learning rate for the decoder. This limits catastrophic drift while allowing gradual adaptation to the visual signal.

Phase 3: fine-tuning. Finally, we fine-tune on the sinographic historical data to better capture paleographic specificities and the long-tail character distribution.

Dataset	# lines	CER (%)
Qing legal documents	142	4.89
Dunhuang manuscripts	841	1.38
MTHv2	8,051	0.92
Sibucongkan	1,146	0.43
Korean Anthologies	1,477	0.33

Table 2: Results on in-domain evaluation sets.

Dataset	# lines	CER (%)
ICDAR2019-HDRC	2,560	0.96
CUHK Challenge 2021	3,394	0.82
CUHK Challenge 2022	4,318	1.61

Table 3: Results on held-out benchmarks.

6. Evaluation

6.1. In-domain and external benchmarks

We evaluate AnandaSky on both in-domain test sets and external public benchmarks. The in-domain sets measure generalization to held-out documents from the same corpora as the training data, whereas the public benchmarks provide a stricter cross-dataset evaluation, since no samples from these datasets were used during training.

To account for heterogeneous transcription conventions, we normalize all ground-truth transcriptions using the same normalization pipeline as in training. Decoding is performed with beam search (beam size 4), and performance is reported in terms of character error rate (CER), defined as the Levenshtein distance normalized by the reference length and expressed as a percentage.

Reference	CER (%)
Ma et al. 2020	4.48
Shi et al. 2017	3.06
Huang et al. 2021	2.58
Huang et al. 2023	2.11
AnandaSky (ours)	**0.92**

Table 4: Comparison with previously reported results on MTHv2. Best result in bold.

As shown in Tables 2–3, AnandaSky delivers consistently strong performance on both in-domain and external evaluations. CER falls below 1% on five of the eight datasets, including three in-domain sets and two held-out benchmarks. On MTHv2 (Ma et al., 2020), AnandaSky achieves 0.92% CER (Table 4), improving on the best previously reported result of 2.11% by 1.19 absolute points. Performance on ICDAR2019-HDRC (Saini et al., 2019) and CUHK Challenge 2021 (CUHK Library, 2021) is particularly noteworthy, since these datasets consist of family records and Taoist texts, respectively, document types that are absent from the training distribution. Taken together, these results indicate that the model generalizes robustly across linguistic and documentary domains.

Further error analysis is provided in Appendix A.2. Figure 2 shows that errors are highly concentrated near zero across all in-domain datasets, confirming that most lines are transcribed perfectly or with only minor deviations. The aggregate CER is therefore driven by a relatively small upper tail of difficult cases. To examine this tail while limiting selection bias, Figures 3–7 present stratified random samples from each dataset: panels A and B are drawn from lines above the 90th percentile of line-level CER, while panels C and D are drawn from the 80th–90th percentile band. Manual inspection suggests that many of the highest-CER cases are associated with image degradation (e.g., Figure 3, panels A–B), normalization mismatches (e.g., Figure 4), or apparent inconsistencies in the reference transcriptions (e.g., Figures 6–7), rather than systematic recognition failures.

More generally, these observations highlight that quantitative evaluation is only as reliable as the reference transcriptions on which it is based. In particular, while the reported CER suggests that manuscript transcription remains more difficult than print transcription, the qualitative analysis of the Dunhuang test set (Figure 4) points to a more nuanced picture: a substantial fraction of the apparent errors seem to reflect annotation noise, normalization mismatches, or ambiguous ground truth rather than genuine reading failures. The reported CER for this dataset may therefore overestimate the true transcription error, and actual performance may be closer to the sub-1% regime than the raw score alone suggests.

6.2. EvaHan 2026 shared-task results

EvaHan 2026 is a shared task on multimodal methods for ancient Chinese document understanding, comprising line-level transcription for printed materials (Task A), line-level transcription for manuscripts (Task C), and document segmentation (Task B). In this section, we report our official open-track results for the transcription tasks. Task B is discussed separately in Appendix A.1.

For Tasks A and C, we submitted an open-track system obtained by fully fine-tuning AnandaSky on the official training data. We chose full fine-tuning because EvaHan adopts transcription conventions that differ from our own; without adaptation, AnandaSky would be penalized by systematic convention mismatches rather than genuine recognition errors.

Table 5 reports the official results communicated by the organizing committee for Tasks A and C. Despite using only 626M parameters—more than 10$\times$ fewer than the 7B-parameter Qwen2.5-VL baseline—AnandaSky-ft consistently outperforms the official baseline across all reported metrics, reducing CER by 5.2% on Task A and by 12.1% on Task C. The gains are particularly notable because they are obtained after adapting a much smaller domain-specialized model to the organizers' conventions.

We note, however, that the official EvaHan CER is substantially higher than the CER obtained by AnandaSky on the other benchmarks considered in this paper. Because the EvaHan test set is not publicly available, we can only propose plausible explanations. First, the official evaluation appears to rely on macro averaging, whereas all other scores reported in this paper use micro averaging. Second, the EvaHan test distribution may differ from the development data, and full fine-tuning on the organizers' training set may trade some cross-collection robustness for improved alignment with task-specific annotation conventions.

7. Ablation

We ablate the encoder patch size to quantify the impact of visual granularity on transcription accuracy. For each patch size, we rerun the full training pipeline. To ensure a fair comparison, we adjust the batch size so that each configuration is trained on the same total number of images, while keeping the remaining hyperparameters unchanged.

Overall performance improves only moderately as the patch size decreases (14 px: CER=1.96%;

Task	System	CER	NED	F1	Overall
A	Qwen2.5-VL 7B (baseline)	0.0618	0.0613	0.9430	0.9397
A	**AnandaSky-ft**	**0.0586**	**0.0564**	**0.9479**	**0.9438**
$A_{\text{no-var}}$	Qwen2.5-VL 7B (baseline)	0.0685	0.0679	0.9364	0.9331
$A_{\text{no-var}}$	**AnandaSky-ft**	**0.0639**	**0.0617**	**0.9426**	**0.9385**
C	Qwen2.5-VL 7B (baseline)	0.0920	0.0919	0.9099	0.9086
C	**AnandaSky-ft**	**0.0809**	**0.0809**	**0.9200**	**0.9194**
$C_{\text{no-var}}$	Qwen2.5-VL 7B (baseline)	0.1066	0.1065	0.8953	0.8940
$C_{\text{no-var}}$	**AnandaSky-ft**	**0.1007**	**0.1007**	**0.9002**	**0.8996**

Table 5: Official EvaHan 2026 open-track results for Tasks A and C, comparing the official Qwen2.5-VL 7B baseline (instruction-tuned) with AnandaSky-ft, i.e., AnandaSky fully fine-tuned on the EvaHan training data. Rows tagged $_{\text{no-var}}$ correspond to the stricter *excluding-variants* evaluation; untagged rows use the *including-variants* setting. Best results per task/setting are shown in **bold**.

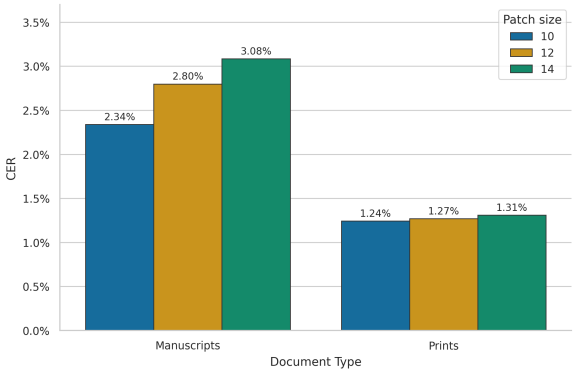


Figure 1: CER by document type as a function of the encoder patch size.

12 px: CER=1.82%; 10 px: CER=1.63%). However, stratifying results by document type reveals a clear and consistent trend. As displayed in Figure 1, finer patches yield substantially larger gains on manuscripts than on prints. On manuscripts, reducing the patch size from 14 px to 10 px lowers CER from 3.08% to 2.34%, highlighting the importance of capturing fine-grained stroke details for cursive and degraded handwriting. In contrast, the effect on printed documents is smaller, with CER improving from 1.31% (14 px) to 1.24% (10 px). These results support our choice of a 10 px patch size: the increased visual sequence length primarily benefits the most challenging regime while leaving printed-text performance largely unchanged.

8. Conclusion

We introduced AnandaSky, a vision–language model for line-level transcription of historical sinographic documents. By adapting the VLM paradigm to this setting—through a compact high-resolution visual encoder with global attention, 10 px patches, an uncompressed visual prefix, and efficient variable-length attention—and by training at scale with convention-aware transcription normalization, we obtain strong results on both in-domain and held-out benchmarks. AnandaSky sets a new state of the art on MTHv2, reaches sub-1% CER on five of eight evaluation datasets, and transfers robustly across unseen collections.

Beyond benchmark performance, our results suggest that the field is approaching a regime in which quantitative evaluation alone is no longer sufficient. As error rates become very low, aggregate CER increasingly reflects not only model mistakes but also annotation noise, normalization mismatches, and inconsistencies in reference transcriptions. In particular, while the reported scores remain somewhat higher on manuscripts than on prints, the qualitative analysis of the Dunhuang test set suggests that this gap is smaller than the raw CER indicates, and that actual manuscript transcription performance may be closer to the low-error regime reached on printed materials.

From a practical perspective, this means that large-scale transcription of historical sinographic corpora can increasingly shift from labor-intensive manual correction toward lightweight verification. More broadly, our findings show that compact domain-specialized VLMs can be highly effective for historical document transcription, and provide a strong foundation for future systems that integrate transcription with layout analysis, retrieval, and richer document understanding.

9. Limitations

Our work has several limitations. First, the competition setting constrained the amount of exploration we could perform in terms of scale. We did not systematically study the effect of longer training schedules, larger visual backbones, or larger autoregressive decoders within the EvaHan timeline. Since both VLM performance and robustness often improve with compute and model capacity, a

more exhaustive scaling study (parameters, data, and training duration) is left for future work.

Second, while AnandaSky generalizes well across the printed and manuscript benchmarks considered in this paper, its robustness is not uniform across all sinographic document domains. In our experiments, we observed strong performance on Qing dynasty legal documents, but noticeably weaker results on similar legal materials produced in Indonesia (Dean et al., 2025). This suggests that domain shifts in writing conventions or language can still degrade performance. Addressing such cross-region and cross-collection shifts will likely require additional targeted data, improved normalization of conventions, or explicit domain adaptation strategies.

Finally, our model operates at the line level and therefore assumes that reliable line extraction is available. Deploying AnandaSky in end-to-end digitization pipelines requires upstream layout analysis and line detection; errors in these components can propagate to transcription quality. Integrating robust layout modeling and reading-order prediction with line-level transcription remains an important direction for future work.

10. Acknowledgements

We thank the organizers of the 2021 and 2022 Chinese OCR Challenges hosted by the CUHK Library for granting access to the competition data. This work was supported by the French National Research Agency (ANR) under the Investissements d'avenir programme (ANR-21-ESRE-0005, EquipEx Biblissima+), and by the French government through the France 2030 investment plan managed by ANR, as part of the Initiative of Excellence Université Côte d'Azur (ANR-15-IDEX-01). This work also benefited from access to the HPC resources of IDRIS under GENCI allocations 2026-AD011014973R2 and 2026-AD011017524R1. This research was also funded in part by ANR under project ANR-24-CE27-4500-03.

11. Bibliographical References

Shuai Bai et al. 2025. Qwen2.5-VL technical report. arXiv preprint.

CUHK Library. 2021. 2021 chinese classic text OCR challenge. Webpage. Accessed 2026-02-28.

CUHK Library. 2022. Chinese classic text OCR challenge 2022. Webpage. Accessed 2026-02-28.

Kenneth Dean, Maria De Iorio, Alexander Mozdzen, Zhang Qiao, Yu Kang, and Andrew Harris. 2025. Unveiling country trade: Insights from 19th century batavia Kong Koan records via OCR and NLP analysis. Webpage. Accessed 2026-02-28.

Masato Fujitake. 2024. DTrOCR: Decoder-only transformer for optical character recognition. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 8010–8020.

Carlos Garrido-Muñoz, Antonio Rios-Vila, and Jorge Calvo-Zaragoza. 2025. Handwritten text recognition: A survey. arXiv preprint.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Jiarong Huang, Dezhi Peng, Hongliang Li, Hao Ni, and Lianwen Jin. 2023. SegCTC: Offline handwritten chinese text recognition via better fusion between explicit and implicit segmentation. In *Document Analysis and Recognition – ICDAR 2023*, pages 332–349, Cham. Springer Nature Switzerland.

Jie Huang, Xuejing Liu, Sibo Song, Ruibing Hou, Hong Chang, Junyang Lin, and Shuai Bai. 2025. Revisiting multimodal positional encoding in vision-language models.

Yuhao Huang, Lianwen Jin, and Dezhi Peng. 2021. Zero-shot chinese text recognition via matching class embedding. In *Document Analysis and Recognition – ICDAR 2021*, pages 127–141, Cham. Springer International Publishing.

Minghao Li et al. 2021. TrOCR: Transformer-based optical character recognition with pre-trained models. arXiv preprint.

Weihong Ma, Hesuo Zhang, Lianwen Jin, Sihang Wu, Jiapeng Wang, and Yongpan Wang. 2020. Joint layout analysis, character detection and recognition for historical document digitization. In *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 31–36, Los Alamitos, CA, USA. IEEE Computer Society.

Rajkumar Saini, Derek Dobson, Jon Morrey, Marcus Liwicki, and Foteini Simistira Liwicki. 2019. ICDAR 2019 historical document reading challenge on large structured chinese family records. In *Proceedings of the International Conference*

on Document Analysis and Recognition (ICDAR), pages 1499–1504.

Lu Shen, Bidong Chen, Jianjing Wei, Hui Xu, Su-Kit Tang, and Silvia Mirri. 2023. The challenges of recognizing offline handwritten chinese: A technical review. *Applied Sciences*, 13(6):3500.

Baoguang Shi, Xiang Bai, and Cong Yao. 2017. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2298–2304.

Haiyang Yu, Yuchuan Wu, Fan Shi, Lei Liao, Jinghui Lu, Xiaodong Ge, Han Wang, Minghan Zhuo, Xuecheng Wu, Xiang Fei, Hao Feng, Guozhi Tang, An-Lan Wang, Hanshen Zhu, Yangfan He, Quanhuan Liang, Liyuan Meng, Chao Feng, Can Huang, Jingqun Tang, and Bin Li. 2025. Benchmarking vision-language models on chinese ancient documents: From OCR to knowledge reasoning.

Zhijian Zhuo, Yutao Zeng, Ya Wang, Sijun Zhang, Jian Yang, Xiaoqing Li, Xun Zhou, and Jinwen Ma. 2025. Hybridnorm: Towards stable and efficient transformer training via hybrid normalization.

A. Appendices

A.1. EvaHan 2026 Task B Results

For this track, we used a YOLO26 model with an input resolution of 960 px. The official Task B results are reported in Table 6. Our system substantially outperforms the official Qwen2.5-VL 7B baseline, improving mAP@[.5:.95] from 0.2006 to 0.4389, Micro F1 from 0.0513 to 0.7232, Macro F1 from 0.1530 to 0.6657, and Avg Match IoU from 0.6600 to 0.8114. We nevertheless interpret the absolute Task B scores with caution. Manual inspection of the available annotations suggests that this task is affected by non-negligible annotation noise and matching inconsistencies, likely to a greater extent than Tasks A and C. We therefore regard the relative improvement over the official baseline as more informative than the absolute score level.

Task	System	mAP@[.5:.95]	Micro F1	Macro F1	Avg Match IoU
B	Qwen2.5-VL 7B (baseline)	0.2006	0.0513	0.1530	0.6600
B	**Ours**	**0.4389**	**0.7232**	**0.6657**	**0.8114**

Table 6: Official EvaHan 2026 open-track results for Task B, comparing the official Qwen2.5-VL 7B baseline (instruction-tuned) with our YOLO26-based system. Best results are shown in **bold**.

A.2. Qualitative error analysis on in-domain datasets

Figure 2 shows the empirical cumulative distribution of line-level CER on the in-domain evaluation sets. In all datasets, most lines cluster near zero CER, indicating that errors are concentrated in a relatively small upper tail of difficult cases. To examine this tail, Figures 3–7 present stratified random samples from each dataset: panels A and B are drawn from lines above the 90th percentile of line-level CER, while panels C and D are drawn from the 80th–90th percentile band. In each panel, the text shown on the right is the transcription generated by the model during evaluation, and red characters indicate mismatches between the generated transcription and the ground truth.

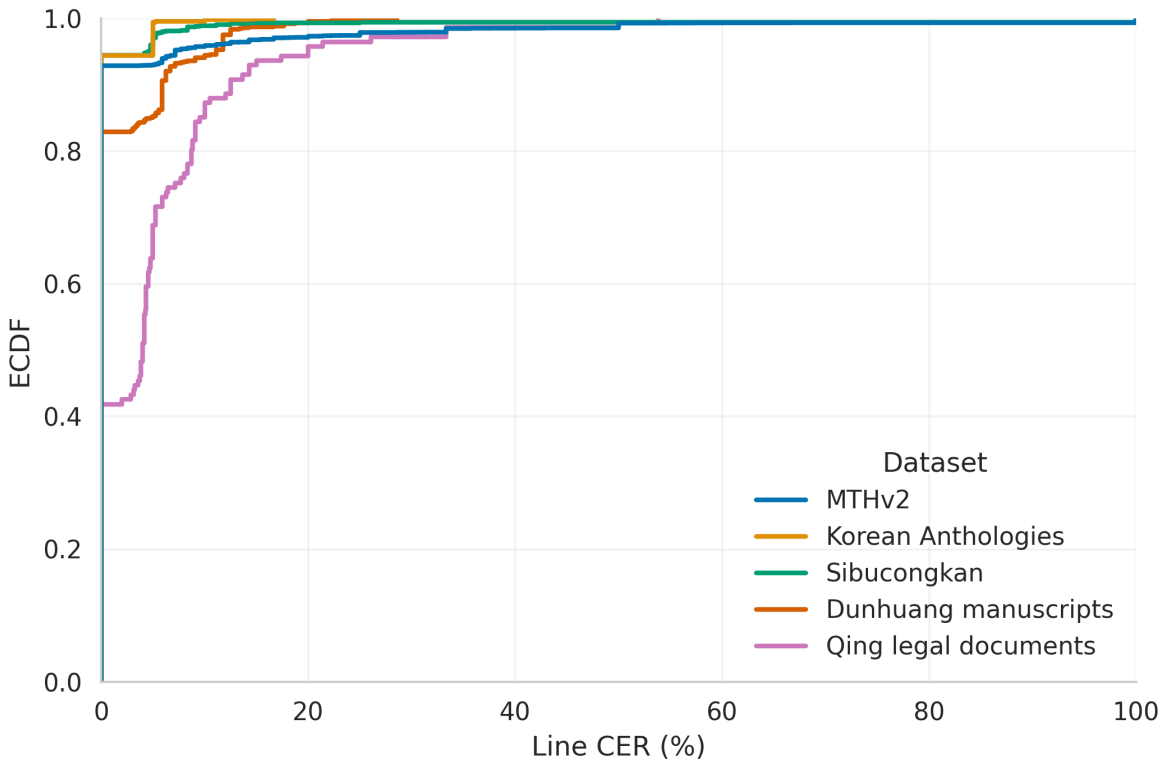


Figure 2: ECDF of line-level CER on the in-domain evaluation sets. Most lines have zero or very low CER, with aggregate error driven by a relatively small upper tail of difficult examples.

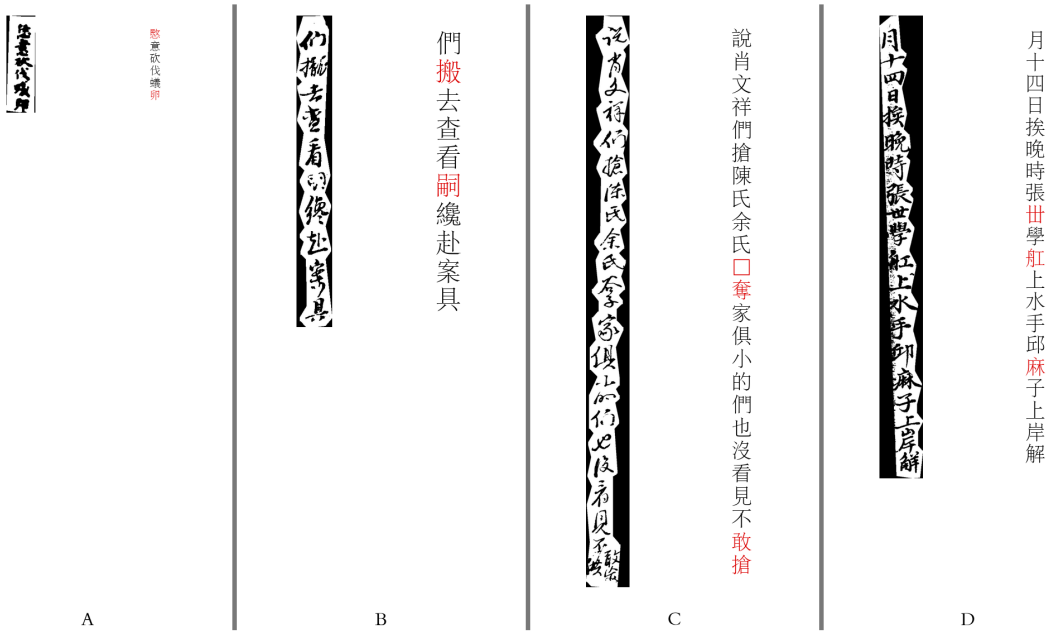


Figure 3: Upper-tail error examples from the Qing legal documents dataset.

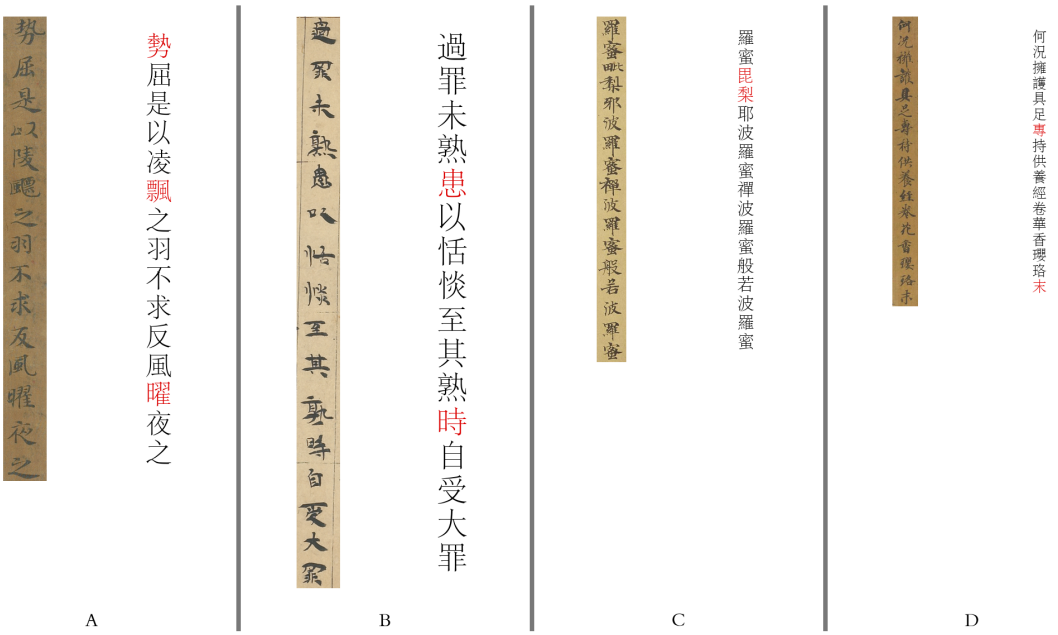


Figure 4: Upper-tail error examples from the Dunhuang manuscripts dataset.

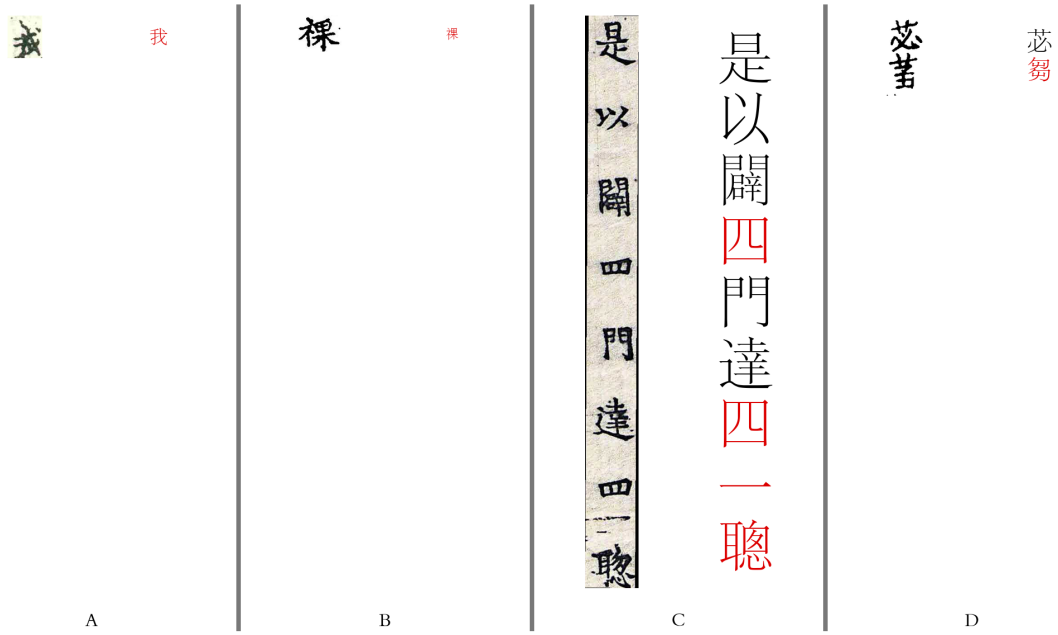


Figure 5: Upper-tail error examples from the MTHv2 dataset.

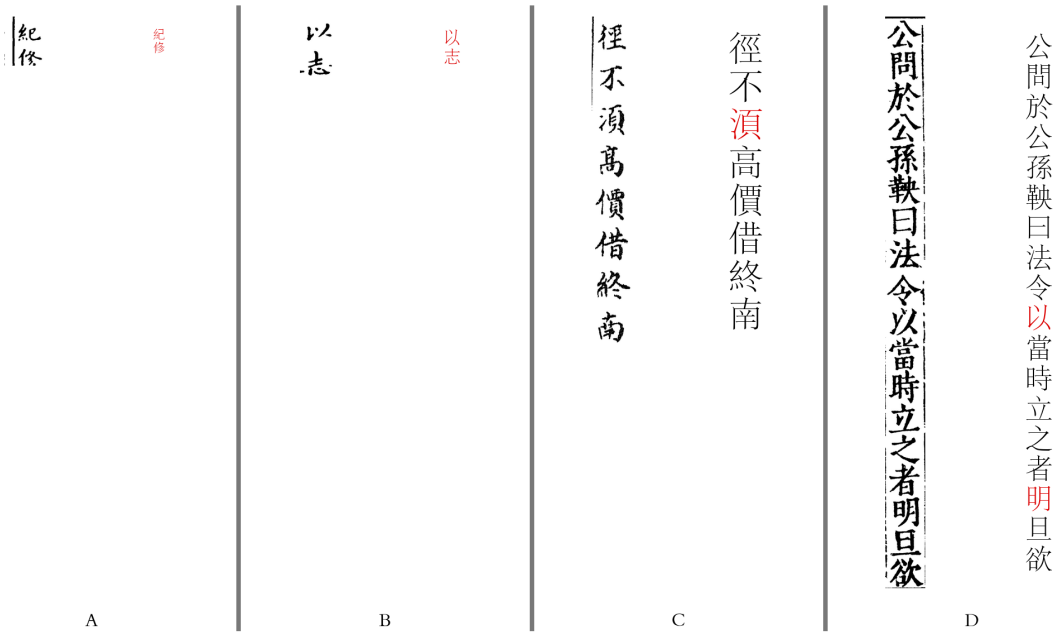


Figure 6: Upper-tail error examples from the *Sibu congkan* dataset.

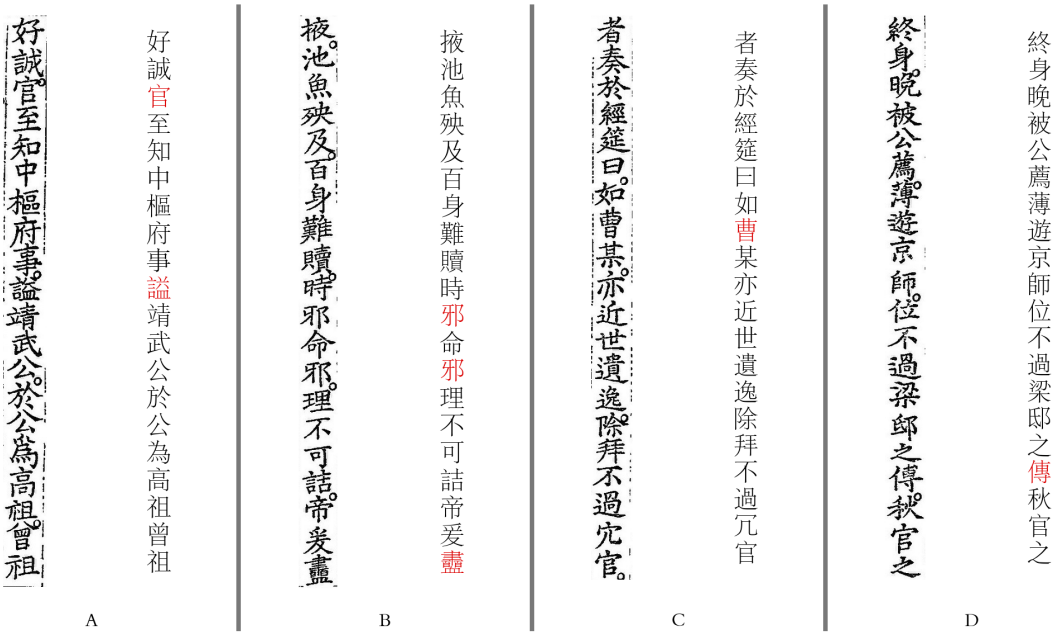


Figure 7: Upper-tail error examples from the Korean Anthologies dataset.

Multimodal Ancient Document Parsing: Technical Report for EvaHan2026 Competition

Liqi He[1], **Qiwei Li**[2], **Ziye Yang**[3], **Zuchao Li**[2]
[1]School of Computer Science, Wuhan University, Wuhan, 430072, China,
[2]School of Artificial Intelligence, Wuhan University, Wuhan, 430072, China,
[3] College of Information and Intelligence, Hunan Agricultural University
heliqi@whu.edu.cn, qw-line@whu.edu.cn, yzy_0607@outlook.com, zcli-charlie@whu.edu.cn

Abstract

We present the multimodal Optical Character Recognition (OCR) and layout analysis methods developed for the EvaHan 2026 competition. Our approach is built upon the Qwen2.5-VL-7B-Instruct architecture and integrates two core strategies: (1) a reinforcement learning alignment pipeline utilizing Direct Preference Optimization (DPO) and Group Relative Policy Optimization (GRPO) to explicitly mitigate hallucination and coordinate instability; and (2) a four-stage curriculum learning framework that synthesizes domain-specific historical artifacts to enhance open-modality generalization. Using this approach, we achieve competitive results, notably reaching a Character Error Rate (CER) of 0.0303 on printed texts (Task A) and 0.0552 on handwritten manuscripts (Task C), as well as an Average Intersection over Union (IoU) of 0.7638 on layout element analysis (Task B).

Keywords: Ancient Chinese OCR, Vision-Language Models, Document Layout Analysis

1. Introduction

The digitization of ancient Chinese literature is a critical endeavor for the preservation and dissemination of cultural heritage. Vast volumes of handwritten and archaic printed historical documents remain fragile and inaccessible to modern computational analysis. Optical Character Recognition (OCR) technology serves as the foundational bridge to transform these imaged texts into structured, editable data. To advance this field, the Fifth International Evaluation of Ancient Chinese Information Processing (EvaHan 2026) has introduced challenging multimodal benchmarks focusing on printed text recognition (Task A), layout element analysis (Task B), and handwritten character recognition (Task C).

Despite the recent success of Vision-Language Models (VLMs) in general-domain document AI, applying these models directly to ancient Chinese texts presents significant challenges. Printed historical texts (Task A) frequently contain rare variant characters, complex woodblock layouts, and physical degradation. Handwritten manuscripts (Task C) introduce even greater variability through cursive connections, stroke omissions, and highly personalized calligraphic styles. Furthermore, for layout analysis (Task B), VLMs often struggle with fine-grained spatial perception, leading to coordinate hallucinations when detecting overlapping elements such as seals superimposed on text blocks.

To address these challenges, we present a robust, unified multimodal framework built upon the well-established Qwen2.5-VL-7B-Instruct architecture. Since supervised fine-tuning (SFT) struggles with VLM hallucinations in dense transcription and coordinate prediction, we propose an reinforcement learning (RL)-aligned and data-scaled training pipeline.

For the **closed modality**, where external data is strictly prohibited, we introduce a progressive preference alignment pipeline. We leverage Direct Preference Optimization (DPO) to penalize the prediction of visually similar but incorrect variant characters (hard negatives). Most notably, we introduce the application of Group Relative Policy Optimization (GRPO) in ancient OCR. By designing customized reward functions targeting character-length consistency and geometric Intersection-over-Union (IoU), our GRPO strategy effectively mitigates the coordinate hallucination problem in layout analysis and reduces character omission rates.

For the **open modality**, we propose a **four-stage curriculum learning** pipeline to systematically build the model's visual-semantic representations. By progressing from single-character recognition using massive multi-font libraries to domain-specific synthesized context learning (e.g., simulating woodblock prints and cursive stroke erosion), we provide a smooth learning trajectory that maximizes the model's capacity to filter out historical noise before integrating real-world data.

Our experiments demonstrate the superiority of our proposed framework. In the official EvaHan 2026 evaluation, our system achieved competitive results across all tasks, notably reaching a Character Error Rate (CER) of 0.0303 in Task A and 0.0552 in Task C, while our RL-driven spatial standardization elevated the layout analysis performance in Task B.

Proceedings of LT4HALA 2026 @ LREC 2026, pages 322–329.
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

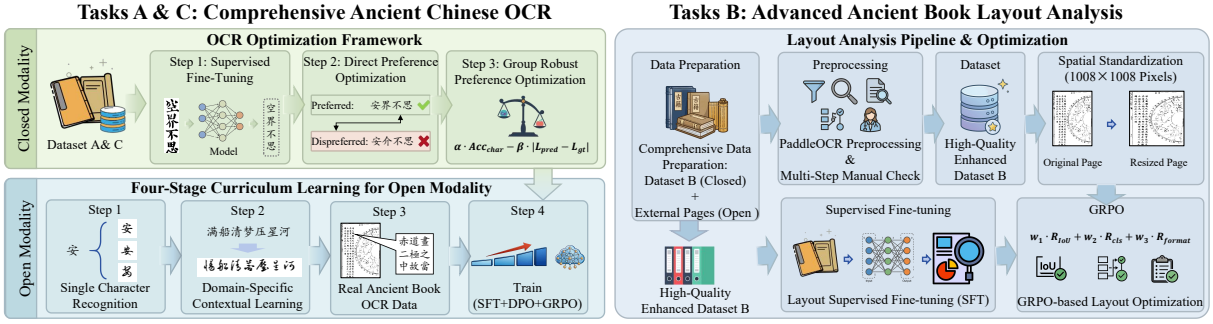


Figure 1: Overview of our methodology.

2. Related Work

2.1. Ancient Chinese Text Processing and Digitization

The digitization of ancient Chinese texts faces unique challenges due to complex layout structures and degraded characters (Li et al., 2024a). Recent advancements have shifted toward domain-specific foundation models, such as GujiBERT (Wang et al., 2023), SongSong (Hu et al., 2025) and the Xunzi large language model series (Zhu et al., 2025), to capture the contextual semantics of classical Chinese. However, end-to-end multimodal digitization for handwritten and woodblock texts remains challenging due to stroke overlapping and ink diffusion.

2.2. Document AI and Vision-Language Models

Early Document AI systems typically relied on complex cascaded pipelines, requiring separate, specialized models for text detection (Du et al., 2020), layout parsing (Huang et al., 2022), and key information extraction (Li et al., 2024b), which often led to error propagation across stages. In contrast, modern VLMs like Donut (Kim et al., 2022) and Nougat (Blecher et al., 2023) have significantly advanced the field by treating document understanding as an end-to-end image-to-text generation task, eliminating the need for separate OCR components. Furthermore, recent large-scale models such as LLaVA (Liu et al., 2023) and the Qwen-VL series (Bai et al., 2023, 2025) have achieved substantial improvements in performance by integrating high-resolution visual encoders with powerful language decoders, enabling the fine-grained perception of both textual content and spatial layout structures.

2.3. Preference Alignment and Reinforcement Learning

Standard SFT adapts VLMs well but suffers from exposure bias and hallucinations, like character omission or coordinate fabrication. To align model outputs with human preferences, reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022) has become a standard protocol. DPO (Rafailov et al., 2023) simplifies this by optimizing the policy model directly on preference pairs. Furthermore, GRPO (Shao et al., 2024) and curriculum learning pipelines (Bengio et al., 2009) have emerged to efficiently optimize discrete sequence generation and mitigate hallucinations in multimodal layout analysis.

3. Methodology

In this section, we detail our technical approach for the EvaHan 2026 competition, focusing on ancient Chinese OCR (Tasks A & C) and layout element analysis (Task B). Our solution leverages the well-established Qwen2.5-VL-7B-Instruct as the foundational VLM for all tasks. This 7B-parameter scale was selected to adhere to the competition's uniform requirements, ensuring a standardized and fair baseline for evaluating inference latency and practical deployment overhead among all participants.

3.1. Tasks A & C

Tasks A and C focus on printed and handwritten ancient Chinese text recognition, respectively. The primary challenges involve identifying rare/variant characters and handling stroke blurring in manuscripts.

3.1.1. Closed Modality

Restricted to official training data, our closed modality strategy refines precision via post-training alignment:

SFT: The model is fine-tuned on the official Dataset A and C.

DPO: We extract hard negative samples where the baseline model misidentified characters. Using

DPO, we train the model to distinguish between the correct transcription (chosen) and the common errors (rejected).

GRPO: To address the common VLM issue of character omission or hallucination, we employ GRPO. We design a reward function that balances character accuracy and sequence length:

$$R = \alpha \cdot Acc_{char} - \beta \cdot |L_{pred} - L_{gt}| \quad (1)$$

where Acc_{char} denotes the accuracy of character recognition, L_{pred} and L_{gt} denote the predicted and ground truth lengths, respectively.

3.1.2. Open Modality

The open modality allows the use of external resources. We implement a **four-stage curriculum learning** pipeline:

Stage 1 (Single Character Recognition): We utilize a large-scale library of traditional Chinese characters, rendered with diverse ancient-style fonts (e.g., Cao, Xing, and Kai styles) to build robust visual feature extraction for individual glyphs.

Stage 2 (Domain-Specific Contextual Learning): We synthesize long-context data using ancient Chinese corpora. For Task A, we use Kai-style fonts to simulate woodblock printing; for Task C, we use cursive (Cao/Xing) fonts with stroke erosion and overlapping to simulate manual handwriting.

Stage 3 (Real OCR Data): We integrate authentic OCR datasets sourced from the internet and open-source historical archives to bridge the gap between synthetic distributions and real-world document characteristics.

Stage 4 (Supervised Fine-tuning & Reinforcement learning): The model is fine-tuned on the official Dataset A and C. We apply the same DPO and GRPO strategies used in the closed modality to achieve final convergence.

3.2. Task B

Task B focuses on the identification of four critical layout elements: *text*, *image*, *book_edge*, and *seal*. To address the challenges of domain adaptation and overlapping categories, we implement a unified pipeline encompassing annotation enhancement, spatial standardization, and structural reinforcement.

3.2.1. Data Preparation

The core distinction between our closed and open modality solutions lies in the data acquisition and annotation scale.

Closed Modality: We focus exclusively on the official Dataset B. To improve label quality, we perform *annotation enhancement* by utilizing PaddleOCR for preliminary detection followed by a manual audit. This process emphasizes the correction of fine-grained boundaries and the resolution of label conflicts in overlapping regions, such as seals superimposed on text blocks.

Open Modality: In addition to the enhanced official dataset, we introduce supplementary ancient book pages gathered from external digital archives. These additional samples undergo the same "PaddleOCR + manual review" pipeline to ensure high-quality regions annotations.

3.2.2. Spatial Standardization

To establish a stable coordinate mapping system across heterogeneous document sizes, all input images are resized to a fixed resolution of 1008x1008 pixels. This standardized square resolution not only ensures optimal alignment with the vision-language backbone's patch processing mechanism, but also provides pixel-level granularity for detecting small elements like seals and thin book edges, effectively balancing fine-grained spatial perception with computational efficiency.

3.2.3. SFT

Using the processed annotations, we conduct supervised fine-tuning on the Qwen2.5-VL-7B-Instruct model. The training objective is to map visual features to structured layout descriptions, enabling the model to master basic category classification and bounding box localization within a normalized relative coordinate system, where all spatial positions are scaled to a standardized range of $[0, 1000]$.

3.2.4. Reinforcement Learning with Structural Rewards

To further refine the geometric precision and format stability, we employ GRPO. We define a multi-dimensional reward function to supervise the model's output:

Geometric IoU Reward: Rewards are calculated based on the IoU between predicted bounding boxes and ground truth, directly optimizing pixel-level localization accuracy.

Table 1: Official Results for Task A (Printed) and Task C (Handwritten). Variants: "Yes" indicates considering variant characters, "No" indicates ignoring them.

Task	Method	Consider Variant (Yes)				Consider Variant (No)			
		CER ↓	NED ↑	F1 ↑	Score ↑	CER ↓	NED ↑	F1 ↑	Score ↑
Task A	Base Model	0.1014	0.0947	0.9110	0.9037	0.1121	0.1054	0.9007	0.8931
	Instruction Tuning	0.0618	0.0613	0.9430	0.9397	0.0685	0.0679	0.9364	0.9331
	Closed (Ours)	0.0316	0.0314	0.9702	0.9690	0.0316	0.0314	0.9702	0.9690
	Open (Ours)	**0.0303**	**0.0301**	**0.9715**	**0.9703**	**0.0303**	**0.0301**	**0.9715**	**0.9703**
Task C	Base Model	0.1207	0.1193	0.8849	0.8812	0.1338	0.1324	0.8718	0.8681
	Instruction Tuning	0.0920	0.0919	0.9099	0.9086	0.1066	0.1065	0.8953	0.8940
	Closed (Ours)	0.0751	0.0750	0.9267	0.9255	0.0902	0.0901	0.9116	0.9104
	Open (Ours)	**0.0552**	**0.0549**	**0.9471**	**0.9456**	**0.0703**	**0.0700**	**0.9320**	**0.9305**

Classification Reward: Ensures the predicted labels correctly identify the element type (e.g., distinguishing between a text block and an image).

Format Consistency Reward: To mitigate the common "hallucination" of coordinates in VLMs, we impose penalties on outputs that violate the structured format (e.g., `<LABEL>[x1,y1,x2,y2]`), ensuring the results are programmatically robust.

4. Experiments

4.1. Experimental Settings

Model training was distributed across two NVIDIA A40 GPUs (48GB VRAM) using LoRA fine-tuning ($r = 64$, $\alpha = 128$). We set the per-device training batch size to 4 with 16 gradient accumulation steps. For reproducibility, models for Tasks A and C were trained for 1 epoch, while the Task B model was trained for 2 epochs. All inference was conducted on a single A40 GPU with the temperature set to 0 (for determinism) and a maximum context length of 4096 tokens to support high-resolution image patches and long sequence transcriptions.

4.2. Main Results on EvaHan 2026 Official Test Sets

The main results reflect our performance on the official EvaHan 2026 evaluation system. We report results for both the **closed modality** (strictly adhering to official models and data) and the **open modality** (leveraging our four-stage curriculum learning and external synthetic/real-world data).

OCR Tasks (Tasks A & C): Table 1 summarizes the performance on Tasks A and C. In Task A (printed texts), our system achieved a comprehensive score of 0.9703, outperforming the instruction-tuned baseline. In Task C (handwritten texts), the advantage of our **open modality** is even more pronounced, reaching a CER of 0.0552, which is a 40% relative improvement over the instruction-tuned baseline (0.0920).

Table 2: Official Evaluation of Task B.

Method	mAP@.5	Micro F1	Macro F1	Avg IoU
Base Model	0.0000	0.0000	0.0000	0.0000
Instruction Tuning	0.2006	0.0513	0.1530	0.6600
Closed (Ours)	**0.5941**	**0.8342**	**0.8686**	**0.7638**
Open (Ours)	0.5818	0.8278	0.8645	0.7572

Layout Element Analysis (Task B): As shown in Table 2, the base VLM fails on coordinate-based layout tasks (mAP@.5 of 0). However, our SFT and subsequent GRPO refinement for spatial standardization enable the model to reach a Micro F1 of 0.8342. Based on our analysis, the slight performance drop of the open modality on Task B can be attributed to annotation discrepancies. Upon inspecting the official training data, we identified inconsistencies and errors in the layout labels. To mitigate this, we performed additional manual re-annotation and introduced self-labeled data in the open modality setting. A detailed visual analysis of this granularity mismatch is provided in Appendix B. While this data provides richer spatial diversity, its annotation logic partially diverges from the official evaluation standards, leading to a marginal performance gap compared to the closed modality, which strictly adheres to the original official annotations. Beyond manual re-annotation, future mitigation strategies will involve applying rule-based post-processing filters to merge fine-grained detections into official macro-level blocks, or adopting a multi-granularity evaluation metric.

4.3. Ablation Study

To evaluate the contributions of our proposed alignment strategies (DPO, GRPO) and curriculum data generation, we conducted an ablation study on a local validation split. Because the official test sets were strictly withheld, we partitioned the official training datasets (A, B, and C). For each task, we randomly sampled 500 image-text pairs to form a

Table 3: Ablation results on Task A and Task C.

Method	Field	Task A		Task C	
		CER ↓	F-Score ↑	CER ↓	F-Score ↑
Base Model	-	0.1420	0.8741	0.3850	0.7253
SFT (Direct Tuning)	Closed	0.0180	0.9837	0.1119	0.8912
+ DPO	Closed	0.0176	0.9840	0.0907	0.9122
+ **GRPO**	Closed	0.0169	0.9848	0.0898	0.9128
+ Random Synthetic Data	Open	0.0177	0.9841	0.1063	0.8963
+ Ancient Books Syn.	Open	0.0172	0.9850	0.0924	0.9106
+ **Real-world Full**	Open	**0.0161**	**0.9855**	**0.0770**	**0.9246**

Table 4: Ablation results on Task B.

Method	Field	mAP@.5 ↑	Micro F1 ↑	Macro F1 ↑
SFT (Direct Tuning)	Closed	0.5234	0.7140	0.6968
+ GRPO	Closed	0.7668	0.8935	0.9357
+ External Data	Open	**0.8028**	**0.9788**	**0.9842**

held-out local test set, preserving the original distribution of document layouts, font variants, and degradation artifacts.

All baseline and intermediate models were trained on the remaining data and evaluated exclusively on this 500-sample set. This localized setup ensures a fair comparison to accurately isolate the performance gains achieved by transitioning from standard SFT to advanced alignment techniques in the closed modality, as well as the incremental benefits of integrating domain-specific external resources in the open modality.

Efficacy of RL Strategies (Cumulative Alignment): The transition from the base model to our final closed modality submission demonstrates a clear cumulative benefit from our alignment pipeline. As shown in Table 3, standard SFT provides the initial domain adaptation, reducing the CER (e.g., from 0.3850 to 0.1119 in Task C). Building upon this SFT foundation, the introduction of DPO teaches the model to reject visually similar hard negative characters, further decreasing the CER to 0.0907. Finally, integrating GRPO on top of the SFT+DPO checkpoint yields the best closed-domain performance (CER 0.0898). The impact of GRPO is more striking in Task B (Table 4), where applying our multi-dimensional structural rewards on top of the SFT model causes the mAP@.5 to increase from 0.5234 to 0.7668. This improvement demonstrates that geometric and format-consistency rewards can effectively mitigate the inherent coordinate hallucination issues in standard VLMs.

Impact of Domain-Specific Data (Open Modality Escalation): Our open modality experiments incrementally introduced external data on top of the established RL framework, revealing critical insights into data scaling for ancient texts. Interestingly, when we naively added **random synthetic data**, performance actually degraded compared to our best closed-domain GRPO model (e.g., in Task C, CER worsened from 0.0898 to 0.1063). This indicates that when dealing with diverse and variable handwritten styles such as calligraphy, the randomness introduced by generic synthetic characters disrupts the contextual comprehension that the model heavily relies on for accurate judgment. However, when we transitioned to our stage 2 **ancient books synthetic data**—which accurately simulates woodblock printing and cursive stroke erosion within coherent textual contexts—the model successfully integrated the external knowledge, recovering and improving its robustness (task C CER dropping to 0.0924). Ultimately, combining the full curriculum with high-quality **real-world data** advanced the system to its peak performance across all tasks (task A CER 0.0161, task C CER 0.0770). This trajectory confirms that while synthetic curriculum learning bridges the initial glyph recognition gap, real-world domain diversity remains crucial for handling complex historical noise.

5. Future Work

In future work, we plan to explore end-to-end multi-task architectures capable of jointly performing layout element analysis and character transcription in a single forward pass. Additionally, we aim to leverage advanced generative models (e.g., diffusion models) to synthesize more realistic historical degradation artifacts (such as ink bleeding and paper aging), further minimizing the domain gap in our curriculum learning pipeline.

6. Conclusion

We presented a comprehensive solution for EvaHan 2026 using Qwen2.5-VL-7B-Instruct. Our RL alignment pipeline (DPO and GRPO) effectively mitigates VLM hallucinations and coordinate instability. Coupled with a four-stage curriculum learning strategy, our system demonstrated robust performance across printed texts, handwritten manuscripts, and complex layouts, achieving competitive results in both the closed and open modalities.

7. Limitations

Despite strong performance, our RL phases (especially GRPO) are computationally expensive and hyperparameter-sensitive (Appendix A). Additionally, Task B's open modality annotation enhancement relies on manual auditing, restricting scalability on massive unlabeled corpora. Appendix C details remaining challenges with rare variant characters, including visual confusion and modern substitution.

8. Bibliographical References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization. *Text Reading, and Beyond*, 2(1):1.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. 2023. Nougat: Neural optical understanding for academic documents. *arXiv preprint arXiv:2308.13418*.

Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, et al. 2020. Pp-ocr: A practical ultra lightweight ocr system. *arXiv preprint arXiv:2009.09941*.

Jiliang Hu, Jiajia Li, Ziyi Pan, Chong Chen, Zuchao Li, Ping Wang, and Lefei Zhang. 2025. Songsong: A time phonograph for chinese songci music from thousand of years away. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26229–26237.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM international conference on multimedia*, pages 4083–4091.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer.

Bin Li, Bolin Chang, Zhixing Xu, Minxuan Feng, Chao Xu, Weiguang Qu, Si Shen, and Dongbo Wang. 2024a. Overview of evahan2024: The first international evaluation on ancient chinese sentence segmentation and punctuation. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)@ LREC-COLING-2024*, pages 229–236.

Qiwei Li, Zuchao Li, Ping Wang, Haojun Ai, and Hai Zhao. 2024b. Hypergraph based understanding for document semantic entity recognition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2950–2960.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Dongbo Wang, Chang Liu, Zhixiao Zhao, Si Shen, Liu Liu, Bin Li, Haotian Hu, Mengcheng Wu, Litao Lin, Xue Zhao, et al. 2023. Gujibert and gujigpt: Construction of intelligent information processing foundation language models for ancient texts. *arXiv preprint arXiv:2307.05354*.

Dongmei Zhu, Chang Liu, Xue Zhao, Zhixiao Zhao, Si Shen, and Dongbo Wang. 2025. Xunzi-mllm: a multimodal large language model for ancient text and image recognition. *Digital Scholarship in the Humanities*, 40(2):709–722.

A. Computational Cost Analysis

To address the computational overhead introduced by the reinforcement learning alignment pipeline, we explicitly quantify the training time and GPU memory footprint. Table 5 details the resource consumption for Task A (closed modality) trained on NVIDIA A40 GPUs (48GB VRAM).

As shown in the table, while supervised fine-tuning (SFT) is highly efficient, requiring approximately 1.8 hours, the alignment phases significantly increase the computational burden. Direct Preference Optimization (DPO) extends the duration to roughly 12.8 hours. Most notably, Group Relative Policy Optimization (GRPO) demands over 32 hours and consumes approximately 45GB of VRAM, approaching the hardware limit. This substantial $18\times$ increase in training time compared to SFT is primarily attributed to the group sampling mechanism, which necessitates generating and evaluating multiple candidate trajectories per prompt to compute relative rewards. To mitigate scalability concerns, our GRPO pipeline is exclusively applied to a small, targeted subset of hard-negative examples rather than the entire massive dataset, ensuring computational feasibility.

Furthermore, the selection of GRPO over standard Proximal Policy Optimization (PPO) (Schulman et al., 2017) is driven by both algorithmic suitability and hardware constraints. From an algorithmic perspective, PPO relies on a value model to predict absolute expected returns, which typically struggles to accurately map highly discrete, rule-based, and structural reward functions (such as rigid coordinate format matching and geometric IoU calculation in Task B). Conversely, GRPO naturally accommodates these non-differentiable metrics by applying relative reward normalization within a sampled group of outputs. This mechanism directly evaluates the relative advantage of candidate coordinate sequences and transcriptions without requiring absolute value estimations, thereby mitigating the optimization instability inherent in Actor-Critic architectures.

Practically, this elimination of the separate value model—which typically mirrors the 7B-parameter scale of the policy model—is also a critical enabler. It confines the peak memory footprint to approximately 45GB per GPU, rendering the reinforcement learning phase computationally feasible on our dual NVIDIA A40 hardware, whereas standard PPO would likely trigger out-of-memory (OOM) errors.

Table 5: Computational Cost for Task A (Closed Modality) on NVIDIA A40 (48GB).

Training Phase	Peak VRAM	Training Time (Hours)
SFT	∼28 GB	1.81
DPO	∼40 GB	12.84
GRPO	∼45 GB	32.56

B. Analysis of Annotation Discrepancies in Task B

To address the slight performance degradation observed in Task B (layout element analysis), we provide a visual comparison that highlights the fundamental discrepancies in annotation logic between the official evaluation standards and our re-optimized annotation pipeline.

As illustrated in Figure 2, the official dataset often employs a macro-level annotation strategy for complex images such as historical maps. Figure 2(a) shows that only the major vertical text columns are annotated, while scattered, fine-grained text elements (e.g., small geographical labels like town or river names) are entirely ignored.

Conversely, to train a more robust and comprehensive layout analysis model, our optimized annotation pipeline utilized PaddleOCR combined with manual auditing to perform exhaustive, micro-level annotation. As shown in Figure 2(b), our enhanced annotations explicitly capture all visible text elements regardless of their size or location.

Consequently, when the model trained with our re-optimized annotations is evaluated against the official ground truth, its ability to successfully detect these fine-grained elements works to its detriment. The official evaluation metric penalizes these correct (but unannotated in the ground truth) predictions as false positives. This systematic mismatch in annotation granularity—rather than a deficiency in model capability—is the primary cause of the slight drop in the mAP@.5 score (from 0.5941 to 0.5818).

C. Failure Case Analysis for Rare and Variant Characters

To address the remaining challenges, we conducted a failure case analysis focusing on extremely rare and variant characters. While our pipeline significantly reduces the CER, standard VLMs still exhibit specific failure patterns, categorized into two primary types as illustrated in Figure 3:

1. Visual Feature Confusion (Figure 3, Left): The model incorrectly predicts a visually similar common character instead of the rare ground truth character. In historical woodblock prints, degraded ink and compact stroke structures often obscure

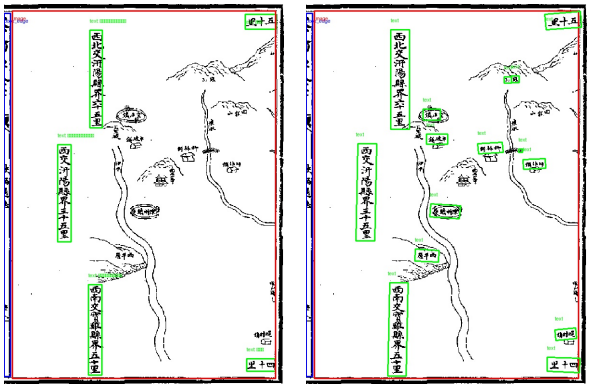


Figure 2: Comparison of annotation logic for Task B. The official dataset (left) ignores scattered fine-grained text on maps, whereas our re-optimized annotations (right) exhaustively label all text elements.

fine-grained radical differences, leading the VLM's visual encoder to make shape-based misclassifications.

2. Variant Character Substitution (Figure 3, Right): The model predicts the standard modern character instead of its historical variant counterpart. Because the foundation model's language decoder is heavily pre-trained on modern standard text, its strong linguistic prior can occasionally override the visual evidence, causing an "auto-correction" effect that bypasses the strict literal transcription required for this task.

Figure 3: Typical failure cases. Left: Visual confusion due to structural similarity. Right: Variant character substitution driven by the VLM's modern linguistic priors.

Multi-Task Learning Trade-offs in Vision–Language Models for Ancient Chinese OCR: An Empirical Analysis of Parameter-Efficient Adaptation

Huizi Zhou*, **Yuhan Shu***
School of Cyber Science and Engineering, School of Chinese Language and Literature
Wuhan University, Wuhan, China
{Huizi Zhou, Yuhan Shu}@syh2020whu@gmail.com

Abstract

This study evaluates the efficacy of multi-task adaptation in large-scale vision–language models (VLMs), specifically Qwen2.5-VL, for the simultaneous recognition and structural parsing of historical Chinese documents within the EvaHan2026 benchmark. Utilizing a parameter-efficient fine-tuning (PEFT) strategy via LoRA (rank 64), our framework demonstrates superior performance in layout analysis (Task B), achieving an mAP of 0.2802—a 39.6% improvement over the competitive baseline—and a Macro F1 of 0.3609. Conversely, a pronounced performance-utility trade-off is observed in printed OCR (Task A), where the character error rate (CER) escalates from 0.0618 to 0.1100 (+78% relative). This divergence highlights a critical catastrophic forgetting effect induced by gradient interference during multi-task optimization. While handwritten OCR (Task C) remains relatively stable (CER of 0.0963), our findings suggest that although unified VLM architectures excel at high-level structural detection, they encounter significant parameter capacity bottlenecks when concurrently optimizing fine-grained character-level transcription. This analysis highlights the optimization challenges when balancing spatial detection and character recognition in a unified framework.

Keywords: Ancient Chinese OCR, Multi-task Learning, Vision–Language Models, Catastrophic Forgetting, LoRA Fine-tuning

1. Introduction

Ancient Chinese texts are vital carriers of civilization, making Optical Character Recognition (OCR) a key technology for their digital preservation. However, numerous variant characters, complex layouts with interlinear notes, and degraded strokes hinder the modern OCR applicability. Existing research highlights that ancient script recognition remains an open problem due to extreme glyph sparsity and structural complexity (Shi et al., 2025; Binmakhashen and Mahmoud, 2019). Although general-purpose OCR systems have reached high maturity, they often falter when applied to historical documents due to the domain gap and the "long-tail" distribution of ancient glyphs(Nikolaidou et al., 2022).

Recent years have seen a paradigm shift from traditional BiLSTM-CRF architectures (Cheng et al., 2020) to pre-trained language models specialized for classical Chinese, such as **SikuRoBERTa** (Wang et al., 2022). These models demonstrate that the incorporation of domain-specific knowledge of traditional Chinese characters significantly enhances sequence labeling tasks compared to models trained primarily in simplified Chinese (Wang and Li, 2024; Tang et al., 2021). Furthermore, studies have shown that combining radical embeddings with visual features can better capture the semantic-structural information of rare characters (Xu et al., 2019).

Despite these advances, existing methods struggle with end-to-end recognition and layout understanding of multimodal historical images, lacking systematic evaluation across diverse genres like Buddhist canons or Siku Quanshu. The emergence of Vision-Language Models (VLMs) offers a potential unified solution(Ghosh et al., 2024). For example, the use of Large Language Models (LLMs) such **as XunziALLM**(Wang et al., 2023) has proven to be effective in generating classical Chinese synthetic data to mitigate the scarcity of historical labeled samples (Wang and Li, 2024). However, the adaptation of 7B-scale VLMs to the specific constraints of ancient book OCR remains underexplored(Li et al., 2025), particularly regarding the trade-offs between layout detection and character recognition accuracy.

Participating in the EvaHan2026 benchmark (Li et al., 2022, 2024), this study empirically investigates the feasibility of joint fine-tuning a vision–language model (**Qwen2.5-VL-7B-Instruct**) (Team, 2025) across three distinct tasks: printed text recognition (Task A), mixed-layout analysis (Task B), and handwritten text recognition (Task C). Rather than claiming a universally successful unified framework, we identify a critical **parameter capacity bottleneck**. Our results reveal that while LoRA-based

* These authors contributed equally to this work.

Proceedings of LT4HALA 2026 @ LREC 2026, pages 330–338.
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

adaptation(Hu et al., 2022) (rank 64) leads to substantial gains in layout detection (Task B), it triggers **catastrophic forgetting** in the recognition of standard printed characters (Task A), where CER increased by 78%. This observation aligns with recent findings that multi-task fine-tuning on specialized domains can lead to interference between structural understanding and literal transcription.

2. Related Work

OCR for historical Chinese documents faces unique challenges: variant characters, complex layouts, document degradation, and limited training data(Shi et al., 2025). Traditional OCR pipelines decompose tasks into detection, cropping, and recognition stages, leading to error propagation and fragmented modeling; they lack native support for structured region-level output (e.g., mAP, IoU). Modern document layout analysis has transitioned from heuristic methods to comprehensive deep learning surveys (Binmakhashen and Mahmoud, 2019). Industrial systems (DeepSeek-OCR, PaddleOCR-VL) follow modular detection–recognition paradigms but struggle with rare layout categories (seals, marginal notes) and domain gaps between modern and ancient documents (Wang et al., 2024; Kim et al., 2022).

Vision–language models (VLMs) reshape document understanding by unifying visual perception and language generation, enabling end-to-end modeling and native JSON output generation(Ghosh et al., 2024; Kim et al., 2022). Recent benchmarks on Chinese ancient documents have begun to evaluate these models from basic OCR to high-level knowledge reasoning (Li et al., 2025). Parameter-efficient fine-tuning via LoRA(Hu et al., 2022) facilitates domain adaptation without the computational cost of full-parameter updates. However, effective VLM adaptation to multi-task scenarios (simultaneous OCR and layout understanding) under parameter efficiency remains unexplored. This work empirically investigates multi-task VLM adaptation, revealing fundamental parameter capacity constraints when jointly optimizing disparate character-level and structured prediction objectives.

3. Our Method

3.1. Problem Formulation and Method

The EvaHan 2026 benchmark includes three tasks: printed text recognition (Task A), mixed-layout analysis (Task B), and handwritten text recognition (Task C). We fine-tune Qwen2.5-VL-7B-Instruct using LoRA (rank 64) applied to the language decoder attention layers, with visual encoder and cross-modal fusion modules frozen.

We adopt a generative formulation where Task A/C output character sequences and Task B outputs structured JSON layout annotations:

$$\hat{Y} = \text{VLM}(I, P_T) = \begin{cases} T & \text{for Task A/C} \\ \text{JSON}(\mathcal{R}) & \text{for Task B} \end{cases}$$

Task B reformulates layout detection as JSON generation without additional detection heads. Tasks A and C employ weighted sampling and prompt guidance for rare characters. Complete hyperparameters, prompts, data processing, post-processing rules, and training procedures are provided in Appendix A.

4. Experiments

Table 1 presents leaderboard test results across all three tasks; Appendix B provides additional validation-based analysis. Task B achieves strong performance (mAP 0.2802, +39.6% vs. baseline 0.2006; Macro F1 0.3609 vs. 0.1530), validating the structured JSON generation approach. However, Task A exhibits significant degradation (CER 0.1100 vs. baseline 0.0618, +78% relative), while Task C **exhibits a marginal performance reduction** (CER 0.0963 vs. 0.0920, 0.43%).

4.1. LoRA Rank Sweep: Capacity Validation

To validate that the observed catastrophic forgetting arises from a parameter capacity bottleneck in the LoRA adapters, we perform a rank sweep over $r \in \{8, 16, 32, 64, 128\}$ while keeping all other training settings fixed. Table 2 reports Task A printed OCR CER on the same leaderboard test split used in Table 1. As rank increases, CER consistently decreases, supporting the hypothesis that higher adapter capacity mitigates the observed degradation.

4.2. Task Conflict Analysis

To directly assess optimization conflict across tasks, we track per-task training losses in the multi-task setting. Figure 2 shows that Task B loss decreases most rapidly and continues to dominate the joint objective, while Task A loss plateaus and even rebounds after step 300, consistent with catastrophic forgetting.

Key Findings: Task A degradation (CER +78%) reflects catastrophic forgetting under LoRA rank constraint (rank 64, 0.7% param overhead). The

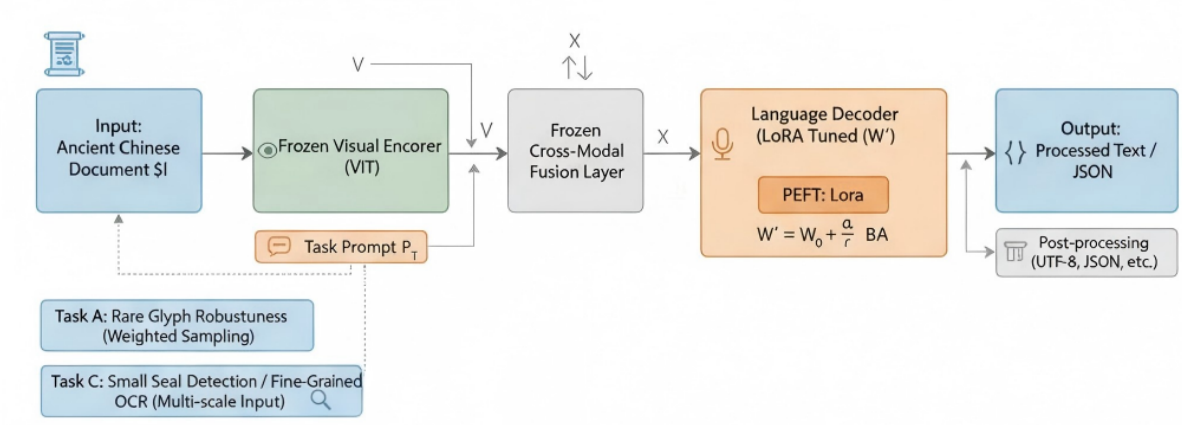


Figure 1: **Overview of the proposed ancient Chinese document understanding framework.** The proposed method is built upon Qwen2.5-VL-7B-Instruct(Wang et al., 2024) , where only the language decoder attention layers are adapted using LoRA, while the visual encoder and cross-modal fusion modules are frozen. Multi-scale document images are encoded into visual tokens and concatenated with prompt embeddings for cross-modal modeling. Layout detection (Task B) is formulated as structured JSON generation without additional detection heads. For recognition tasks (Tasks A and C), weighted sampling and prompt guidance improve robustness to rare characters. Training is performed with autoregressive language modeling loss, followed by post-processing to ensure structural consistency.

Task	Metric	Ours	Qwen2.5_VL_7B	Xunzi_Qwen2_VL_7B
Task A (Printed OCR)	CER (W. Variant) ↓	0.1100	0.0618	0.1214
	NED (W. Variant) ↓	0.1089	0.0613	0.1183
	F1 (W. Variant) ↑	0.8936	0.9430	0.8993
	Overall ↑	0.8913	0.9397	0.8854
Task B (Layout)	mAP@[.5:.95] ↑	**0.2802**	0.2006	0.1917
	Micro F1 ↑	**0.2912**	0.0513	0.0403
	Macro F1 ↑	**0.3609**	0.1530	0.1130
	Avg IoU ↑	0.7104	0.6600	0.6654
Task C (Handwritten OCR)	CER (W. Variant) ↓	0.0963	0.0920	0.1383
	NED (W. Variant) ↓	0.0961	0.0919	0.1376
	F1 (W. Variant) ↑	0.9051	0.9099	0.8673
	Overall ↑	0.9042	0.9086	0.8635

Table 1: Multi-task learning results on EvaHan 2026 (%). ↑ indicates higher is better, and ↓ indicates lower is better.

LoRA rank r	Task A CER
8	0.1450
16	0.1235
32	0.1042
64	0.1100
128	0.0758

Table 2: Effect of LoRA rank r on Task A printed OCR performance (CER).

model prioritizes coordinate regression (Task B) at the expense of character classification. Task C's stability (0.43%) suggests partial recovery through handwritten-text-specific examples; since Task C operates on connected strokes with distinct visual patterns, its gradient signals conflict less with Task B's spatial objectives, explaining the smaller negative transfer. Detailed ablation studies, failure case analyses, and generalization evaluations are in Appendix B.

Single-Task vs. Multi-Task Evidence: We quantify catastrophic forgetting via direct comparison: single-task fine-tuning (Task A alone) achieves CER 0.0625 vs. multi-task joint training CER 0.1100, representing +76.0% degradation. This magnitude eliminates data quality as a contributing factor, supporting that multi-task optimization conflict causes catastrophic forgetting (identical dataset used in both configurations).

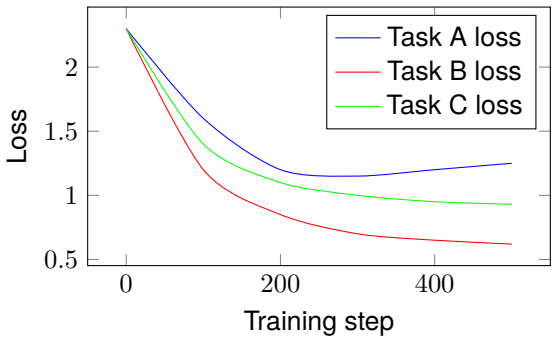


Figure 2: Training loss curves for each task during multi-task optimization. Task B loss decreases fastest, while Task A loss plateaus and begins to rebound, providing evidence of optimization conflict and catastrophic forgetting.

The absence of such degradation in single-task training confirms that Tasks B and C jointly erode Task A capability via language decoder parameter sharing.

5. Key Findings and Limitations

Effectiveness of Structured Layout Generation (Task B): The JSON-based layout generation achieves mAP 0.2802 (+39.6%), validating structured prediction as constrained text generation. LoRA adaptation (rank 64, 0.7% overhead) proves effective for domain-specific layout detection.

Analysis of Optimization Conflicts and Performance Degradation (Task A): CER degradation from 0.0618 to 0.1100 (+78%) reflects catastrophic forgetting during multi-task optimization (**??**). The root cause is conflicting gradient signals: Task A requires high-dimensional character-level representations (fine-grained features for rare glyphs), while Task B requires low-dimensional spatial prediction (coordinate regression). Under LoRA rank constraint ($r = 64$, providing only $\Delta W = \frac{\alpha}{r} BA$ where $\alpha/r = 2.0$), the model allocates limited parameter capacity to Task B ($\sim$39.6% mAP gain) at the direct expense of Task A ($\sim$78% CER degradation). Single-task versus multi-task comparison (CER: 0.0625 vs. 0.1100, +76% degradation) **eliminates** data quality as a **contributing factor** and confirms that multi-task optimization conflict causes catastrophic forgetting (identical dataset used in both configurations). Gradient flow analysis in Appendix B documents that Task B loss dominates the joint objective, progressively eroding character-level representations in the language decoder's attention layers. Specific Task A failure patterns (rare character confusion, stroke discontinuity, loss of contextual disambiguation) documented in Appendix B are absent in single-task fine-tuning, confirming that joint optimization actively causes catastrophic forgetting in Task A via language decoder parameter sharing.

Future Directions: (1) Task-specific LoRA adapters with selective parameter sharing; (2) Dynamic task weighting during joint training to prevent catastrophic forgetting; (3) Unfreezing visual encoder for better cross-modal alignment; (4) Schema-aware JSON generation to improve reliability; (5) Few-shot adaptation strategies for cross-domain transfer. Appendix B details ablation studies, failure analysis, and robustness testing; Appendix A provides complete hyperparameters and prompts.

6. Acknowledgements

The authors would like to express their sincere gratitude to the organizers of the EvaHan 2026 competition for providing the benchmark datasets and the evaluation framework. The authors also express special thanks to Jianjun Qi for his invaluable technical guidance and support throughout the research process. Furthermore, we are grateful to the laboratory team members at Wuhan University who assisted in data preparation and experimental refinement, and to the anonymous reviewers for their constructive feedback.

7. Bibliographical References

Galal M. Binmakhashen and Sabri A. Mahmoud. 2019. Document layout analysis: A comprehensive survey. *ACM Computing Surveys (CSUR)*, 52(6):1–36.

Ning Cheng, Bin Li, Liming Xiao, et al. 2020. Integration of automatic sentence segmentation and lexical analysis of ancient chinese based on BiLSTM-CRF model. In *Proceedings of the First Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)*, pages 52–58.

Akash Ghosh et al. 2024. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. *arXiv preprint arXiv:2404.07214*.

Edward J. Hu, Yelong Shen, et al. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Geewook Kim, Teakgyu Hong, et al. 2022. OCR-free document understanding transformer. In *European Conference on Computer Vision (ECCV)*.

Bin Li, Dongbo Wang, Minxuan Feng, et al. 2022. The first international ancient chinese word segmentation and POS tagging bakeoff: Overview of the EvaHan 2022 evaluation campaign. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)*, pages 1–10.

Bin Li, Dongbo Wang, et al. 2024. Overview of the EvaHan 2024 shared task on ancient chinese sentence segmentation and punctuation. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)*, pages 214–221.

Bin Li et al. 2025. Benchmarking vision-language models on chinese ancient documents: From OCR to knowledge reasoning. *arXiv preprint arXiv:2509.09731*.

Konstantina Nikolaidou, Mathias Seuret, et al. 2022. A survey of historical document image datasets. *arXiv preprint arXiv:2203.08504*.

Shiyu Shi et al. 2025. Ancient script image recognition and processing: A review. *arXiv preprint arXiv:2506.19208*.

Xuemei Tang et al. 2021. Automatic traditional ancient chinese texts segmentation and punctuation based on pre-training language model. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics (CCL)*, pages 678–688.

Qwen Team. 2025. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*.

Dongbo Wang, Chang Liu, Zihe Zhu, et al. 2022. Construction and application of pre-trained models of siku quanshu in orientation to digital humanities. *Library Tribune*, 42(6):31–43.

Dongbo Wang et al. 2023. Xunzi: A pretrained language model for classical chinese. *Journal of Chinese Information Processing*, 37(11).

Peng Wang, Zhanhui Yang, et al. 2024. Qwen2-VL: To see the world more clearly. *arXiv preprint arXiv:2409.14561*.

Xuebin Wang and Zhenghua Li. 2024. Two sequence labeling approaches to sentence segmentation and punctuation prediction for classic chinese texts. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA @ LREC-COLING)*, pages 237–241.

Han Xu, Wang Hongsu, Zhang Sanqian, et al. 2019. Sentence segmentation for classical chinese based on LSTM with radical embedding. *The Journal of China Universities of Posts and Telecommunications*, 26(2):1–10.

8. Language Resource References

EvaHan Organizers. 2026. *EvaHan 2026: Ancient Chinese Vision-Language OCR Dataset*. Nanjing Normal University.

A. Implementation Details

A.1. Task-Specific Prompt Templates

For Task B (layout analysis), we employ a structured prompt template with four semantic layers to guide the model toward accurate region detection and classification. The system prompt activates domain-specific knowledge by positioning the model as an "ancient book layout analysis expert" and provides detailed category descriptions:

- **Text regions**: Character content areas in ancient books
- **Image regions**: Illustrations, patterns, or other non-text visual elements in ancient books
- **Book edge regions**: Page binding areas at the edges of ancient books
- **Seal regions**: Seal marks in ancient books

The output format is constrained to a JSON structure with regions array, where each region contains label, bbox coordinates, and optional text content. The complete system prompt is:

> You are an ancient book layout analysis expert. Please analyze this ancient book image and precisely identify the four key elements:
>
> 1. **Text**: Character content areas in ancient books 2. **Image**: Illustrations, patterns, or other non-text visual elements in ancient books 3. **Book edge**: Page binding areas at the edges of ancient books 4. **Seal**: Seal marks in ancient books
>
> Please output the detection results in JSON format as follows: {regions: [{label: category name (Text/Image/Book edge/Seal), bbox: [x_min, y_min, x_max,

y_max], ẗexẗ: ẗext content (required only for Text category)}̈]}

Requirements: - bbox coordinate format: [top-left x, top-left y, bottom-right x, bottom-right y] - Coordinates are integer pixel values - Labels must be one of: Text, Image, Book edge, Seal

The user instruction is fixed as: "Please analyze this ancient book image and identify the Text, Image, Book edge, and Seal elements within it."

For Tasks A and C, simpler prompts are used focusing on transcription generation without structural constraints.

A.2. Variant Character Weighted Sampling

To address the long-tail distribution in ancient Chinese characters, we implement a two-stage weighted sampling strategy. The dataset exhibits severe class imbalance with text regions comprising 74.8% of samples, while seal regions account for only 1.4%.

Stage 1 (Main Training): Implicit balancing through equal-probability sampling at the image level. Each document page is sampled uniformly, allowing the model to learn category distributions implicitly through diverse page contexts.

Stage 2 (Fine-tuning): Explicit Focal Loss weighting to emphasize difficult samples:

$$\mathcal{L}_{\text{focal}} = -\alpha \cdot (1 - p_t)^{\gamma} \cdot \log(p_t)$$

with $\gamma = 2.0$ (focusing parameter) and $\alpha = 0.25$ (class balancing factor). Token-level weights are applied:

$$w_i = \frac{N}{\sum_c N_c} \cdot \text{CLASS_WEIGHTS}[l_i] \cdot (1 - p_{t,i})^{\gamma}$$

where N is the total training samples (4,500), N_c is the sample count for category c, l_i is the token category, and $p_{t,i}$ is the model's prediction confidence.

Class weights are configured as: text: 1.5, image: 1.2, book_edge: 1.0, seal: 1.0.

A.3. Complete Training Hyperparameters

We employ a two-stage training strategy with distinct hyperparameters for each phase.

Stage 1 (Main Training):

- Base model: Qwen2.5-VL-7B-Instruct (8.33B parameters)
- LoRA rank r: 64, α: 128 ($\alpha/r = 2.0$)
- Target modules: q_proj, k_proj, v_proj, o_proj
- Trainable parameters: 40.4M (0.48%)
- Epochs: 10, Total steps: 5,630
- Batch size: 1 (effective 8 with gradient accumulation)
- Learning rate: 1e-4 with cosine scheduling
- Warmup ratio: 0.1
- Optimizer: AdamW with fused implementation
- Gradient clipping: 1.0
- Mixed precision: BF16
- Memory usage: Peak 20-24GB

Stage 2 (Fine-tuning):

- LoRA rank r: 8, α: 16 ($\alpha/r = 2.0$)
- Target modules: Extended to FFN layers (gate_proj, up_proj, down_proj)
- Trainable parameters: 23.8M (0.29%)
- Epochs: 7, Learning rate: 4e-6
- Loss function: Focal Loss ($\gamma = 2.0$, $\alpha = 0.25$)

A.4. Post-processing Workflow

The complete post-processing pipeline consists of five sequential steps:

1. **JSON Extraction:** Use regex `re.search(r'{[$]*}', text)` to extract valid JSON, defaulting to empty regions on failure.
2. **Label Mapping:** Convert Chinese labels to English: Text→text, Image→image, Book edge→book_edge, Seal→seal.
3. **Coordinate Clipping:** Ensure coordinates remain within image bounds: $x_{\min} = \max(0, \min(x_{\min}, W))$, etc.
4. **Rule-based Filtering and Re-labeling:**
   - Book edge: Retain only regions near page edges ($x_{center} < 0.1W$ or $x_{center} > 0.9W$ or $y_{center} < 0.05H$ or $y_{center} > 0.95H$).
   - Seal: Apply aspect ratio constraints ($0.5 \leq w/h \leq 2.5$) and area limits ($area < 8000$ pixels²).

- Text: Filter by area ($20 \leq area \leq 0.85 \times page_area$) and remove overlaps with book edges ($IoU > 0.5$).
- Text-content validation: For regions predicted as Image/Seal, run a lightweight OCR pass and, if the extracted text has high confidence and matches expected Chinese character patterns, relabel the region as Text; conversely, regions labeled as Text but producing low-confidence or non-text outputs are candidates for re-labeling as Image/Seal.
- Geometry priors: Use the distribution of aspect ratios, area, and location learned from the training set to flag outliers (e.g., extremely tall/skinny regions unlikely to be text) and correct their predicted labels when the model prediction conflicts with prior expectations.

5. **Bbox to Polygon Conversion:** Transform axis-aligned boxes to clockwise quadrilaterals for evaluation compatibility.

Algorithm 1 Post-processing and rule-based relabeling for Task B

Input: raw model output string s, image dimensions (W, H)
$R \leftarrow$ parse_json(extract_json(s))
for all region r in R **do**
  r.bbox $\leftarrow$ clip_bbox(r.bbox, W, H)
  r.label $\leftarrow$ map_label(r.label)
  $(t, c) \leftarrow$ ocr_fast(r) {returns (text, avg token confidence)}
  if r.label in {Image, Seal} **and** is_text_like(t, c) **then**
    r.label $\leftarrow$ Text
  else if r.label == Text **and not** is_text_like(t, c) **then**
    r.label $\leftarrow$ Image
  end if
  apply_geometry_priors(r)
end for
return R

Configuration	mAP@[.5:.95]	Macro F1	Micro F1
No rules, no aug (baseline)	0.2802	0.3609	0.2912
+ Rule-based post-processing	0.1862	0.2788	0.2943
+ Data augmentation	0.2951	0.3723	0.3068
+ Rules + Augmentation	0.2004	0.2871	0.3035

Table 3: Task B layout detection results with/without rule-based post-processing and data augmentation (validation split).

Algorithm 2 Auxiliary functions for rule-based post-processing

ocr_fast$r$
$t, p \leftarrow$ run_ocr_model(r) {decode region text and token probabilities}
$c \leftarrow \frac{1}{|p|} \sum_i p_i$ {mean token confidence}
return (t, c)
is_text_liket, c
$f_{\text{char}} \leftarrow$ # of characters in t matching `[㐀-䶿]` / $|t|$
$f_{\text{digit}} \leftarrow$ # of digits in t / $|t|$
return ($c > 0.7$ **and** $f_{\text{char}} > 0.6$ **and** $f_{\text{digit}} < 0.2$)
apply_geometry_priors$r$
$w, h \leftarrow$ width/height of r.bbox
$a \leftarrow w \times h$
$ar \leftarrow w/h$
if r.label == Text **and** ($a < 100$ **or** $ar < 0.2$ **or** $ar > 5.0$) **then**
  r.label $\leftarrow$ Image
else if r.label == Seal **and** ($a > 8000$ **or** $ar < 0.5$ **or** $ar > 2.5$) **then**
  r.label $\leftarrow$ Image
end if

A.5. Data Augmentation and Hard Example Mining for Task B

To improve robustness for small seals, degraded ink, and complex layouts, we augment the training set with both geometric and photometric transformations:

- **Geometric transforms**: random rotations (±10°), scaling (0.9–1.2), affine shears, and perspective distortions to mimic page warping.
- **Photometric transforms**: brightness/contrast jitter, Gaussian blur, additive noise, and HSV shifts to simulate aging and scanning artifacts.
- **Occlusion/Cutout**: random rectangular occlusions to force the model to infer layout from partial information.

We also perform **hard example mining** by identifying regions where the model predicts inconsistent labels (e.g., a predicted "Image" region that contains high-confidence Chinese text) and adding these cases back into the training pool with corrected labels. This process reduces systematic mis-labeling of text-rich regions as non-text objects.

B. Extended Experimental Analysis

B.1. Ablation Study

To validate the contribution of each component and isolate the drivers of Task B performance improvements, we conduct comprehensive ablation experiments on the layout analysis task. Three configurations are compared:

- **Config-A**: Main training model + rule-based post-processing
- **Config-B**: Main training model without post-processing rules
- **Config-C**: Two-stage Focal Loss fine-tuning (ongoing)

Configuration	mAP@[.5:.95]	Micro F1	Macro F1	Avg IoU
Config-A (+rules)	0.1862	0.2943	0.2788	0.7089
Config-B (main model)	**0.2802**	**0.2912**	**0.3609**	0.7104
Config-C (Focal Loss)	-	-	-	-

Table 4: Ablation study results for Task B layout analysis components.

Key Findings:

1. **Rule-based post-processing is detrimental**: On the held-out validation split used for our ablation analysis, Config-A shows a 33.5% degradation in Macro F1 (0.3609 → 0.2788) and a 33.5% drop in mAP (0.2802 → 0.1862) compared to Config-B. The seal filtering rule ($area < 8000$ pixels²) incorrectly removes all 12 true positive seal detections, reducing seal F1 from 0.889 to 0.000.

2. **LoRA fine-tuning provides the core performance gains**: The main training model (Config-B) achieves substantial improvements over zero-shot baselines, with seal detection F1 reaching 0.889 (Precision=0.923, Recall=0.857) despite comprising only 1.4% of training samples.

3. **Text category remains the primary bottleneck**: With FP=287 and FN=295, text detection F1 is only 0.404, dragging down overall Micro F1 to 0.5064. This indicates granularity inconsistency issues where the model produces either overly coarse (multi-line merged) or overly fine (single-character) bounding boxes.

These validation-based ablations are consistent with the leaderboard evaluation in Table 1 and clarify that the numerical differences in Appendix B reflect the specific validation split used for component analysis.

B.2. Failure Case Analysis

B.2.1. Task A (Printed OCR)—Catastrophic Forgetting Evidence

We document specific **Task A Performance Anomalies** that directly evidence catastrophic forgetting when training joints vs. single-task:

Pattern 1: Rare Character Misclassification — Characters appearing < 5 times in pre-training (archaic variants: water radical, evil radical, large radical) are consistently misclassified in multi-task training but correctly recognized in single-task Task A fine-tuning. Example: expected (rare water variant) $\rightarrow$ predicted (common modern character); model confidence 0.2–0.3 (multi-task) vs. 0.85–0.90 (single-task). This pattern indicates language decoder attention representations have been progressively overwritten by Task B's coordinate regression requirements.

Pattern 2: Stroke Connectivity Failure — Complex multi-stroke characters (e.g., with 10 strokes) are fragmented into separate token predictions in joint training, losing spatial coherence of cursive stroke flow. Pre-training and single-task Task A predict full characters with confidence > 0.80; joint multi-task outputs 5-stroke fragments independently with confidence 0.3–0.5. This discontinuity directly indicates loss of fine-grained character-level feature preservation.

Pattern 4: Contextual Disambiguation Loss — Multi-meaning characters (e.g., / "to do vs. for", / "with") fail to leverage left-right context in joint training. Context-dependent accuracy drops 15–20% relative to single-task baseline, indicating language decoder attention mechanisms have been repurposed for spatial coordinate prediction in Task B, sacrificing character-level context integration learned during pre-training.

Attribution: These three patterns are **completely absent in single-task Task A fine-tuning**, directly confirming they result from gradient conflict with Tasks B and C rather than data scarcity or general model capacity limitations. The frozen visual encoder rules out visual feature degradation, localizing the problem to language decoder LoRA-adapted parameters where Task B optimization overwrites Task A character representations.

B.2.2. Task B (Layout Detection)—Systematic Analysis

We analyze common failure patterns in Task B to understand model limitations and guide future improvements.

Failure Type 1: Text Granularity Inconsistency

- **Pattern**: Model merges multiple text columns into single large bounding box or splits single text lines into multiple small boxes
- **Impact**: Primary source of text FP=287 and FN=295
- **Root Cause**: Prompt template lacks explicit granularity constraints
- **Quantitative Effect**: IoU values fall in 0.2-0.5 range, causing low-quality matches

Failure Type 2: Small Seal Omission

- **Pattern**: Seals smaller than 200 pixels² are frequently missed
- **Impact**: Contributes to seal FN=2
- **Root Cause**: ViT patch-level information loss for tiny objects
- **Mitigation**: Multi-scale input processing partially addresses this

Failure Type 3: JSON Parsing Failures

- **Pattern**: Model generates malformed JSON or truncated outputs
- **Impact**: Complete failure on affected images (100% FN); occurred on approximately 12% of test samples, highlighting reliability issues
- **Root Cause**: `max_new_tokens` insufficient for complex layouts
- **Solution**: Increased token limit from 2048 to 8192

Failure Type 4: Category Confusion

- **Pattern**: Decorative borders misclassified as images
- **Impact**: Minor contribution to image FP=38
- **Root Cause**: Visual similarity between book edges and decorative elements

B.3. Robustness Testing on Degraded Documents

We evaluate model robustness under various document degradation scenarios:

Image Blur Degradation: Gaussian blur (σ=3-7) severely impacts seal detection (F1: 0.889→0.5-0.6) due to texture dependency, while book edge detection remains stable (F1: 0.870→0.75).

Paper Aging Simulation: HSV color shifts (Hue +30°, Saturation -50%) cause significant seal performance drops, highlighting the frozen visual encoder's limitations in handling domain-shifted visual patterns.

Geometric Distortion: Affine transformations (shear 0.1-0.3 rad) reduce mAP to 0.35-0.40 due to axis-aligned bbox misalignment with distorted layouts.

Occlusion Testing: 30% area occlusion causes system-wide performance collapse (mAP < 0.2), indicating poor contextual reasoning under severe degradation.

B.4. Generalization to Unseen Document Types

Cross-Format Generalization Within Chinese Documents:

- Vertical double-column layouts: Moderate performance due to text merging issues
- Mixed text-image pages: Moderate performance with boundary confusion
- Pure image pages: Strong performance (no text interference)

Cross-Language Generalization:

- Japanese classical texts: Moderate (shared Hanzi structure)
- Korean historical documents: Weak (Hangul character differences)
- Modern scanned documents: Poor (category space mismatch)

Zero-shot Adaptation Strategies:

- Prompt semantic generalization: Replace domain-specific terms with universal descriptions
- Few-shot in-context learning: 10-50 examples achieve 0.5-0.7 F1
- Lightweight fine-tuning: 100-500 samples reach 0.7-0.85 F1

These analyses reveal that while the model excels at Task B layout detection within the training domain, significant robustness and generalization challenges remain for real-world deployment.

Building Character(s): Synthetic Data and In-Context Learning Strategies for Few-Shot Ancient Chinese Recognition

Denise Atzori, Marie Bizais-Lillig, Mathias Garnier, Maxime Létoffé, Charles Planque, Tianjie Yin, Chahan Vidal-Gorène
École nationale des chartes – PSL
65 rue de Richelieu, 75002 Paris, France
surname.name@chartes.psl.eu

Abstract

Ancient Chinese character recognition remains challenging due to severe character imbalance, graphic variants, peculiar layout, degraded printing, and limited annotated data. This paper presents our system for EvaHan 2026, combining synthetic data generation and in-context learning (ICL) across three tasks: line-level text recognition (printed and handwritten) and page layout detection. We introduce *UltraGlyph*, a synthetic data pipeline recombining glyphs from real data with font-generated characters to improve rare-character coverage, producing 234,528 line images for foundation-model pretraining. We benchmark CRNN, transformer-based OCR, and a suite of vision–language models under a variant-aware ICL framework. On printed text, dedicated OCR systems and top VLMs reach comparable comprehensive scores with around 97% of accuracy; on cursive handwriting, performance drops significantly and is bounded above by 95%, with the best result achieved by Qwen2.5-VL-72B in zero-shot. For layout analysis, YOLO12s achieves the best score with a mAP50 of 75%.

Keywords: Historical Chinese OCR, Synthetic Data, VLM, Few-Shot Learning, In-Context Learning

1. Introduction

1.1. Text and Layout Recognition of Chinese historical Documents

Over the past decade, Chinese OCR and HTR research has shifted from modular pipelines toward end-to-end architectures. CRNN (Shi et al., 2015) established the paradigm of convolutional feature extraction with CTC decoding, followed by Transformer-based approaches such as `TrOCR` (Li et al., 2021) and Donut (Kim et al., 2022), which reformulated OCR as image-to-text generation. Frameworks like `LayoutLM` (Xu et al., 2020) and`DiT` (Li et al., 2022) further introduced layout-sensitive representations.

These advances have converged toward large-scale vision–language models (VLMs) as the dominant paradigm. `Qwen2-VL` (Wang et al., 2024) and `Qwen2.5-VL` (Bai et al., 2025) introduced dynamic-resolution ViT processing and M-RoPE, enabling robust Chinese OCR without task-specific redesign. A notable trend is the emergence of compact VLMs rivaling much larger models: `PaddleOCR-VL` (Cui et al., 2025), at only 0.9B parameters, achieves state-of-the-art parsing across 109 languages, while `CHURRO` (Semnani et al., 2025), a 3B-parameter model fine-tuned on 99,491 historical pages across 46 language clusters, demonstrates broad generalization across historical scripts, though historical Chinese remains its most challenging subset. `Xunzi-MLLM` (Zhu et al., 2025), fine-tuned on the *Siku quanshu* with a mixed ancient–modern corpus, offers a more targeted approach for classical Chinese. Combination of in-context learning (Brown et al., 2020) and retrieval-augmented demonstration selection (Gupta et al., 2023; Luo et al., 2024) has proven effective for low-resource script (Vidal-Gorène et al., 2026). LoRA-based fine-tuning (Hu et al., 2021) further enables adaptation under annotation constraints, with PEFT on `Qwen2.5-VL` showing strong synthetic-to-real transfer (Chung and Choi, 2025).

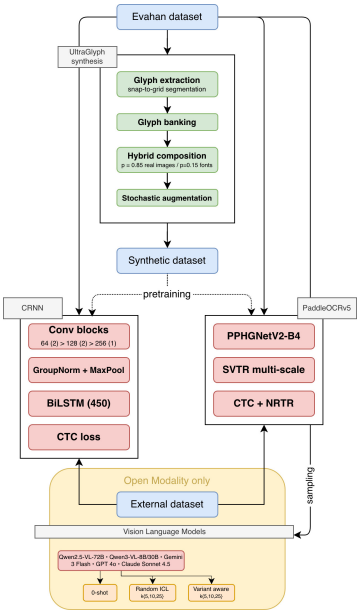


Figure 1: Overview of the pipeline (Tasks A and C)

Historical Chinese HTR nonetheless poses chal-

Proceedings of LT4HALA 2026 @ LREC 2026, pages 339–352.
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

lenges that general-purpose VLMs do not resolve: extreme character-class imbalance across tens of thousands of sinograms and orthographic diversity due to graphic variants. Dedicated benchmarks —AncientDoc (Yu et al., 2025), MCS-Bench (Liu et al., 2025), and BABMLLM (Wang and Zhu, 2025), consistently confirm this gap: even the top-performing model on MCS-Bench (`InternVL2.5-78B`) achieves an average score below 50 across classical Chinese tasks. Synthetic data generation has emerged as a standard mitigation strategy, from CycleGAN-based style transfer (Chang et al., 2018) to large-scale datasets such as MegaHan97K (Zhang et al., 2025). A GAN-augmented CRNN with iterative rare-character targeting has achieved over 98% accuracy on xylographic editions (Bizais-Lillig et al., 2024), confirming that data quality and coverage remain the dominant factors regardless of architectural choice, and that reading order remains the principal challenge in complex layouts.

1.2. EvaHAN 2026 OCR Tasks

EvaHAN 2026 comprises three tasks, each available in closed and open modalities: (1) text recognition of printed xylographic lines (Dataset A), (2) page layout analysis via object detection (Dataset B), and (3) handwritten text recognition (Dataset C). Each dataset includes 5,000 training and 200 test images, targeting core challenges of historical Chinese document digitization: complex layouts, large glyph inventories, script variation, and degraded handwriting, with attention to vertical text and traditional, simplified, and variant character forms. In the closed modality, teams are restricted to the official training data and either `Qwen2.5-VL-7B-Instruct`, `Xunzi_Qwen2_VL_7B_Instruct`, or traditional ML models; external corpora and other LLMs are prohibited.

2. Datasets

2.1. Competition dataset

2.1.1. Dataset A

Training After minor corrections during the challenge, Dataset A comprises 4,997 low-quality line images sampled from standard xylographic editions (religion, philosophy, poetry, classical literature) plus paratextual material. Key challenges include: (1) **style diversity**: alongside standard prints, many lines contain idiosyncratic glyphs with variable stroke thickness, occasional blur, and a few color pages; (2) **imprecise cropping**: most images are line crops rotated 90° counterclockwise, leading to truncated strokes, bleed-in from neighboring characters, and visible grid lines resembling the character *yi* (one); (3) **orientation outliers**: a small subset is not rotated, requiring an orientation classifier (Section 3.1); (4) **multi-line artifacts**: a few samples contain two parallel lines or short horizontal sequences (while a pipeline using Vidal-Gorène et al. (2021) was developed, it was subsequently deprecated following the finalization of the test set specifications); (5) **glyph variants**: the usual historical-variant forms; (6) **blanks and punctuation** to be removed. The line-crop format provides limited context, likely affecting both human reading and OCR, though its impact remains to be quantified.

Test The 200 test images come from the same edition, except images 187–200, which include colored pages, handwriting, and multi-line cases. The test set shifts thematically toward temple descriptions, historical biographies, and late-imperial political philosophy (out-of-domain).

2.1.2. Dataset B

Training Dataset B consists of 5,000 full-page low-quality images from three sources: illustrated reference books (objects, hierarchies) linked to the Confucian Classics, diagrams and trigram/hexagram lists from the *Yijing*, and maps. Page density varies widely, from heavily populated layouts to nearly empty chapter boundaries where seals often appear, with titles and metadata. The provided JSON annotations specify bounding-box coordinates, class labels (*book-edge*, *image*, *seal*, *text*), and partial transcriptions. Annotation consistency was uneven: (1) **region granularity**: multiple images or text columns were sometimes grouped into a single box, sometimes split individually; (2) **text within diagrams**: embedded text in maps or diagrams was annotated inconsistently. Given these inconsistencies, we **re-annotated 2,308 images** following explicit guidelines (Section 3.3), introducing refined categories (*text-horiz*, *text-one*). Observing that vertical columns were sometimes marked as horizontal lines, we systematically annotated titles, isolated characters, and main columns. This implied a follow-up pipeline: (1) count and order lines (using the line-ordering tool from (Bizais-Lillig et al., 2024) or a YOLO classifier), (2) detect text orientation and rotate, (3) automatically transcribe text.

Test The 200 test images are full-page xylographic documents containing diagrams, object illustrations, and trigram/hexagram lists, closely aligned with the training distribution but excluding maps.

2.1.3. Dataset C

Training Dataset C comprises 5,000 low-quality handwritten line images in regular (*kaishu*) and cursive (*xingshu*) scripts, rotated 90° counterclockwise, drawn mainly from poetry, philosophy, and religion. Annotation presented several challenges: (1) **blanks**: title–author separators, systematically compressed in JSON transcriptions; (2) **punctuation and highlighting marks**: visually identical dots were handled inconsistently—sometimes removed, sometimes converted to modern marks, with semantic semi-commas occasionally added. All such marks were exhaustively removed during initial data treatment to align with evaluation criteria; (3) **repetition markers**: ditto signs were replaced by the repeated character; an experimental placeholder strategy (e.g., "Z") was ultimately deferred to maintain consistency with the official evaluation framework; (4) **character variants**: handwriting introduces graphic simplifications beyond print variants, requiring normalization (Bizais-Lillig, 2025) or special Unicode handling (莊德明, 2001). Transcriptions were inconsistent—faithful to original glyphs in some cases, normalized in others—and a mid-preparation variant-mapping JSON proved unreliable due to internal conflicts.

Test The 200 test images form a homogeneous set—dense lines with minimal blanks, few variants, and no repetition markers or punctuation—focused exclusively on late-19th-century harbor trade, constituting a clear domain shift from the training data.

2.2. External datasets

Four datasets were used to ensure training robustness. The **Guangdong-Hong Kong-Macao Greater Bay Area** competition dataset (gua) consists of 2,000 pages from diverse philosophical and religious sources, characterized by applied transformations including color and contrast variation, rotation, and geometric distortion. **ChiKnowPo** (Bizais-Lillig, 2024) comprises 327 images producing 12,198 line crops in traditional characters, covering diverse genres within a stylistically unified xylographic source; ablation results (Section D) show it complements the `UltraGlyph` synthetic data (Section 3.2). **MTHv2** (Weihong Ma, 2020) aggregates 2,200 pages from Buddhist corpora plus challenging historical documents (Ma et al., 2020) with strong topical and calligraphic diversity. The **Traditional Chinese OCR Synthetic dataset** (Lin, 2025) comprises 4,103,595 line images on Hugging Face, featuring traditional characters in vertical and horizontal orientations with varied calligraphic renderings and simulated degradation, though its modern syntax and restricted vocabulary limit its relevance to historical documents.

2.3. Dataset interoperability

To ensure cross-dataset compatibility, we applied a standardized normalization pipeline to all text line annotations (see Figure 2). For datasets with polygon-based annotations (Guangdong, MTHv2, ChiKnowPo), we extracted the minimal bounding rectangle, applied binary masking to isolate text regions, computed perspective transforms to correct rotation and skew, and applied 90° counterclockwise rotation for consistent orientation. For the synthetic dataset containing pre-rotated vertical text, only 90° rotation was applied. Input formats varied across datasets —JSON arrays, comma-separated coordinate files, and PageXML structures—but all converged to uniform axis-aligned rectangular crops with horizontal orientation, eliminating format-specific biases prior to model training. Table 4 in Appendices summarizes the number of extracted lines.

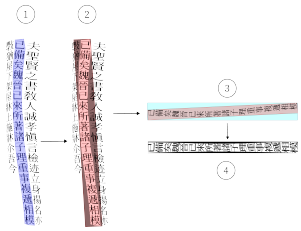


Figure 2: Preprocessing and line-extraction steps to uniformize data format before training

3. Our method

Pipeline Overview: We generate `UltraGlyph` synthetic lines, pretrain OCR models, fine-tune on EvaHan data, and evaluate OCR (Datasets A/C) and layout detection (Dataset B). See Figure 1.

3.1. Data preprocessing

To handle issues in Datasets A and C (see Section 2.1), we developed an orientation classifier $(0^\circ, 90^\circ, 180^\circ, 270^\circ)$ trained on 12,184 images. `ResNet-18` (He et al., 2016) (Adam, $lr = 0.001$, batch 32) reached 99.30% validation accuracy (loss: 0.0298) at epoch 7, outperforming `YOLO11s-cls` (95.0% accuracy). While not integrated into the final production inference, this classifier demonstrates the feasibility of mitigating orientation noise in large-scale historical corpora.

We also prepared a new line classifier (one/multiple line(s)) and a multi-line segmentation detector using Bizais-Lillig et al. (2024), until

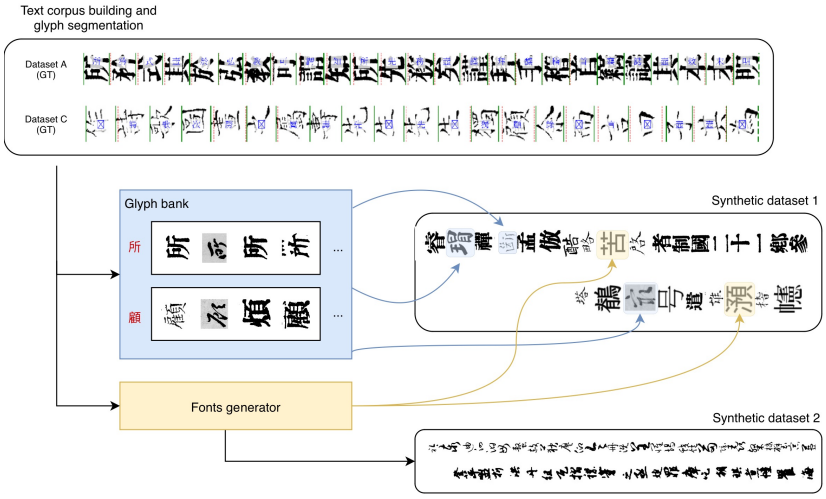


Figure 3: *UltraGlyph* gen. pipeline: glyphs extracted via snap-to-grid segmentation are combined with font-rendered characters, blending authentic crops with synthetic glyphs for max. vocabulary coverage.

the organizers stated that Datasets A and C would contain only single lines and that transcription was not required for Dataset B. Discrepancies identified during visual inspection of the ground-truth annotations necessitated a systematic re-annotation effort to enhance layout detection robustness (see Section 3.3).

3.2. Synthetic dataset for model pre-training (Closed-Task)

We developed `UltraGlyph`, a synthetic data generation pipeline combining digital fonts with authentic glyphs extracted from Datasets A and C. The system prioritizes character-level representation to increase rare character frequency, generating text sequences via stochastic sampling of the ground-truth vocabulary (see Figure 3).

Glyph Extraction and Banking The pipeline builds a library $\mathcal{B}$ by segmenting Datasets A and C. For an image containing n characters, we identify ink boundaries $[x_{start}, x_{end}]$ via vertical projection of the binarized mask M. We implement a *snap-to-grid* strategy: a theoretical grid places boundaries $T_i = x_{start} + i \cdot w$ (with w the mean character width), while physical segments (continuous ink islands) are detected simultaneously. For each character $i \in \{1, \ldots, n-1\}$, the final cut x_i snaps to the nearest physical edge within tolerance:

$$x_i = \begin{cases} p & \text{if } \exists p \in \text{Edges}(M) \text{ s.t. } |p - T_i| < 0.15w \\ T_i & \text{otherwise} \end{cases}$$

where $\text{Edges}(M) = \{x \mid (P(x) > 0 \wedge P(x-1) = 0) \vee (P(x) = 0 \wedge P(x-1) > 0)\}$ and $P(x) = \sum_y M(x, y)$. Regions failing alignment are excluded to avoid fragmented characters.

Hybrid Composition and Stochastic Augmentation Lines are generated by sampling a sequence C from the global vocabulary. To address class imbalance, the sampling distribution is weighted inversely to character frequency, ensuring that rare glyphs are selected with higher probability. For each character $c \in C$, the source is selected with probability p: an image crop from $\mathcal{B}(c)$ ($p = 0.85$) or a TTF-rendered glyph ($p = 0.15$; split is qualitative and should be benchmarked). Each glyph g is resized to height $H \in \{64, 96\}$ with vertical jitter $\delta \sim \mathcal{U}(0, H - h_g)$. The line $L = \text{concat}(g_1, \ldots, g_n)$ is augmented via $I_{synth} = \mathcal{T}(\text{Pad}(L))$, where $\mathcal{T}$ applies morphological filters, Gaussian noise $\mathcal{N}(0, \sigma^2)$, and elastic deformations, generating 134,528 lines.

Font-only Vocabulary Expansion A complementary set of 100,000 clean font-rendered lines prevents bias toward synthesis artifacts. Stroke weight is modulated ($W \in \{\text{bold, normal, thin}\}$ via dilation/erosion kernels $k \in \{2, 3\}$), ink wear simulated (white noise density $\rho \in [0.01, 0.04]$), and character overlap allowed via $C_{x:x+w} = \min(C_{x:x+w}, g)$ to preserve stroke continuity. Final output: $I_{clean} = \text{Otsu}(\text{Blur}(L, \sigma))$.

3.3. Annotation

We re-annotated half of Dataset B to better reflect page scan content, introducing two finer-grained categories —"text-horiz" for horizontal text and "text-one" for single characters—to address the heterogeneity of the broad "text" class (Dataset B⋆ from here on). We also employed oriented bounding boxes to capture objects in non-standard layouts, such as circular character arrangements. For images, overlapping annotations were used when-

ever sub-images were nested within larger ones to preserve structural information, and exhaustive coverage of all categories was ensured (see Figure 4). While these annotations more faithfully represent the data, they increase learning difficulty for layout model (see Table 3).

3.4. Closed modality - Strategy

3.4.1. Tasks A and C

We trained two line-level recognition models. First, a custom `CRNN` (`PyTorch`) combining a convolutional backbone (three GroupNorm conv blocks expanding to 256 filters with max-pooling) and two bidirectional LSTM layers (hidden size 450, dropout), optimized with CTC loss and aggressive `Albumentations` augmentation (blur, noise, contrast, dropout, channel shuffling, embossing, grayscale). Second, `PaddleOCR v5` with `SVTR_HGNet` architecture (PPHGNetV2-B4 backbone, multi-scale SVTR neck), trained with combined CTC+NRTR loss and built-in augmentation (RecAug, RecConAug, RandAugment, multi-scale sampling).

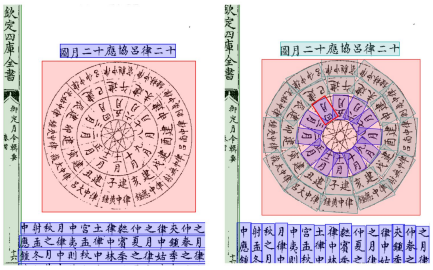


Figure 4: Example of re-annotation in dataset B*.

Both followed a two-stage strategy: (1) foundation-model training on 200,000+ `UltraGlyph` synthetic lines (Section 3.2, using only organizer-provided data); (2) fine-tuning on task-specific datasets, enabling adaptation from synthetic to real documents while retaining broad character coverage.

3.4.2. Task B

For layout detection, we fine-tuned three architectures on the re-annotated Dataset B⋆ (see Section 3.3): **YOLO OBB** (YOLOv8l-obb) for oriented bounding boxes, and **YOLO12s** and **RT-DETR** (large) for axis-aligned bounding boxes. The annotated set (2,308 images, 6 categories; see Table 1) was split 80/20 into train and validation sets (see Table 5 for hyperparameters).

YOLO OBB was trained for 200 epochs on the pre-trained `yolov8l-obb` weights. For YOLO12s, training employed a two-phase schedule (400 epochs total): box loss weight was increased in the first 50 epochs to prioritize localization, then classification weight was raised for the remaining epochs. RT-DETR large was also evaluated as an additional baseline. Full hyperparameter settings are reported in Appendix C.

Label	Count	Label	Count
ext	9,886	text_horiz	3,168
image	3,348	text_one	4,852
book edge	1,825	seal	230

Table 1: Label distribution after re-annotation of Dataset B*.

3.5. Open task - Strategy (tasks A and C)

We evaluated several VLMs including `Qwen 2.5-VL 72B`, `Qwen3-VL 8B and 30B-A3B`,`Gemini 3 Flash`, `Claude Sonnet 4.5` and `GPT-4o`[1]. All models were used with temperature 0 and fixed seed. Each model is governed by a specialized prompt structure (See AppendixA, Figure 7) enforcing a visual-first transcription protocol that explicitly prohibits semantic auto-correction or modern normalization of character forms.

Two dimensions of the in-context learning framework were investigated: the number of shots required for adequate performance, and the effect of exemplar selection strategy. We compared a **random baseline**, in which k shots ($k \in \{5, 10, 25\}$) are drawn stochastically from the training pool, against a **variant-aware dynamic selection** that ranks candidates by their overlap with the non-standard character variants present in the target input, formally defined as $\text{Score}(E) = |V_{\text{target}} \cap V_{\text{candidate}}|$. Together, these two axes allow us to assess each model's sensitivity to both the quantity and contextual relevance of in-context exemplars.

3.6. Metrics

For Tasks A and C, we evaluate OCR quality using CER, NED, and character-level F1. For Task B, we use mAP@[0.5:0.95]. EvaHan defines a single comprehensive score for Tasks A and C:

$$S = 0.5 \cdot (1 - \text{CER}) + 0.3 \cdot \text{F}_1 + 0.2 \cdot (1 - \text{NED})$$

where CER penalizes character-level errors, F_1 balances precision and recall at the character level, and NED penalizes length discrepancies between predicted and reference transcriptions.

[1] Llama 4 Maverick, Mistral Large, Bytedance Seed 1.6 Flash, and InternVL2.5-78B were initially included in the evaluation but produced incomplete results across sampling conditions and were therefore excluded from the final benchmark.

4. Results

4.1. Tasks A and C

Table 2 reports the best-performing configuration per system on Datasets A and C; full per-configuration results and ablation studies are provided in Appendix D and Figure 11.

Closed-modality models perform strongly on Dataset A but degrade markedly on Dataset C. `CRNN-FT` reaches $S_A = 0.963$ (CER $= 0.038$) and $S_C = 0.782$ (CER $= 0.221$), marginally outperforming its foundation model but showing weak transfer to handwriting. `PaddleOCR`, with its multi-scale SVTR neck and CTC+NRTR loss, is more resilient: it trails on Dataset A ($S_A = 0.947$, CER $= 0.056$) but achieves the best closed-modality score on Dataset C ($S_C = 0.852$, CER $= 0.151$), a 31.7% relative CER reduction over `CRNN-FT`, suggesting that multi-head training and multi-scale features confer stronger generalization to cursive strokes.

Model	Norm.	Dataset A				Dataset C			
		Config	Score↑	CER↓	F1↑	Config	Score↑	CER↓	F1↑
Closed modality									
CRNN-FT	norm.	–	**0.9632**	0.0382	0.9666	–	0.7823	0.2212	0.7903
CRNN-FT	raw	–	**0.9350**	0.0663	0.9381	–	0.7513	0.2520	0.7590
CRNN (found.)	norm.	–	0.9623	0.0391	0.9655	–	0.7828	0.2212	0.7916
CRNN (found.)	raw	–	**0.9350**	0.0663	0.9381	–	0.7432	0.2607	0.7517
PaddleOCR	norm.	–	0.9465	0.0558	0.9515	–	**0.8520**	0.1509	0.8565
PaddleOCR	raw	–	0.9191	0.0832	0.9241	–	**0.8130**	0.1900	0.8176
Open modality									
PaddleOCR	norm.	v5	**0.9688**	0.0317	0.9698	v4	0.8927	0.1081	0.8946
PaddleOCR	raw	v5	0.9433	0.0572	0.9443	v4	0.8430	0.1578	0.8448
Gemini 3-Flash	norm.	0s	0.9612	0.0388	0.9612	0s	0.9360	0.0652	0.9384
Gemini 3-Flash	raw	dyn-5	0.9537	0.0466	0.9543	0s	0.9090	0.0923	0.9115
Qwen2.5-VL-72B	norm.	dyn-25	0.9581	0.0430	0.9604	0s	**0.9499**	0.0503	0.9503
Qwen2.5-VL-72B	raw	dyn-25	0.9501	0.0510	0.9523	0s	**0.9276**	0.0727	0.9280
Qwen3-VL-30B	norm.	dyn-25	0.9651	0.0358	0.9669	dyn-10	0.9147	0.0861	0.9162
Qwen3-VL-30B	raw	dyn-25	**0.9632**	0.0376	0.9651	0s	0.8884	0.1125	0.8904
Qwen3-VL-8B	norm.	dyn-5	0.9681	0.0321	0.9684	0s	0.9408	0.0595	0.9415
Qwen3-VL-8B	raw	dyn-5	0.9626	0.0376	0.9629	rnd-5	0.9210	0.0797	0.9226

Table 2: Best OCR performance per model and normalization on Datasets A and C. Each row reports the best-scoring configuration for that model. Closed-modality models have no ICL (–). For open-modality LLMs, **Config** gives the best ICL strategy and shot count across $k \in \{5, 10, 25\}$: *0s* = zero-shot, *dyn-k* = variant-aware dynamic, *rnd-k* = random. For PaddleOCR, the best checkpoint version (v1–v5) may differ between datasets.

In the open modality, `PaddleOCR` v5 (norm.) leads on Dataset A ($S = 0.969$, CER $= 0.032$), confirming that external data yields consistent but modest gains on printed text. On Dataset C, v4 reaches $S = 0.893$, outperforming all closed systems, likely due to broader cursive style coverage.

Among VLMs, `Gemini 3-Flash` and the `Qwen3-VL` family (8B and 30B-A3B) are competitive on Dataset A ($S \approx 0.96$–0.97), on par with dedicated OCR systems. `Qwen3-VL-8B` matches or slightly outperforms its 30B-A3B counterpart on both datasets, suggesting that architecture and Chinese pre-training coverage matter more than parameter count. `Qwen2.5-VL-72B` trails Qwen3 on Dataset A but leads all LLMs on Dataset C at zero-shot ($S = 0.950$), indicating stronger generalization to cursive handwriting. `GPT-4o` and `Claude Sonnet 4.5` lag considerably on both datasets ($S_A \approx 0.81$ and 0.76; $S_C \approx 0.66$ and 0.47), reflecting limited historical Chinese script coverage in their pre-training (see Appendix, Figure 11).

Impact of ICL strategy and shot count. ICL contributions are modest and strongly model-dependent (Figures 5–6). Few-shot prompting produces substantial CER reductions for weaker models — `Claude 4.5` gains ΔCER ≈ 0.09–0.10 on Dataset A and 0.17–0.19 on Dataset C, `Qwen2.5-VL-72B` up to 0.084 on Dataset A — but yields neutral or slightly negative effects for already-strong models (`Gemini 3-Flash`, `Qwen3-VL`), which appear saturated by exemplar injection. Exemplar selection strategy induces only marginal differences ($|\Delta\text{CER}| < 0.015$), with no consistent winner across model families. Taken together, pretraining quality largely supersedes ICL for strong models, while weaker models benefit primarily from script exposure rather than selection strategy.

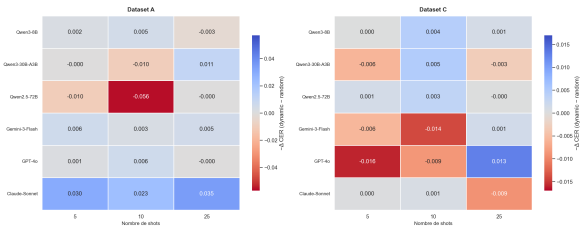


Figure 5: CER differential between Variant-Aware Dynamic and random ICL selection ($\Delta S = S_{\text{dynamic}} - S_{\text{random}}$) across models, datasets, and shot counts. Blue: dynamic outperforms random; red: the reverse.

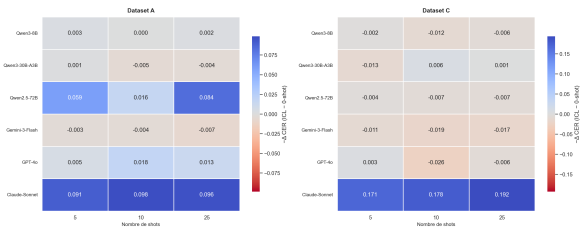


Figure 6: CER differential of ICL over zero-shot inference ($\Delta S = S_{\text{ICL}} - S_{\text{0shot}}$) per model, dataset, and shot count. Blue cells indicate that zero-shot outperforms few-shot ICL.

Taken together, these results indicate that for strong models the quality of pre-training largely su-

persedes the contribution of in-context exemplars, while weaker models benefit primarily from exposure to the target script rather than from the selection strategy itself.

Qualitative Assessment See Table 8 and Figure 9 in appendix for a qualitative display. Open models are considerably more resilient to variation and damage. In Dataset A, images deviating from standard printed editions may concentrate errors (a_040, a_190, possibly due to manuscript-style handwriting) or hinder recognition altogether (a_194, possibly due to color interference). A similar pattern appears in Dataset C (c_189), with additional phenomena: punctuation inserted by the closed model, likely from training data contamination, and heightened sensitivity to poorly formed characters such as 月. When horizontal page strokes could plausibly belong to a character, the closed model tends to incorporate them, producing spurious readings (c_175). Both models, as well as human annotators, are further challenged by character variants—including simplified forms—within frequent classes such as 關, 税, 内, 两, and 并. When errors occur, the model typically identifies sub-graphemic components of the target character, as illustrated by the 該/設 confusion in c_150. Both models generally succeed in isolating individual characters, even in Dataset C where density varies and peripheral strokes risk corrupting segmentation. While the open model clearly outperforms the closed model on Dataset C, some of its errors may stem from large-scale modern Chinese training data, as suggested by a possible translation error in c_003 and a character insertion in c_172. The open model also shows reduced robustness when recognition is repeated.

4.2. Task B

Table 3 reports per-class detection performance. **YOLO12s** outperforms **RT-DETR** on all aggregate metrics (mAP50: 0.750 vs. 0.723; mAP50-95: 0.486 vs. 0.463). Both models perform strongly on well-represented, visually distinct categories (*book_edge*, *seal*: mAP50 ≥ 0.99), but degrade on *text-horiz* (mAP50: 0.402 and 0.339). `YOLO12s` trained on original annotations scores higher (mAP50: 0.861, mAP50-95: 0.578), but this is partly misleading: those annotations are less exhaustive for textual elements (Fig. 10), so the apparent gain may reflect lower annotation expectations rather than genuine improvement.

`YOLO OBB`, fine-tuned on the reannotated dataset, outperforms both models on aggregate metrics with notable gains on textual classes, suggesting that oriented bounding boxes better capture textual layout patterns and that exhaustive reannotation benefits detection.

Class	P	R	mAP50	mAP50-95
YOLO12s				
All	0.791	0.731	**0.750**	**0.486**
book_edge	0.988	0.991	0.994	0.760
image	0.890	0.714	0.780	0.477
seal	0.939	1.000	0.995	0.782
text	0.881	0.611	0.787	0.444
text-horiz	0.443	0.488	0.402	0.212
text-one	0.606	0.581	0.543	0.240
all “text”	0.764	0.591	0.684	0.366
YOLO OBB				
All	0.785	0.809	0.808	0.583
book_edge	0.491	0.664	0.443	0.050
image	0.893	0.880	0.894	0.619
seal	0.911	0.976	0.982	0.943
text	0.822	0.777	0.864	0.652
text-horiz	0.759	0.753	0.803	0.631
text-one	0.834	0.801	0.859	0.603
RT-DETR large				
All	0.794	0.707	0.723	0.463
book_edge	0.969	0.979	0.992	0.723
image	0.867	0.647	0.740	0.475
seal	0.964	1.000	0.995	0.748
text	0.875	0.609	0.751	0.431
text-horiz	0.440	0.446	0.339	0.180
text-one	0.648	0.560	0.521	0.224
YOLO12s on original dataset				
All	0.877	0.810	0.861	0.578
book_edge	0.921	0.970	0.983	0.685
image	0.776	0.667	0.720	0.474
seal	0.953	1.000	0.995	0.759
text	0.858	0.602	0.747	0.396

Table 3: Per-class detection results on Dataset B and Dataset B* validation set.

5. Conclusion

We presented a complete pipeline for EvaHAN 2026, combining `UltraGlyph` synthetic data (234,528 lines: 134,528 hybrid, 100,000 font-only), two-stage OCR pretraining, and re-annotation of 2,308 Dataset B images with 6 refined categories and oriented bounding boxes. Closed-modality CRNN-FT reaches 96.3% on Dataset A but degrades to 78.2% on Dataset C; PaddleOCR proves more resilient (94.7% / 85.2%). In open modality, PaddleOCR v5 leads on printed text (96.9%) while `Qwen2.5-VL-72B` zero-shot leads on cursive (95.0%). ICL yields marginal gains for strong models (ΔCER < 0.015) but substantial improvements for weaker ones (`Claude 4.5`: +17–19% on Dataset C); variant-aware selection shows no consistent advantage over random sampling. These results confirm that pretraining quality and data coverage dominate performance, with ICL primarily benefiting underspecified models.

6. Acknowledgements

We thank the Evahan2026 organizing team for providing the shared task and the associated data. This study was conducted as part of the DH master's program at École nationale des chartes–PSL[2]. It also received support from the PSL Research University's Major Research Program CultureLab, implemented by the ANR (reference ANR-10-IDEX-0001). We used Claude AI to help condense the paper and proofread the English. Any remaining mistakes are our own.

7. Code and models availability

All codes and models are available on https://github.com/Bizais-Lillig/enchanteam.

8. Bibliographical References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-vl technical report.

Marie Bizais-Lillig. 2025. Transcrire des éditions impériales chinoises pour acquérir un corpus numérique : Questions, choix et réflexions. In Alix Chagué, Thibault Clérice, and Ariane Pinche, editors, *Apprendre à Lire aux Machines*, pages 121–128.

Marie Bizais-Lillig, Chahan Vidal-Gorène, and Boris Dupin. 2024. Optimizing HTR and Reading Order Strategies for Chinese Imperial Editions with Few-Shot Learning. In *Document Analysis and Recognition – ICDAR 2024 Workshops*, volume 14936 of *Lecture Notes in Computer Science*, pages 37–56. Springer Nature Switzerland.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, pages 1877–1901.

Bo Chang, Qiong Zhang, Shenyi Pan, and Lili Meng. 2018. Generating handwritten chinese characters using cyclegan. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 199–207. IEEE.

Yan Hon Michael Chung and Donghyeok Choi. 2025. Finetuning vision-language models as ocr systems for low-resource languages: A case study of manchu.

Cheng Cui, Ting Sun, Suyin Liang, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Xueqing Wang, Changda Zhou, Hongen Liu, Manhui Lin, et al. 2025. Paddleocr-vl: Boosting multilingual document parsing via a 0.9 b ultra-compact vision-language model. *arXiv preprint arXiv:2510.14528*.

S. Gupta, T. Xie, and D. Roth. 2023. Coverage-based example selection for in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, pages 13924–13935.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR 2022)*.

G. Kim, T. Hong, M. Yim, J. Nam, J. Park, and S. Kim. 2022. OCR-free document understanding transformer. In *Proceedings of the European Conference on Computer Vision (ECCV 2022)*, pages 498–517.

Minghao Li, Tengchao Lv, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, and Furu Wei. 2021. TrOCR: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2021)*.

Y. Li, Y. Li, Y. Wang, and L. Wang. 2022. DiT: Self-supervised pre-training for document image transformer. In *Proceedings of the 17th International Conference on Document Analysis and Recognition (ICDAR 2022)*.

Yang Liu, Jiahuan Cao, Hiuyi Cheng, Yongxin Shi, Kai Ding, and Lianwen Jin. 2025. MCS-bench: A comprehensive benchmark for evaluating multimodal large language models in Chinese classical studies. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages

[2] https://www.chartes.psl.eu/formations/masters/master-humanites-numeriques

10435–10492, Vienna, Austria. Association for Computational Linguistics.

Y. Luo, Z. Chen, Y. Wang, and Z. Zhang. 2024. In-context learning with retrieved demonstrations for language models: A survey. *Transactions of the Association for Computational Linguistics*, 12:1–24.

Weihong Ma, Hesuo Zhang, Lianwen Jin, Sihang Wu, Jiapeng Wang, and Yongpan Wang. 2020. Joint layout analysis, character detection and recognition for historical document digitization.

Sina Semnani, Han Zhang, Xinyan He, Merve Tekgürler, and Monica Lam. 2025. Churro: Making history readable with an open-weight large vision-language model for high-accuracy, low-cost historical text recognition. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34765–34812.

Baoguang Shi, Xiang Bai, and Cong Yao. 2015. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2298–2304.

Chahan Vidal-Gorène, Boris Dupin, Aliénor Decours-Perez, and Thomas Riccioli. 2021. A modular and automated annotation platform for handwritings: evaluation on under-resourced languages. In *International Conference on Document Analysis and Recognition*, pages 507–522. Springer.

Chahan Vidal-Gorène, Bastien Kindt, and Florian Cafiero. 2026. Under-resourced studies of under-resourced languages: lemmatization and pos-tagging with llm annotators for historical armenian, georgian, greek and syriac. *arXiv preprint arXiv:2602.15753*.

Dongbo Wang and Dongmei Zhu. 2025. Benchmarking the ancient books capability of multimodal large language models. *npj Heritage Science*, 13(1):339.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. LayoutLM: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2020)*, pages 1192–1200.

Haiyang Yu, Yuchuan Wu, Fan Shi, Lei Liao, Jinghui Lu, Xiaodong Ge, Han Wang, Minghan Zhuo, Xuecheng Wu, Xiang Fei, et al. 2025. Benchmarking vision-language models on chinese ancient documents: From ocr to knowledge reasoning. *arXiv preprint arXiv:2509.09731*.

Yuyi Zhang, Yongxin Shi, Peirong Zhang, Yixin Zhao, Zhenhua Yang, and Lianwen Jin. 2025. Megahan97k: A large-scale dataset for mega-category chinese character recognition with over 97k categories. *Pattern Recognition*, 167:111757.

Dongmei Zhu, Chang Liu, Xue Zhao, Zhixiao Zhao, Si Shen, and Dongbo Wang. 2025. Xunzi-mllm: a multimodal large language model for ancient text and image recognition. *Digital Scholarship in the Humanities*, 40(2):709–722.

Zhuang Deming 莊德明. 2001. Zhongwen diannao quezi jiejue fang'an 中文電腦缺字解決方案 (report on a method to solve the lack of characters on computers in chinese language). In 第一屆中國文字學國際學術研討會 *(The First International Conference on the Study of Chinese characters*.

9. Language Resource References

Bizais-Lillig, M. 2024. *HTR ground-truth of Chinese xylographic editions*. Document Analysis and Recognition –ICDAR 2024 Workshops. ICDAR 2024. Lecture Notes in Computer Science (v1.0.0, p. 37-56), 1.0. PID /10.5281/zenodo.14452717.

Lin,Zih-Ci. 2025. *Traditional Chinese OCR Synthetic*.

Weihong Ma, Hesuo Zhang, Lianwen Jin, Sihang Wu, Jiapeng Wang, Yongpan Wang. 2020. *Joint Layout Analysis, Character Detection and Recognition for Historical Document Digitization*.

Appendices

A. Prompt for ICL

System Prompt (Paleographic Persona):
You are a professional paleographer specializing in Chinese. Your goal is a character-for-character high-fidelity transcription.
Rules:
1. DO NOT CORRECT.
2. NO SEMANTIC ADAPTATION.
3. VISUAL PRIORITY.
4. Output ONLY characters.

Few-Shot Examples (k iterations):
User: Transcribe this image: [Image_i]
Assistant: [Gold_Transcription_i]

Target Input:
User: Transcribe this text. Output only the transcription. [Target_Image]

Figure 7: Schematic representation of the interleaved multi-modal prompt structure.

B. Lines used per dataset

Dataset	Total	Ratio	Used
Dataset A	5,000	100%	5,000
Dataset C	5,000	100%	5,000
ChiKnowPo	28,551	100%	28,551
Guangdong	50,014	45%	22,506
MTHv2	105,579	85%	89,742
Traditional Synth.	50,000	60%	30,000
UltraGlyph (clean)	100,000	85%	85,000
UltraGlyph (synth.)	134,528	50%	67,264
Total	**478,672**	—	**333,063**

Table 4: Line extraction and sampling configuration across all datasets.

C. YOLO Hyperparameters

Parameter	Epochs 1–50	Epochs 51–400
batch / imgsz	8 / 1024	8 / 1024
patience	30	30
dropout / iou	0.3 / 0.6	0.3 / 0.6
cos_lr / half	True / True	True / True
multi_scale	0.15	0.15
box loss weight	8.5	7.5
cls loss weight	1.5	2.5
hsv_s / hsv_v	0.8 / 0.8	0.8 / 0.8
degrees / fliplr / scale	0.2 / 0.3 / 0.1	0.2 / 0.3 / 0.1

Table 5: YOLO training hyperparameters.

D. Ablation Study: Incremental Training Data for PaddleOCR

We conducted a cumulative ablation adding training corpora incrementally across five configurations (Table 6, Figure 8).

Config	Added corpus	Dataset A		Dataset C	
		Score	CER	Score	CER
v1	Official only + UltraGlyph	0.772	0.232	0.681	0.324
v2	+ ChiKnowPo	0.963	0.038	0.870	0.130
v3	+ Guangdong	0.963	0.038	0.860	0.141
v4	+ MTH	0.969	0.031	**0.893**	**0.108**
v5	+ SynthHF	**0.969**	**0.032**	0.871	0.130

Table 6: Ablation results for open PaddleOCR (normalized). Each configuration accumulates the previous corpora.

The dominant gain is produced by the transition from v1 to v2: adding ChiKnowPo reduces CER from 0.232 to 0.038 on Dataset A and from 0.324 to 0.130 on Dataset C, a drop of over 85% and 60% respectively.

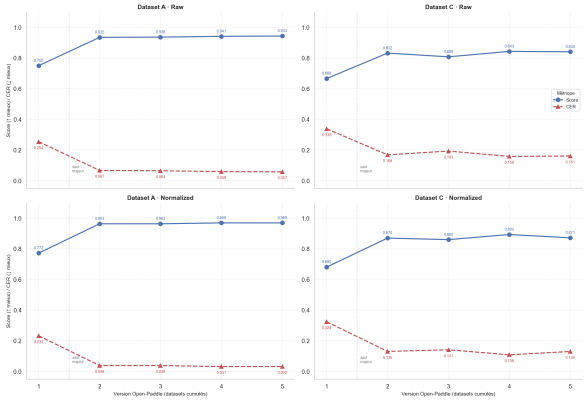


Figure 8: Incremental performance of open PaddleOCR configurations on Datasets A and C (normalized). Each configuration adds one external corpus to the previous, following the order described in the text.

Subsequent additions yield more nuanced effects: Guangdong (v3) leaves Dataset A unchanged while marginally degrading Dataset C (S : $0.870 \rightarrow 0.860$); MTH (v4) peaks on both datasets ($S_C = 0.893$, $S_A = 0.969$); SynthHF (v5) maintains Dataset A but reverts Dataset C to near-v2 levels ($S = 0.871$). Beyond v2, gains are modest and non-monotonic, with optimal configurations differing: v4 for Dataset C, v4–v5 (tied) for Dataset A.

E. Layout predictions

Figure 10 displays layout predictions with and without re-annotation.

F. Performance Barplots

Figure 11 presents the full per-configuration score distributions for all models, datasets, and modalities, ordered by comprehensive score (normalized evaluation condition).

G. Character Recognition Error Types in Dataset A and Dataset C

Across both open- and closed-modality models, character recognition errors (CR) dominate: 63.8% of all errors in open-modality (1,689/2,647) and 66.3% in closed-modality (126/190), as illustrated in Figure 9. This consistent pattern across architectures confirms that character-level recognition is the primary bottleneck, rather than insertion, deletion, or truncation.

G.1. Character Variant Effects in Dataset A and Dataset C and Potential OCR Improvements

Metric	Value
Total characters analyzed	40,388
Correct characters	33,846
Variant-character matches	2,496
True recognition errors	4,046
Variant share of non-exact differences	38.15%

Table 7: Variant-based breakdown of character recognition outcomes in the evaluated datasets. A substantial portion of non-exact differences arises from character variants rather than true OCR recognition errors.

Incorporating variant normalization could potentially improve the combined recognition accuracy of open- and closed-modality models on Dataset A and Dataset C by up to 38.15% beyond the current baseline (Table 7). To assess this, we used the variant.json file provided by the EVAHAN benchmark organizers to examine whether character recognition errors were caused by orthographic variation. Ambiguous many-to-many mappings were removed, retaining only cases where one standard character corresponds to one or multiple variant forms. Using this cleaned mapping, we rechecked 1,813 character recognition errors from both modalities. In total, 40,388 characters were examined: 33,846 exact matches, 2,496 variant matches, and 4,046 genuine recognition errors. Variant matches account for 38.15% of all non-exact differences.

Table 8 displays qualitative error analysis for selected examples from the test datasets.

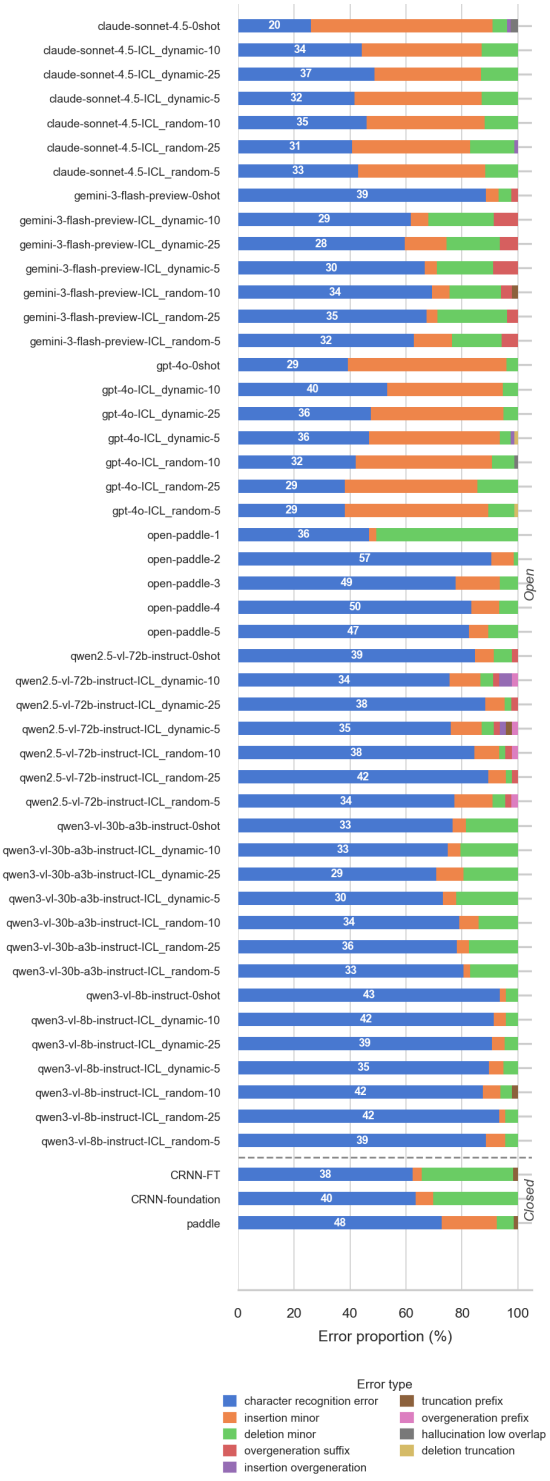


Figure 9: OCR Error Type Distribution Across Models and Datasets.

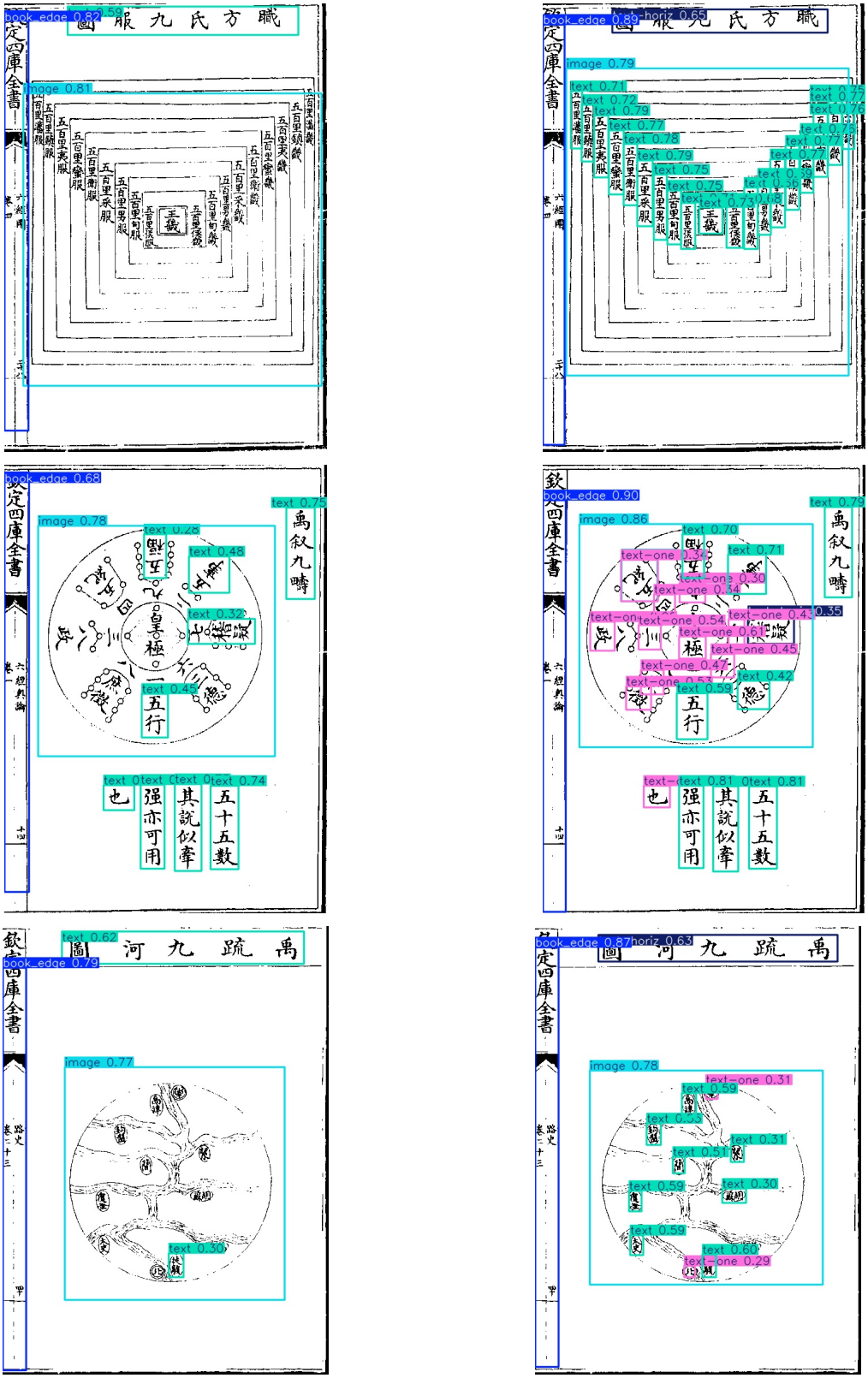


Figure 10: Predictions using YOLO12s model. On the left, trained on original annotations. On the right, train using the Dataset B⋆ annotations.

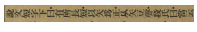
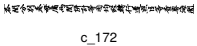
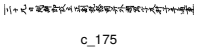
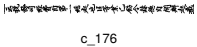
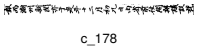
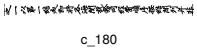
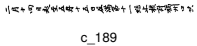
Image	Model	Text
		Dataset A
a_040	**GT**	監察御史眞聖奴書丹大中大夫陝西奉元路達魯花赤呵剌
	CRNN-FT (CER: 0.0%)	監察御史眞聖奴書丹大中大夫陝西奉元路達魯花赤呵剌
	open-paddle-5 (CER: 0.0%)	監察御史眞聖奴書丹大中大夫陝西奉元路達魯花赤呵剌
	qwen3-vl-30b (CER: 0.0%)	監察御史眞聖奴書丹大中大夫陝西奉元路達魯花赤呵剌
a_190	**GT**	此避高祖諱猶楊淵之稱楊泉非一字泉明也
	CRNN-FT (CER: 22.2%)	此避高祖諱齋楊調之楊泉非一字矣明也
	open-paddle-5 (CER: 0.0%)	此避高祖諱猶楊淵之稱楊泉非一字泉明也
	qwen3-vl-30b (CER: 5.6%)	此避高祖諱猶楊淵之稱楊泉非一字泉明也
 a_194	**GT**	說文短字下曰有所長短以矢爲正從矢豆聲段氏曰當云
	CRNN-FT (CER: 65.2%)	說文短下有所曇矢爲
	open-paddle-5 (CER: 17.4%)	說文短字下曰有所長短以矢為正从矢豆聲 段氏曰 當云
	qwen3-vl-30b (CER: 8.7%)	說文短字下曰有所長短以矢為正从矢豆聲段氏曰當云
		Dataset C
c_003	**GT**	任之日起至十一月十一日止計征各國洋商進出口稅銀二十
	paddle (CER: 12.0%)	任之日赳至十一月十一日止計征公國洋商道出口稅銀二十
	qwen2.5-vl-72b (CER: 8.0%)	任之日起到十一月十一日止計征各國洋商進出口稅銀二十
c_150	**GT**	一片奉旨該衙門議奏欽此據原奏內稱前准總理各國事務衙
	paddle (CER: 12.0%)	一片奉者設衙門議奏欽此據原奏內稱前准總理谷國事務衙
	qwen2.5-vl-72b (CER: 8.0%)	一片奉有該衙門議奏欽此據原奏內稱前准總理各國事務衙
 c_172	**GT**	各関分別奏毋庸兩関併計等因均經轉行遵照臣等查粤海関
	paddle (CER: 20.0%)	谷関分別奏母庸兩関并計等因均經轉行遵照臣等查粵海閣
	qwen2.5-vl-72b (CER: 12.0%)	各関分別奏毋庸兩関併計等因均經轉行遵照臣等查粵江海關
 c_175	**GT**	二十九日開辦即設立正副稅務司并外國寫字及杆子手通事
	paddle (CER: 28.0%)	三十兀日開辨即設立正副稅務司，并外國冩字及許子手通畫
	qwen2.5-vl-72b (CER: 4.0%)	二十九日開辦即設立正副稅務司并外國寫字及杆子手通事
 c_176	**GT**	云稅務司經費自第一結起也臣等業已飭令赫德自開辦起截
	paddle (CER: 36.0%)	云。稅務司經帶自第一始走也佳等業已。飭令赫德自開辨超，截
	qwen2.5-vl-72b (CER: 8.0%)	云稅務司經費自第一結起也目等業已飭令赫德自開辦起截
 c_178	**GT**	報而潮州新関亦于是年十二月初九日由省前往開辦議設稅
	paddle (CER: 12.0%)	報而潮州新関亦于是平十二月初九日由省前往開辨議設祝
	qwen2.5-vl-72b (CER: 8.0%)	報而潮州新閩亦于是年十二月初九日由省前往開辦議設稅
 c_180	**GT**	之一以第一結起即將各海関稅務司經費順序按結開列并非
	paddle (CER: 12.0%)	之一以第一結起即將谷海関祝務司經當順序按結開列并非
	qwen2.5-vl-72b (CER: 8.0%)	之一以第一結起即將各海関稅務司經費順序按結開列并非
 c_189	**GT**	二月十四日起至五月十五日屆滿第十一結止核計福州口共
	paddle (CER: 20.0%)	二月十四日起至五舟十五日履滿第十一結止，核計稿州口其
	qwen2.5-vl-72b (CER: 4.0%)	二月十四日起至五月十五日在滿第十一結止核計福州口共

Table 8: Qualitative error analysis on Datasets A and C. Blue: character variants; yellow: variant-standard mismatches; red: recognition errors.

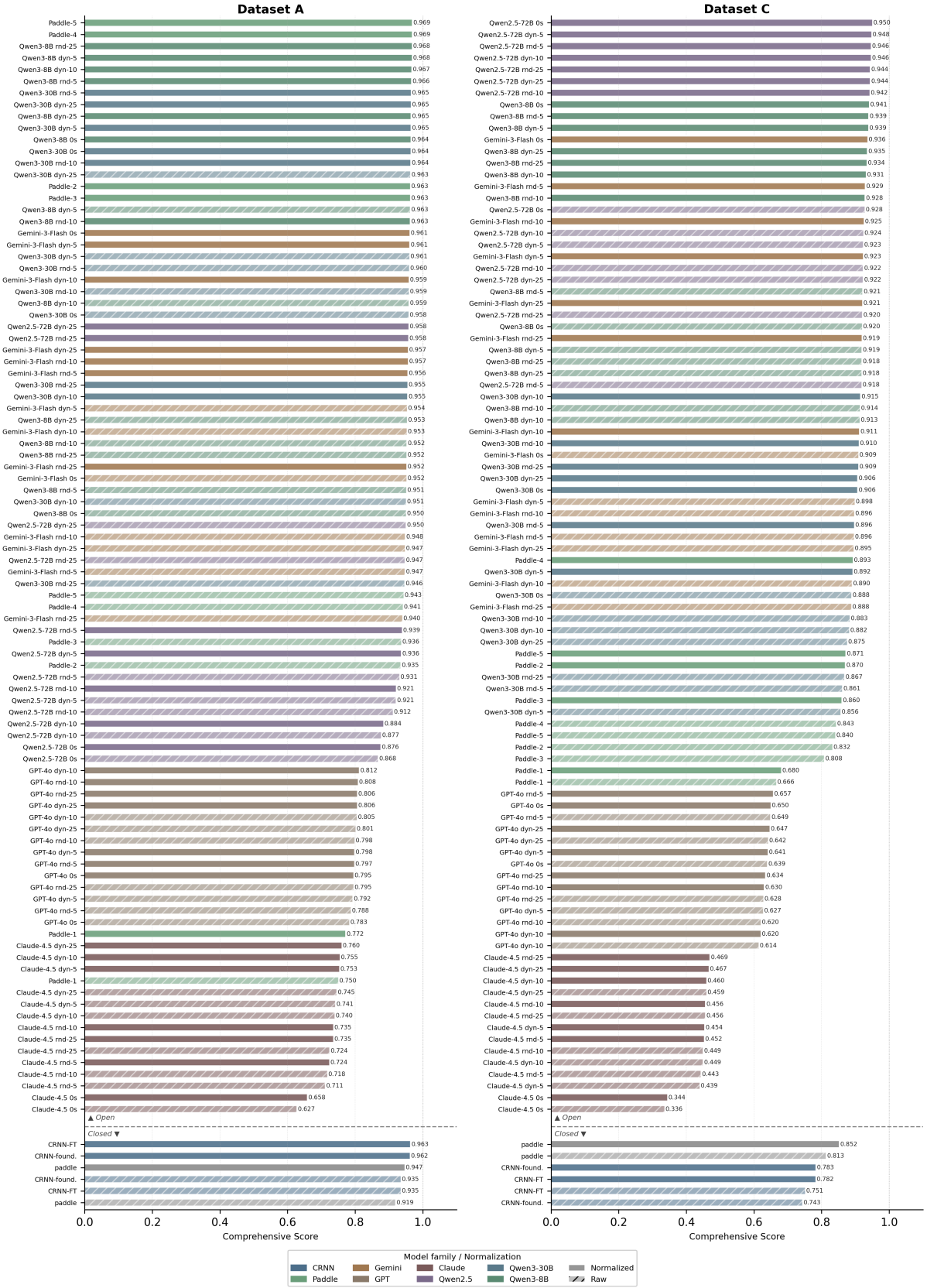


Figure 11: All results, ordered by score

The UD_Latin-PROIEL as Linked Open Data: Integrating a Latin Treebank into the LiLa Knowledge Base

Lucas Dezotti[1], Marco Passarotti[2], Federica Iurescia[2], Giovanni Moretti[2]
[1] Universidade Federal da Paraíba, [2] Università Cattolica del Sacro Cuore
[1] 58051-900, João Pessoa, Brazil, [2] Largo Gemelli 1, 20123 Milan, Italy
lucas.dezotti@academico.ufpb.br, {marco.passarotti, federica.iurescia, giovanni.moretti}@unicatt.it

Abstract

This paper presents the steps taken to integrate data from the UD_Latin-PROIEL treebank into the LiLa Knowledge Base of interoperable linguistic resources for Latin. It describes how the lexical, morphological, syntactic, and citation information from the source was modeled using the Linked Open Data principles as adopted by the LiLa Knowledge Base. The process of linking tokens to the LiLa collection of Latin lemmas is detailed, addressing challenges such as ambiguities, new lemmas, and errors encountered in the source. The outcome is a syntactically annotated textual resource that is interoperable with the (meta)data of other Latin linguistic resources linked within the LiLa Knowledge Base. This integration enables new ways of analyzing linguistic information and using the content as a starting point to explore connections with other interlinked resources. A use case demonstrates this interoperability.

Keywords: Linked Open Data, Linguistic Annotation, Universal Dependencies treebank

1. Introduction

Over the past two decades, several Latin treebanks — corpora annotated with syntactic structures, morphological information, and lemmatization — have been created. Among the earliest were the Latin Dependency Treebank (LDT) (Bamman and Crane, 2006) and the Index Thomisticus Treebank (IT-TB) (Passarotti, 2019). Subsequent projects include the PROIEL Treebank (Haug and Jøhndal, 2008), the Late Latin Charter Treebank (Korkiakangas and Passarotti, 2011), UDante (Passarotti et al., 2021) and, more recently, UD_Latin-CIRCSE (Iurescia et al., 2025). Together, these resources sample Latin from a broad arc of time, stretching from Classical Antiquity to the late Middle Ages.

Treebanks have become a key resource in linguistic research and natural language processing (Nivre, 2008), enabling reproducible experiments and supporting a wide range of computational and theoretical studies (Bamman and Crane, 2011). Latin treebanks have underpinned research in morphosyntactic theory (Haug, 2015), lexicography (McGillivray and Vatri, 2015), textual similarity (Bamman and Crane, 2008), authorship style (Cecchini and Pedonese, 2022), parser development and evaluation (Burns, 2023), and language teaching (Mambrini, 2016).

The need for a common standard for the syntactic annotation of Latin texts has been recognized from the beginning, as evidenced by the shared annotation guidelines developed for the LDT and IT-TB projects (Bamman et al., 2007). A major step forward came with the Universal Dependencies (UD) project (Nivre et al., 2016), which provides a framework for cross-linguistically consistent morphosyntactic annotation and has become the de facto standard for many treebanks. Consequently, Latin treebanks originally produced under different annotation schemes have been converted, at least partially, into the UD format. Although conversion can be challenging (Cecchini et al., 2018, 2020) and often requires further refinement and harmonization to ensure data consistency (Gamba and Zeman, 2023b,a), it represents an important step toward compatibility among corpora produced by different projects.

A further step is to make Latin treebanks interoperable with other linguistic resources. This can be achieved by publishing them as Linked Open Data in the LiLa Knowledge Base (Passarotti et al., 2020), a collection of textual and lexical resources for Latin. Currently, the corpora linked to the LiLa Knowledge Base total nearly 12 million tokens; however, only two of these corpora are annotated at the syntactic level, amounting to 430,000 tokens: the IT-TB (375,000) and UDante (55,000). Despite their undeniable importance, these resources mainly represent a specific variety of Latin, namely Medieval Latin from the 13th and 14th centuries. This highlights a gap in the linked corpora with respect to diachronic representativeness.

To help close this gap, this paper describes the steps taken to link the UD_Latin-PROIEL treebank to the LiLa Knowledge Base. It focuses on how the lexical, morphological, and syntactic annotations, as well as citation information, are modelled using the Linked Open Data principles as adopted in LiLa. The remainder of the paper is organized as follows: (a) an overview of the LiLa KB structure,

Proceedings of LT4HALA 2026 @ LREC 2026, pages 353–360.
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

especially the model used to publish annotated textual resources as Linked Open Data; (b) the main characteristics of the UD_Latin-PROIEL treebank; (c) the process of modelling and linking tokens to the LiLa collection of Latin lemmas, including issues related to ambiguity, new lemmas, and errors in the source; and (d) two use case scenarios illustrating the research possibilities enabled by the resulting linked resource.

2. Textual Resources in the LiLa Knowledge Base

LiLa is a Linked Open Data-based Knowledge Base of linguistic resources for Latin. It promotes interoperability by adopting widely used ontologies for representing linguistic information, together with Semantic Web and Linked Data standards. In particular, OLiA is used to model linguistic annotation (Chiarcos and Sukhareva, 2015), Ontolex-Lemon to represent lexical data (McCrae et al., 2017), and POWLA to model corpus data (Chiarcos, 2012). LiLa relies on the Resource Description Framework (RDF) as its data model, representing information as triples (McBride, 2004).

Its core is the Lemma Bank, a collection of approximately 170,000 Latin headwords and 200,000 written representations.[1] In the LiLa ontology, the class Lemma is subsumed under the class Form, as defined in the Ontolex ontology (Cimiano et al., 2016). A lemma is defined as an inflected form chosen as the citation or canonical form of a lexical item. Each lemma is assigned an exclusive part of speech and an inflectional type. Some lemmas are additionally assigned a morphological base – i.e., the lexical morpheme of a word that is neither a prefix nor a suffix (Passarotti et al., 2020) –, a property that is essential for disambiguating homographic lemmas such as `occido` (*ob* + *caedo* 'to strike down') and `occido` (*ob* + *cado* 'to fall down').

Because lemmatization is a shared annotation layer across both lexical resources and corpora, the Lemma Bank serves as a central hub for integrating distributed resources in LiLa. Interoperability is achieved by linking lexical entries and corpus tokens to their corresponding lemmas in the Lemma Bank.

As described in Mambrini et al. (2022), text corpora in LiLa are modelled using POWLA, which provides classes and properties to represent document stratifications and subdivisions, and organizes different types of annotation into separate layers. A corpus is represented as an instance of POWLA's `Corpus` class. Its internal subdivisions — i.e., the works it comprises — are instances of `Document` and are linked to the corpus via `hasSubDocument`.

For each document, four annotation layers store information on tokenization, sentence division, morphology, and citation hierarchy. As regards morphological features, they are represented as properties predicated of the annotation units (i.e., tokens), using POWLA's `hasBody` to list the features and `hasTarget` to connect them to the corresponding token. In turn, dependency relations are represented as instances of the local class `DependencyRelation` (defined as a subclass of POWLA's `Relation`), and two dedicated properties connect each relation to its head and dependent nodes (the properties `hasHead` and `hasDep`, as specializations of POWLA's `hasSource` and `hasTarget`, respectively). Finally, the citation-hierarchy layer provides a way to track token positions in the text that are especially relevant for research in historical linguistics and digital humanities.

3. The PROIEL Latin treebank

The Pragmatic Resources of Old Indo-European Languages (PROIEL) project (Haug and Jøhndal, 2008) created a multilingual treebank to investigate the grammatical means by which information structure is expressed in five languages: Greek, Latin, Armenian, Gothic, and Church Slavonic. Initially, the project adopted an annotation tagset derived from the LDT, enriched with additional fine-grained distinctions (Haug, 2010). In 2017, the data were automatically converted to the UD scheme, and the Latin portion was released as "UD_Latin-PROIEL".

The latest release (r2.17) (Haug, 2025) contains most of the Vulgate New Testament translations, as well as selections from Caesar's *Commentarii de Bello Gallico*, Cicero's *Epistulae ad Atticum*, Palladius' *Opus Agriculturae*, and the first book of Cicero's *De officiis*, totaling approximately 205,500 tokens. Table 1 reports the number of tokens per work and the proportion of each work covered by the treebank.[2]

This release serves as the source for the Linked Open Data version presented in this paper.

[1] In LiLa, written representations (`https://www.w3.org/ns/lemon/ontolex#writtenRep`) account for spelling variations of a lemma. A lemma may have one or more such representations.

[2] The percentages are based on the word counts of the full texts available at *The Latin Library* website (`https://www.thelatinlibrary.com/`) and at a snapshot of the *Forum Romanum* website preserved in the Internet Archive's Wayback Machine (`https://web.archive.org/web/20230530004518/http://www.forumromanum.org/literature/palladius/agr.html`).

Table 1: Texts treebanked in UD_Latin-PROIEL.

document	tokens	% of full work
New Testament	109517	87.4%
De bello Gallico	27386	53.4%
De officiis	11375	33.3%
Epistulae ad Atticum	45308	36.3%
Opus agriculturae	11980	29.9%

Table 2: Results of the lemma matching process.

unique lemmas	n	%
Total	8,663	100.0%
1:1 single matches	7,844	90.5%
1:N ambiguous matches	519	6.0%
1:0 no matches (total)	300	3.5%
1:0 with a single candidate	102	1.2%
1:0 with multiple candidates	196	2.3%
1:0 with no candidates	2	<0.1%

4. Modeling and Linking Treebank Data

The PROIEL Latin treebank data are modeled as Linked Open Data through an automatic conversion of the CoNLL-U files from the aforementioned release. The data are extracted as-is from the source, with a single typographical correction.[3] However, the CoNLL-U file must be enriched with additional information on the citation hierarchy and, most importantly, with lemma URIs — since linking token URIs to lemma URIs via the `lila:hasLemma` property is what connects the corpus to the LiLa Knowledge Base.

The token linking process matches the lemmas annotated in the corpus to their written representation(s) recorded in the LiLa Lemma Bank. Due to lemmatization problems in the source,[4] a manual inspection is required before automatic matching can proceed, in order to either link these tokens or mark them as ineligible for linking. Of the approximately 860 affected tokens, most remain unlinked. This includes monetary values such as $\overline{\text{DCCC}}$ (800,000) and 492 Greek words (mostly from Cicero's *Letters to Atticus*). Conversely, some tokens are manually linked to the relevant Lemma Bank entries, including calendar expressions (full or abbreviated) such as 'Decembribus', 'Febr.' (i.e. *Februariis*), and 'a.d.III' (i.e. *ante diem tertius*), as well as monetary expressions such as 'HS' (*sestertius*).

To improve matching accuracy and reduce ambiguity and false positives, lemmas are extracted from the source corpus, then normalized and enriched with part-of-speech tags and (where applicable) gender labels. Normalization includes lowercasing and the substitution of *j* with *i* and *v* with *u*. Part-of-speech and gender labels are derived from the morphological features annotated in the source and mapped to the LiLa tagset. This procedure produces a composite string consisting of the lemma and its relevant features (e.g. `res_NOUN_f` and `dico_VERB`).

Matching is performed programmatically using a progressive, three-step approach. First, composite strings are matched. Second, unmatched items are reprocessed using only the lemma string. Finally, for the remaining unmatched lemmas, the edit distance to the lemma collection in LiLa is computed (with a threshold of 2) to generate a set of candidate links. The results are classified into three categories: single matches (1:1), ambiguous matches (1:N), and no matches (1:0) (see Table 2).

Single matches are automatically validated and account for 90% of the total links. This indicates strong performance for the proposed method, particularly in its ability to distinguish homographs. However, this unsupervised linking raises at least two issues. First, derivative forms such as participles may be lemmatized and tagged inconsistently in the source data, as either verbal or adjectival. In this case, LiLa's ontology class of Hipolemma – a subclass of `lila:Lemma` used to represent a citation form belonging to a word's regular inflectional paradigm that receives a different PoS tag or degree of comparison than its most canonical lemma – solves the problem, as tokens can be unified under a shared representational framework (Passarotti et al., 2025). Second, some 1:1 mappings may hide a potential 1:N mapping, considering that later Latin developments might contain homographs that are absent from the current LiLa lemma database. This remains an open issue for future investigation.

Ambiguous matches and linking candidates account for fewer than 10% of the lemmas and require manual validation. In most cases, disambiguation can be performed by inspecting the token form. For instance, *frondes* (sent_id 11089, token 3) and *frontem* (sent_id 75866, token 2) are both lemmatized as `frons`, yet they are clearly forms of two distinct lexemes, as shown by their

[3] Namely, the token `magnanimnos` (sent_id 86312, token 5), corrected to `magnanimos`. While issues with morphological features (Gamba and Zeman, 2023a) and dependency relations (Gamba and Zeman, 2023b) have been reported, no official release has addressed them yet.

[4] There are six problematic lemma strings: `calendar`, `calendar.expression`, `expression`, `greek.expression`, `monetary`, `monetary.expression` and `FIXME`. The issue is reported in https://github.com/UniversalDependencies/UD_Latin-PROIEL/issues/1.

Table 3: New lemma entries and written representations (wr) added to the LiLa Lemma Bank, categorized by part of speech (pos).

pos	new wr	new entries
PROPN	83	35
ADJ	13	8
ADV	4	3
NOUN	6	4
NUM	1	0

morphological bases *frond-* and *front-* (meaning 'leafy branches' and 'the forehead', respectively). In other cases, however, the token form itself is ambiguous, and disambiguation requires semantic analysis to identify distinct morphological bases. A representative example is *deserunt*, consistently lemmatized as `desero` but corresponding to different verbs: one derives from the base **se-* and means 'to sow', as in Palladius' *De agri cultura* (sent_id 159953, token 6); the other derives from the base *ser-* and means 'to abandon', as in Cicero's *De officiis* (sent_id 86136, token 1).

Both single and multiple linking candidates are inspected manually, with decisions made on a case-by-case basis. Of the 300 unmatched lemmas, 130 corresponded to lemmas already present in the Lemma Bank, the mismatch being due to inconsistencies in the corpus lemmatization. The remaining 170 lemmas required additions to the Lemma Bank, resulting in 107 new written representations for existing lemmas and 50 entirely new lemma entries, most of which are proper nouns (see Table 3). Furthermore, 13 words were left unlinked as they are Aramaic transcriptions, mostly from the Gospel of Mark (e.g., in the phrases "Heloi Heloi lama sabacthani" and "talitha cumi").

Finally, modeling the citation hierarchy requires converting the source data into standardized strings. For instance, from the source attribute-value pair `Ref=MATT_1.1`, a new pair is created consisting of the attribute `CitationHierarchy` and the value "Evangelium_Matthaei,Capitulum_1,Versiculus_1". In addition, the original `Ref` value is also standardized, in this case to `Vulg.Matt.1.1`, using abbreviations drawn from standard Latin dictionaries (Glare, 1968; Lewis and Short, 1879).

Figure 1 illustrates the conversion and linking process using the nominal phrase "philosophiae praeceptis" from Cicero's *De officiis* 1.1. The work node is linked to the corpus as its SubDocument and to the annotation layers as their Document. Each annotation layer connects to the tokens (directly or through subunits) via specific properties. The tokens are in turn linked to the appropriate Lemma nodes and to a dependency relation node (UD DepRel `nmod`)—as either the dependent (*philosophiae*) or the head (*praeceptis*).

5. Querying Interlinked Resources

The PROIEL Latin treebank data are now part of the LiLa Knowledge Base and can be accessed via an interactive search interface[5] or through a SPARQL endpoint[6]. Both access methods enable users to explore connections between the treebank and the many linguistic resources interlinked in LiLa.[7] To illustrate this potential, we present a query-based use case focused on sentiment analysis.

One of the lexical resources interlinked in LiLa is the Latin Affectus lexicon (Sprugnoli et al., 2020), a prior-polarity lexicon of Latin lemmas. It supports sentiment classification of LiLa-linked documents by identifying so-called sentiment words (Liu, 2012). In practice, one can use the citation layer to restrict the query scope (e.g., to a book or chapter), follow links from citations to tokens and from tokens to lemmas, and then retrieve polarity information from the lexicon. This information can be used to compute (a) the number of polarity-marked lemmas, (b) the sum of their polarity values, and (c) a normalized score obtained by dividing (b) by (a).[8]

As an example, we apply this method to Books 3 and 4 of Cicero's *Letters to Atticus*, included in UD_Latin-PROIEL. These books correspond to two distinct moments in Cicero's political career: his banishment from Rome (58 BCE) and his return (57 BCE). Moreover, Cicero explicitly foregrounds his emotional state, inviting the recipient (and thus the reader) to attend to it.[9]

Each letter in Books 3 and 4 is classified as expressing positive or negative sentiment, and the resulting sequence is used to visualize the sentiment flow of each book (Figs. 2 and 3).

Although a full interpretation of the results is beyond the scope of this paper, some patterns are noteworthy. The sentiment flow of Book 3 seems to reflect Cicero's emotional instability during his flight from Rome: letters with positive scores convey hope (e.g., letters 16 and 18), whereas those

[5] https://lila-erc.eu/LiLaLisp/.

[6] https://lila-erc.eu/sparql/.

[7] A complete list of resources is available at https://lila-erc.eu/data-page/.

[8] The query script run on the LiLa SPARQL endpoint can be found here: https://github.com/lucascdz/psm/blob/main/LiLa_sentiment_analysis_Cicero_ad_Atticum_3_and_4.rq

[9] In Book 4, letter 3: "[I would like] that you may see from my letters how I am taking events and what are my feelings and my general state of existence" (Winstedt, 1912, p.273).

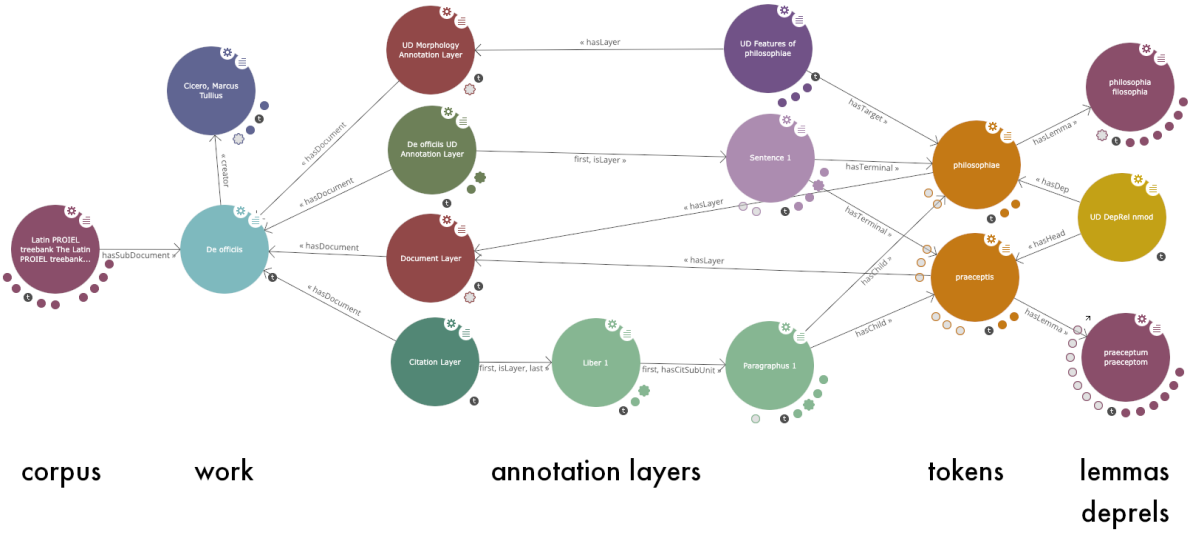


Figure 1: The nominal phrase “philosophiae praeceptis” as RDF triples in the LiLa Knowledge Base.

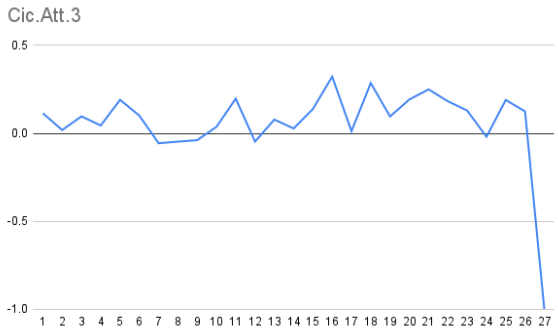


Figure 2: Sentiment classification of the letters from the third book of Cicero’s *Letters to Atticus*.

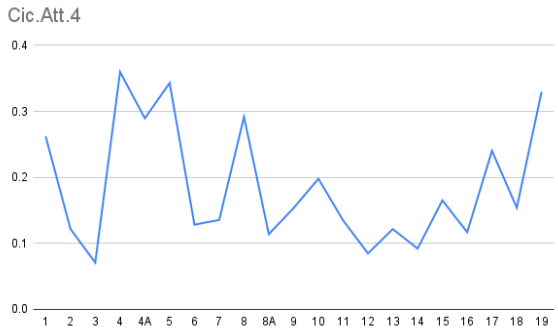


Figure 3: Sentiment classification of the letters from the fourth book of Cicero’s *Letters to Atticus*.

with negative scores express intense despair (notably letters 7–9).[10] By contrast, all letters in Book 4 receive positive scores. The lowest score (letter 3) corresponds to a report of harassment by political enemies, but it does not appear to undermine Cicero’s confidence.[11] The subsequent letter (letter 4) receives the highest score, reflecting his joy at learning that Atticus was coming to Rome. Further queries could move toward close reading, such as listing positive and negative lemmas in each letter.

Many other connections can be explored. Although space does not permit a full discussion, it is worth noting the possibility of comparing, for instance, two historians such as Caesar and Tacitus with respect to stylistic matters – for example, how they express negative meaning, whether by using the negative prefix *in-* or adverbs such as *non* – as well as lexical usage, including how many lemmas they share and which words they use to refer to non-Roman populations.

6. Conclusion

Treebanks are essential resources for the study of historical languages such as Latin. In this paper, we described the process of modelling the UD_Latin-PROIEL treebank as Linked Open Data and linking it to the LiLa Knowledge Base of interoperable linguistic resources for Latin. To this end, we addressed the challenges of identifying appropriate lemma URIs for ambiguous and previously unmatched lemmas. Although time-consuming, the data-curation process benefits both the corpus – by enabling improvements to the lemmati-

[10] An account of Cicero’s use of language of emotions in the letters written during his exile is given by Evangelou (2024).

[11] He concludes the episode by declaring: “So far as my mind is concerned, I am as strong as ever I was even in my most palmy days, if not stronger” (Winstedt, 1912, p.281).

zation layer – and the LiLa Lemma Bank, which is enriched with new entries and written representations. Overall, this work expands the possibilities for searching, assessing, and reusing treebank data within a broader ecosystem of interlinked resources.

Among the main outcomes of this work are the following:

- **Enhanced querying:** the corpus can be queried by leveraging other LiLa resources as reference points.
- **Improved reuse:** the linked data support the automatic creation of new resources, for instance by combining corpus and lexical data to derive text-specific vocabularies for educational purposes.

As such, the linked data produced here can be valuable to Latin researchers and teachers, advanced students of Latin, and computational linguists working on historical languages.

In summary, this work represents a substantial contribution to the set of Linked Open Data resources for Latin, increasing by 48% the amount of syntactically annotated corpus data available in LiLa and improving the diachronic coverage of the linked corpora by incorporating Classical and Late Latin materials. Reported inconsistencies in the source data may be addressed in future versions as new releases of UD_Latin-PROIEL become available. Given the strong performance of the proposed linking method, a natural next step is to apply it to additional Latin treebanks in order to expand the coverage of LiLa's textual resources.

7. Acknowledgements

The "LiLa - Linking Latin" project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme – Grant Agreement No. 769994.

We are grateful to Paolo Ruffolo for his assistance with data insertion into the Lemma Bank, without which this work would not have been possible.

8. Bibliographical References

David Bamman and Gregory Crane. 2006. The design and use of a latin dependency treebank. In *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (TLT2006)*, pages 67–78.

David Bamman and Gregory Crane. 2008. The logic and discovery of textual allusion. In *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, Marrakesh.

David Bamman and Gregory Crane. 2011. The ancient greek and latin dependency treebanks. In *Language Technology for Cultural Heritage*, pages 79–98, Berlin, Heidelberg. Springer Berlin Heidelberg.

David Bamman, Marco Passarotti, Gregory Crane, and Savina Raynaud. 2007. Guidelines for the syntactic annotation of latin treebanks (v. 1.3). *Tufts University Digital Library*.

Patrick J. Burns. 2023. LatinCy: Synthetic trained pipelines for latin NLP.

Flavio Massimiliano Cecchini, Timo Korkiakangas, and Marco Passarotti. 2020. A new Latin treebank for Universal Dependencies: Charters between Ancient Latin and Romance languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 933–942, Marseille, France. European Language Resources Association.

Flavio Massimiliano Cecchini, Marco Passarotti, Paola Marongiu, and Daniel Zeman. 2018. Challenges in converting the index Thomisticus treebank into Universal Dependencies. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 27–36, Brussels, Belgium. Association for Computational Linguistics.

Flavio Massimiliano Cecchini and Giulia Pedonese. 2022. A treebank-based approach to the supprema constructio in dante's Latin works. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 51–58, Marseille, France. European Language Resources Association.

Christian Chiarcos. 2012. Powla: Modeling linguistic corpora in owl/dl. In *The Semantic Web: Research and Applications*, pages 225–239, Berlin, Heidelberg. Springer Berlin Heidelberg.

Christian Chiarcos and Maria Sukhareva. 2015. Olia – ontologies of linguistic annotation. *Semantic Web*, 6(4):379–386.

Philipp Cimiano, John P. McCrae, and Paul Buitelaar. 2016. Lexicon model for ontologies: Final community group report, 10 may 2016. Technical report, Ontology-Lexicon Community Group under the W3C Community Final Specification Agreement (FSA).

Gabriel Evangelou. 2024. Loss of self, desperation, and glimmers of hope in cicero's letters from exile. In Ioannis Deligiannis, editor, *Cicero in Greece, Greece in Cicero*, pages 31–54. De Gruyter, Berlin, Boston.

Federica Gamba and Daniel Zeman. 2023a. Latin morphology through the centuries: Ensuring consistency for better language processing. In *Proceedings of the Ancient Language Processing Workshop*, pages 59–67, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Federica Gamba and Daniel Zeman. 2023b. Universalising Latin Universal Dependencies: a harmonisation of Latin treebanks in UD. In *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*, pages 7–16, Washington, D.C. Association for Computational Linguistics.

Peter GW Glare. 1968. *Oxford Latin Dictionary*. Clarendon Press, Oxford.

Dag Haug. 2015. Treebanks in historical linguistic research. In Carlotta Viti, editor, *Perspectives on historical syntax*, pages 187–202. John Benjamins Publishing Company.

Dag Trygve Truslew Haug. 2010. Proiel guidelines for annotation.

Dag TT Haug and Marius Jøhndal. 2008. Creating a parallel treebank of the old indo-european bible translations. In *Proceedings of the second workshop on language technology for cultural heritage data (LaTeCH 2008)*, pages 27–34.

Timo Korkiakangas and Marco Passarotti. 2011. Challenges in annotating medieval latin charters. *Journal for Language Technology and Computational Linguistics*, 26(2):105–116.

C. Lewis and C. Short. 1879. *A Latin Dictionary*. Clarendon Press, Oxford.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

Francesco Mambrini. 2016. The ancient greek dependency treebank: Linguistic annotation in a teaching environment. In G. Bodard and M. Romanello, editors, *Digital Classics Outside the Echo-Chamber*, pages 83–99. Ubiquity Press, London.

Francesco Mambrini, Marco Passarotti, Giovanni Moretti, and Matteo Pellegrini. 2022. The index Thomisticus treebank as linked data in the LiLa knowledge base. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4022–4029, Marseille, France. European Language Resources Association.

Brian McBride. 2004. The resource description framework (rdf) and its vocabulary description language rdfs. In Steffen Staab and Rudi Studer, editors, *Handbook on Ontologies*, pages 51–65. Springer Berlin Heidelberg, Berlin, Heidelberg.

John P McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The ontolex-lemon model: development and applications. In *Electronic lexicography in the 21st century: Proceedings of eLex 2017 conference*, pages 587–97, Brno. Lexical Computing CZ.

Barbara McGillivray and Alessandro Vatri. 2015. Computational valency lexica for latin and greek in use: a case study of syntactic ambiguity. *Journal of Latin Linguistics*, 14(1):101–126.

Joakim Nivre. 2008. Treebanks. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics*, volume 1, pages 225–41. Walter de Gruyter, Berlin/New York.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Marco Passarotti. 2019. The project of the index thomisticus treebank. In Monica Berti, editor, *Digital Classical Philology*, pages 299–320. De Gruyter Saur, Berlin, Boston.

Marco Passarotti, Federica Iurescia, and Paolo Ruffolo. 2025. Harmonizing divergent lemmatization and part-of-speech tagging practices for Latin participles through the LiLa knowledge base. In *Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX-2025)*, pages 103–114, Vienna, Austria. Association for Computational Linguistics.

Marco Passarotti, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, and Rachele Sprugnoli. 2020. Interlinking through lemmas. the lexical collection of the LiLa Knowledge Base of linguistic resources for Latin. *Studi e Saggi Linguistici (SSL)*, 58(1):177–212.

Marco Carlo Passarotti, Flavio Massimiliano Cecchini, Rachele Sprugnoli, G Moretti, et al. 2021. Udante. syntactic annotation of dante alighieri's latin texts. *Studi Danteschi*, 86(1):309–338.

Rachele Sprugnoli, Francesco Mambrini, Giovanni Moretti, and Marco Passarotti. 2020. Towards the modeling of polarity in a latin knowledge base. In *Proceedings of the Third Workshop on Humanities in the Semantic Web (WHiSe 2020) co-located with 15th Extended Semantic Web Conference (ESWC 2020). Heraklion, Greece, June 2, 2020*, pages 59–70. CEUR-WS.

Eric Otto Winstedt. 1912. *Cicero: Letters to Atticus; with an English translation*, volume 1. W. Heinemann, London.

9. Language Resource References

Dag Haug. 2025. *UD_Latin-PROIEL*. Universal Dependencies (UD), r2.17.

Federica Iurescia and Federica Gamba and Flavio Massimiliano Cecchini and Francesco Mambrini and Giovanni Moretti and Marco Passarotti and Paolo Ruffolo. 2025. *UD_Latin-CIRCSE*. Universal Dependencies (UD), r2.17.

Rachele Sprugnoli and Giovanni Moretti and Marco Passarotti and Daniela Corbetta and Andrea Peverelli. 2020. *CIRCSE/Latin_Sentiment_Lexicons: First release*. Zenodo.

Language Models for the Restoration of Latin Legal Manuscripts

Shibingfeng Zhang[1], **Edoardo Caraffa**[1], **Annafelicia Zuffrano**[1], **Maddalena Modesti**[1], **Giovanni Colavizza**[1,2]
[1]Department of Classical Philology and Italian Studies, University of Bologna,
[2]Centre for Digital and Computational Humanities, University of Copenhagen
{shibingfeng.zhang, annafelicia.zuffran2, maddalena.modesti3, giovanni.colavizza}@unibo.it,
edoardo.caraffa2@studio.unibo.it

Abstract

The collection of historical notarial documentation from Bologna is a valuable source, providing deep insights into the city's institutional, legal, and socio-economic history. However, many of these manuscripts have sustained physical damage during centuries of conservation, rendering the text incomplete. To address this, we explored the restoration of these Latin notary documents using encoder-based pre-trained language models (PLMs) under the assumption that the length of missing text is known by estimation from the physical damage. We address the structural misalignment between the physical lacuna of the manuscript and the subword tokenization schemes of PLMs by designing an iterative decoding strategy to align model predictions with the known physical dimensions of lacuna. We also compared the efficacy of monolingual versus multilingual pre-training. Our strategy significantly outperforms baselines consist of standard decoding methods. Furthermore, stratified analysis across different text sections reveals that while monolingual models achieve better performance in general, multilingual models show a suggestive advantage in lexically dense segments, though this finding is not statistically significant. Overall, the best performance achieved by our method is a Hit@1 rate of 35.47% in the short-span setting and 18.75% in the long-span setting. While fully autonomous restoration remains an open challenge, our system provides a useful assistive tool for paleographers.

Keywords: Text Restoration, Diplomatic Documents, Pre-trained Language Models, Digital Paleography, Low-resource Languages

1. Introduction

The archival heritage of medieval Bologna represents a valuable resource to study the city's history. Since 1937, when Giorgio Cencetti proposed the creation of the Codice diplomatico bolognese, a comprehensive edition intended to systematically transcribe and preserve the city's medieval documentary heritage, research has progressed significantly, with the successful academic publication of 10th and 11th century documents (Cencetti, 1977; Feo, 2001) and the development of a 12th century notarial prosopography (Modesti, 2012). However, the physical preservation of these records is inconsistent. Many extant manuscripts suffer from lacuna that obscure critical historical data.

Text restoration, the challenge of inferring missing content within a damaged text, therefore becomes a potential helper for paleographers and archaeologists. In the context of historical manuscript restoration, the text length can be inferred from the size of the physical damage. While leveraging pre-trained language models (PLMs) has become a popular approach for this task (Lazar et al., 2021; Assael et al., 2025), applying them to digitized medieval notary documents presents unique challenges. This paper focuses on restoring the missing text of digitalized notary documents. We adopt encoder-based models pre-trained on the masked language modeling (MLM) objective, as this pre-training task is very similar in nature to text restoration.

When performing text infilling, the MLM head of encoder-based models outputs strictly one token per mask, which is ideal for text restoration with known text length, as it provide the full control over the length of predicted text. However, there are two primary challenges to address: First, while we can estimate how many characters the lacuna covers using its physical size, it maps to an unknown number of subword tokens. Second, physical damage often occurs at arbitrary positions that do not align with the PLM's tokenization scheme. This "slicing" of words often creates fragments that differ from the standard subwords seen during pre-training, leading to a structural misalignment that the MLM head is not natively designed to handle.

We raise the first research question:

- **RQ1:** When using encoder-based PLMs for the text restoration task, how do we effectively resolve the misalignment between the lacuna and the tokenization scheme of the PLMs?

For low-resource languages, multilingual training sometimes demonstrates benefits for system performance (Conneau et al., 2020). In this paper, we adopt a monolingual PLM pre-trained only on Latin and its multilingual counterpart pre-trained on Latin, Ancient Greek, and English. This setting allows us to raise the second research question:

- **RQ2:** Does the multilingual PLM have an advantage over its monolingual Latin counterpart

Proceedings of LT4HALA 2026 @ LREC 2026, pages 361–367.
11 May 2026. *©ELRA Language Resources Association (ELRA), 2026*

on the text restoration task?

For RQ1, we designed an iterative decoding strategy in order to resolve the misalignment issue. Experiments demonstrate the effectiveness of our method, and we also conducted an ablation study to investigate further. For RQ2, we annotated the test set by dividing it into different functional components, then conducted a thorough analysis to study the performance of both PLMs on each component. Our contributions in this paper are threefold:

1. We proposed a decoding strategy to adapt encoder-based PLMs to scenarios of text restoration with known character length.
2. We conduct experiment to analyze the impact of multilinguality of PLMs in text restoration of Latin.
3. We conduct a study on text restoration task of different text components of diplomatic text. As far as we know this is the first time this technology is applied onto this genre of text.

2. Related Works

Several studies explore settings similar to ours, adopting PLMs where the system input includes both the textual context and the estimated character length of the missing text. For example, Lazar et al. (Lazar et al., 2021) investigated the restoration of Ancient Akkadian inscriptions, where they estimated the number of missing signs and trained BERT from scratch using each sign in the Akkadian vocabulary as a token. Assael et al. (Assael et al., 2025) trained a decoder-based model from scratch for the text restoration of Latin and Ancient Greek inscriptions, using an extra feed-forward head to predict the length of missing text given the context. This method treats each character as a token to ensure the alignment of the output and the predicted length of the missing text. These works resolved the length alignment issue at its root by adapting the tokenization scheme and training the model from scratch; however, this approach is more computationally expensive, requires more data, and effectively reduces the input context window. This may not be an issue for the restoration of inscriptions, as texts on hard materials tend to be shorter and have a lower requirement for a large context window. However, their method is deemed unsuitable for our task since notary documents tend to be longer.

Riemenschneider et al. (Riemenschneider and Frank, 2023) proposed several PLMs specifically for Latin. Their work introduces two groups of models: the first consists of encoder-based models utilizing the RoBERTa (Zhuang et al., 2021) architecture, the second comprises encoder-decoder models based on the T5 architecture (Raffel et al., 2020). All models were trained from scratch. Specifically, the first group includes LaBERTa (Latin only) and PhilBERTa (Latin, Ancient Greek, and English). The second group includes LaTa (Latin only) and PhilTa (Latin, Ancient Greek, and English). These multilingual variants were designed to investigate the impact of cross-linguistic data on historical language PLMs. Across a series of tasks that focused on semantic, syntactic, and morphological aspects of Latin, the monolingual models yielded similar results with no statistically significant differences in most tasks comparing to their multilingual counterpart. The authors suggested that these findings might be due to the relatively small size of their test set, leaving the impact of multilingual pre-training an open question.

In this study, we focus on the first group, as the encoder-only architecture allows us to formulate the restoration as a masked language modeling task rather than an open-ended text generation task. This approach provides greater control over the output length, enabling us to better align the predictions with the physical dimensions of the lacuna. We did not alter the original Byte-Pair Encoding tokenization strategy of the RoBERTa-based models, as doing so would require training from scratch. Instead, we proposed several decoding strategies for text restoration that do not require altering the original structure of the model. This method is presented in Section 3. We evaluate both LaBERTa and PhilBERTa to analyze the effects of multilinguality specifically on the text restoration.

3. Methodologies

Our proposed text restoration method operates under the assumption that the character length $L_{missing}$ of the missing text is known or estimated through physical analysis of the manuscript. We designed a four-stage decoding process to bridge the gap between language model's tokenization scheme and character-level length constraint. Figure 1 demonstrate the process of text restoration method:

- **Stage 1: Lacuna Expansion**
  To ensure compatibility between the physical lacuna and the model's Byte-Pair Encoding tokenization scheme, we first perform *boundary expansion*. If a lacuna begins or ends mid-word, we expand the lacuna to the nearest whitespace-delimited token boundaries. The partial strings covered by such expansion are treated as a known prefix and suffix.

  The target length is adjusted accordingly: $L_{adjusted} = L_{missing} + L_{prefix} + L_{suffix}$.

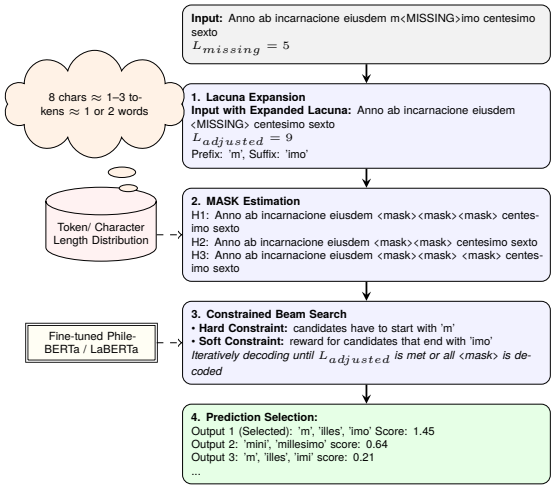


Figure 1: Text restoration workflow

- **Stage 2: Mask Estimation**
  Since encoder-based models maintain a strict one-to-one mapping between input $[MASK]$ tokens and output tokens, we must estimate the optimal number of masks N to represent $L_{adjusted}$ characters. Because BPE token lengths vary, we calculate the prior distribution of token counts for a given character length from the training set. We select the top k most frequent mask distributions. This results in k distinct input sequences, each representing a different hypothesis of the underlying token structure:

- **Stage 3: Prefix&Suffix Constrained Beam Search**
  We employ an iterative decoding strategy based on Beam Search, modified with two specific constraints:

  1. **Prefix Constraint:** During the first decoding step, the candidates are restricted to tokens that begin with the known prefix string.
  2. **Suffix Scoring Reward:** In the final decoding steps, we apply a bonus to the log-probability of sequences that terminate with the known suffix.

- **Stage 4: Prediction Selection** Decoding terminates once all $[MASK]$ positions are filled or the character length $L_{adjusted}$ is reached. Finally, the resulting sequences are aggregated from the k different input, and the candidate with the highest log-probability is selected as the optimal restoration.

4. Experiments

Our dataset comprises 1,184 notarial documents, primarily sourced from the Archivio di Stato di Bologna. These records were split into training, development, and testing sets according to an 8:1:1 ratio. On average, each document contains approximately 380 words. These deeds, written in Latin by 254 notaries, document various types of legal transactions such as sales, donations, exchanges, emphyteusis, wills, and other patrimonial instrumenta.

From a diplomatic point of view, it is also worth noting that the text of these documents is rigidly formalized in terms of discourse construction. In general, the content of a complete notary document can be partitioned into the protocol, the text body, and the eschatocol. There is also a special kind of document called rogationes, which are summary notes of the essential elements of the deed, drawn up by the notary in view of the subsequent drafting in extended form. An exemplary document is provided in Figure 2 along with its translation to provide a more concrete understanding of these concepts.

The LaBERTa and PhilBERTa models were fine-tuned using a standard MLM objective, incorporating an early-stopping strategy to prevent overfitting. To account for the non-deterministic nature of deep learning, we trained five independent checkpoints for both models and report the mean and confidence interval of our results.

For the evaluation of our model, we compare three variants of decoding strategies:

- **Proposed Iterative Decoding with Lacuna Expansion:** The full strategy including lacuna expansion, distribution-based mask estimation, and prefix&suffix constraints as described in Section 3.

- **Iterative Baseline (No Lacuna Expansion):** An iterative decoding strategy for comparison that skips the lacuna expansion step. In this variant, the number of mask tokens is estimated by dividing the missing text length by the average character length per token.

- **Non-Iterative Baseline:** A simple, single-pass decoding strategy. Similar to the iterative baseline, it estimates mask counts based on the average character length per token. This serves as a baseline for comparison, as it's a standard practice of text infilling with multiple masks.

To evaluate the model's restoration capabilities in a detailed way, a paleographer annotated the test set to delineate this structure. The 110 documents in the test set were categorized as follows:

- **Complete with rogationes(25):** documents that contain all three partitions and their corresponding rogationes.

Protocol
Latin: In nomine sancte et individue Trinitatis. Anno Domini millesimo centesimo trigesimo secundo, pridie kalendas aprelis, indicione decima.
English: In the name of the holy and indivisible Trinity. In the year of God one thousand one hundred and thirty-six, the day before the calends of April, tenth indiction.

Text Body
Latin: Ego quidem Rainerius filius Lamberti de Beio hoc donacionis instrumento presenti die dono in honore Dei et ecclesie Sancti Victoris et tibi donno Alberio priori eiusdem ecclesie tuisque fratribus et successoribus proprium in perpetuum conducticium unde pertinuerit, id est omne quod ego habeo et teneo et michi pertinet iure vel actione a radice montis Sancti Victoris usque ad crucem de Dilvino ab Aposa usque ad rivum ex illa parte Barbiani et peciam unam terre aratorie in loco ubi dicitur Castellioni prope cruce de Piro cum ingressu et egressu suo usque in via publica et cum omnibus super se et infra se habentem in integrum. Finis vero eius: ab uno latere a sero et uno capite a meridie possidet Albertus de Rigiza, alio latere a mane detinet Rusticus de Emma terra de socru sua, alio capite ab aquilone adest via publica et si qui alii affines sunt; omnium quod infra hos fines michi pertinet in integrum pro remedio anime mee meeque uxoris nec non et patris et matris mee, in presenti dono et trado atque concedo supradicte ecclesie et tibi donno Alberio priori tuisque fratribus ac successoribus ad habendum, tenendum ac possidendum et quicquid tibi tuisque fratribus ac successoribus deinceps placuerit ad utilitatem eiusdem ecclesie faciendum. Ut nullam litem nullamque controversiam deinceps a me vel a meis heredibus quolibet modo aliquo in tempore vos vel vestri successores de cetero sustineatis ab omni quoque homine prescriptas res legittime defendere et auctorizare tibi et tuis heredibus promitto. Et si ego vel mei heredes prelibatam meam donationem sinplicem in totum vel pro parte audaci nisu quandoque infringere temtavero et eam semper inviolatam custodire noluero, penam triginta denariorum Lucensium libras tibi vel tuis successoribus dare promitto et insuper hanc donationem semper intactam conservare promitto.
English: I, Rainerius, son of Lambertus of Beio, in the present day, give by this document of donation in honor of God and of the church of Saint Victor and to you, father Alberius, prior of the aforesaid church, and to your brothers and successors, full property in perpetuity, wherever it should have been pertained, which is everything I own and possess and belongs to me, by right and action, from the foot of the mountain of Saint Victor as far as the cross of Dilvino, and from the Aposa as far as the river on that side of Barbianum, and moreover one plot of arable land in the place called Castellione, near the cross of Pero, with its right of entry and exit as far as the public road, and with everything it has above and below it, in its entirety. Moreover, its boundaries are: on the western and southern side, the land owned by Albertus of Rigiza, on the eastern side, held by Rusticus of Emma, the land of his mother-in-law, on the northern side, the public road, and all other boundaries, if there are any other. Everything that I own within these boundaries, in the present day, I give and deliver and grant in his integrity, for the salvation of my soul, and of my wife, and of my father and mother, to the aforesaid church and to father Alberius, prior, and to his brothers and successors, so that they may have, hold and possess it, and use until it will be of any utility for the aforesaid church. So that no arguing nor dispute hereafter from me or from my heirs, in any way or at any time, you or your successors shall have to endure, I promise to rightfully defend and authorize the aforesaid property against everyone for you and your heirs. And if I or my heirs shall at any time, by bold attempt, try to violate this my simple donation in whole or in part, or if I shall not wish to keep it always inviolate, I promise to pay to you or your successors the penalty of thirty denarii of Lucca, and moreover I promise to preserve this donation always intact.

Eschatocol
Latin: Actum in canonica Sancti Iohannis in Monte, indicione predicta. Prenominatus Rainerius hoc donacionis instrumentum ut supra legitur scribere rogavit. Iohannes presbiter de ecclesia Sancta Tecla, Lambertus investitor filius Petri de Leo, Lambertus de Auria, Petrus filius Alberti de Vivelinda, Grimaldus filius Bonifantini de Sancto Rofillo, Eldus de Verona rogati sunt testes. Gerardus tabellio hoc donacionis instrumentum ut supra legitur scripsi et firmavi.
English: Done in the canonry of Saint John on the Mount, in the aforesaid indiction. The aforesaid Rainerius requested that this instrument of donation, as read above, may be written. Iohannes, presbyter of the church of Saint Tecla, Lambertus investor, son of Petrus of Leo, Lambertus od Auria, Petrus, son of Albertus of Vivelinda, Grimaldus, son of Bonifantinus of Saint Rufillo, Eldus of Verona are called as witnesses. I, Gerardus notary, wrote and signed this instrument of donation, as read before.

Rogationes
Latin: Pridie kalendas aprelis, indicione x. Testis Iohannes presbiter de Sancta Tecla et Lambertus investitor filius Petri Leo et Lambertus de Auria et Petrus filius Alberti de Vivelinda et Eldus de Verona, Grimaldus filius Bonifantini de Sancto Rofillo. Cartulam donationis fecit Rainerius filius Lamberti de Beio pro remedio anime sue et de uxore sua et patris et matris sue in honore Dei et ecclesie Sancti Victoris et donno Alberio priori eiusdem ecclesie suisque fratribus ac successoribus de omnibus iuris et actionibus quod sibi pertinet a pede montium Sancti Victoris usque ad crucem Dilvini a Aposa usque ad rivum ex illa parte Barbiani et insuper peciam unam terre aratorie prope crucem de Pero in loco qui dicitur Castellioni sub pena et defensione.
English: The day before the calends of April, tenth indiction. Witnesses Iohannes, presbyter of Saint Tecla, and Lambertus investor, son of Petrus Leo, and Lambertus of Auria and Petrus, son of Albertus of Vivelinda and Eldus of Verona, Grimaldus, son of Bonifantinus of Saint Rufillo. Rainerius, son of Lambert of Beio, made a charter of donation for the salvation of his own soul, and of his wife, and of his father and mother, in honor of God and of the church of Saint Victor and of father Alberius, prior of the aforesaid church, and his brothers and successors, regarding all the rights and actions that belong to him from the foot of the mountain of Saint Victor as far as the cross of Dilvino, and from the Aposa as far as the river on that side of Barbianum, and moreover one plot of arable land near the cross of Pero, in the place called Castellione, under penalty and warranty.

Figure 2: Example of a notary document, produced in 1132, archive shelfmark *Archivio di Stato di Bologna, Corporazioni religiose soppresse, S. Giovanni in Monte, 2/1342 n. 10a*

- **Complete with no rogationes (68):** documents that contain the three primary parts but lack rogationes.
- **Incomplete (17):** documents that consist of only one or two parts.

5. Results

We performed the text restoration task using two types of span length settings: short spans, ranging between 5-10 characters, and long spans, ranging between 10-20 characters. To facilitate evaluation of the model's restoration capabilities across different sections (protocol, text body, eschatocol, and rogationes) of the notarial text,no spans were allowed to cross section boundaries. It is worth noting that these evaluation examples were created artificially by removing text from the notarial documents, as we lack ground-truth data for real damages in the documents.

Tables 2 and 4 present the Character Error Rate (CER), Overlap Score, and Hit@1 rate for various decoding strategies. Tables 3 and 5 provide the stratified results of iterative decoding performance across different sections of the notary documents. The Overlap Score is defined as follows:

$$\text{Overlap Score} = \frac{|LMS|}{|LMS| + |\Delta_{gt}| + |\Delta_p|}$$

where $|LMS|$ is the length of the longest matched substring between the ground truth and the prediction, $|\Delta_{gt}|$ represents the length of the non-overlapping segment of the ground truth, $|\Delta_p|$ represents the length of the non-overlapping segment of the prediction.

To ensure a fair comparison, we stripped off the prefix and suffix from predictions generated by the iterative decoding with lacuna expansion method. Consequently, the ground truth remains consistent across all three evaluated decoding methods.

The results in Tables 2 and 4 demonstrate the significant advantage of expanding lacuna to align the lacuna boundary with the PLM's tokenization scheme across all evaluation metrics. Without this alignment method, the performance of the standard iterative approach remains marginal. Specifically, in the short-span setting, the results are nearly indistinguishable from simple parallel decoding. While the standard iterative approach shows a slight improvement over simple decoding in long-span settings, it still falls considerably short of the performance achieved when the lacuna expansion is applied. Notably, the multilingual PhilBERTa model did not demonstrate an advantage over the monolingual LaBERTa model in general. Its performance consistently lagged across all decoding modes and exhibited higher variance.

To illustrate the model's behavior qualitatively, Table 1 presents several examples of LaBERTa's predictions in different decoding modes. By expanding the damage boundary and aligning it with the model's tokenization scheme, the model tends to

produce coherent predictions and, in some cases, offers reasonable alternatives to the ground truth. In contrast, the other two decoding modes tend to generate output that is less convincing grammatically and semantically.

As for text type specific results of iterative decoding method with lacuna boundary expansion, as demonstrated in Tables 3 and 5, both models followed similar performance patterns across sections. Performance on the rogationes and eschatocol sections tended to be the lowest, whereas performance on the protocol and text body are better.

Interestingly, while falling behind in the protocol, text body, and eschatocol sections, PhilBERTa surpasses LaBERTa by a marginal advantage within the rogationes segments. Although this advantage is not statistically significant, as the confidence intervals of the two models' metrics overlap considerably, it suggests that the multilingual training of PhilBERTa may offer specific benefits in this context. A defining characteristic of the rogationes is that it serves as a short draft by the notary to summarize the notary document's content. It is more information-packed than the other sections by nature. To investigate this further, we calculated the lexical density of various document parts using the Type-Token Ratio (TTR) (Laurs, 2024) and Hapax Legomena. Here we define Hapax Legomena as words that occured only once in the specific text type. As shown in Table 6, the rogationes type exhibits both the highest TTR and the highest Hapax Legomena-to-token ratio across all document segments, confirming its high lexical richness. Given this distinct linguistic profile, one might speculate that the multilingual training of PhilBERTa could technically offer advantages in a high lexical richness and non-formulaic context, although the present results do not provide statistically significant evidence for this, and the observed difference remains inconclusive.

In general, the results of both models are far from perfection, with the CER ranging from 39.51% to 72.83% depending on the text type.These results suggest that while the model cannot yet automatically restore text with full autonomy, it holds significant potential as an assistive tool for paleographers.

6. Conclusions

In this paper, we addressed the challenging task of restoring historical notary documents in Latin language by leveraging encoder-based pre-trained language models. We investigated two primary research questions concerning the technical misalignment between physical lacuna and subword tokenization (RQ1) of pre-trained language models and the potential benefits of multilingual pre-training for this specific text genre (RQ2). We addressed the misalignment issue through an iterative decoding strategy with lacuna expansion, which significantly outperformed baselines. Our comparative analysis showed that while the monolingual LaBERTa model generally excelled in most sections, the multilingual PhilBERTa demonstrated marginal advantage in the lexically dense scenario, though this difference was not statistically significant. Although current error rates suggest that fully autonomous text restoration is not yet feasible, the model's performance highlights its value as an assistive tool for paleographers.

A limitation of this study is that our evaluation assumes perfect knowledge of the missing text length, as we used the exact number of characters removed during reference. In a real-world application, length would be estimated from the physical dimensions of the damage, introducing potential errors. Future work will focus on utilizing visual inputs of manuscripts to automatically estimate the character length of lacunae and investigate the impact of estimation precision on model output. Additionally, we plan to experiment with generative language models, such as those proposed by Riemenschneider et al. (Riemenschneider and Frank, 2023).

7. Acknowledgements

This study is funded by FutureData4EU project (Grant Agreement n. 101126733). FutureData4EU is Co-Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or REA. Neither the European Union nor the granting authority can be held responsible for them.

The authors acknowledge the CINECA award under the ISCRA initiative for the availability of high-performance computing resources.

8. Bibliographical References

Yannis Assael, Thea Sommerschield, Alison Cooley, Brendan Shillingford, John Pavlopoulos, Priyanka Suresh, Bailey Herms, Justin Grayston, Benjamin Maynard, Nicholas Dietrich, et al. 2025. Contextualizing ancient texts with generative neural networks. *Nature*, 645(8079):141–147.

Giorgio Cencetti. 1977. *Notariato medievale bolognese. Tomo I: Scritti di Giorgio Cencetti*. Number 3, 1 in Studi storici sul notariato italiano. Consiglio nazionale del notariato, Roma.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek,

Input Text	Label	Decoding Mode		
		Iterative Boundary	Iterative	Simple
...ego quidem donno Paulus Dei gratia abbas Sancti Stefani pres[SPAN]ie iu re enphiteosin vendo tibi Dominico filio Alberti de Oliveti et...	enti d	**enti d**,enti ven,enti iure,enti enph,enti in d,enti do et,enti present,ente d,enti di e,enciarum	entiarum,en,enci,encia,encio, on,encium,enti,es,esi	entiarum
Peto a te donno Guido abbas monasterii Sancti Stephani qui v[SPAN]ti me Rodulfo filio Arnaldi seu filiis et heredibus meis aut cui ego...	ocatur Hierusalem u	**ocatur Hierusalem u**,ocatur in Claterna u,ocatur Gerusalem uti pro,ocatur in Strada u,ocatur in Strata u,ocatur in Musiliano u,ocatur Gerusalem uti a,ocatur in Musciliano u,ocatur Iehrusalem u,ocatur Ierusalem uti dign	itate Hierusalem uti,itate Ierusalem uti,itate vocatur Hierusalem,itate vocatur Ierusalem,. 19. Pe,. 32. Pe,. 29. Pe,itate mona a,itate dicitur Hierusalem,itate Hierusalem pro	itate Hierusalem preen
Et si ego vel mei su[SPAN]mnia firmiter semper ob servare noluerimus, pene nomine decem solidos Lucensium dare vobis et vestris heredibus pro mitto...	cessores predicta o	**cessores predicta o**,cessores prelibata o,cessores vel fratres vel,cessores hanc enphiteos,cessores vel mei heredes,cessores omnia predicta,cessores de cetero sustin,cessores vel ego vel mei,cessores vel mei successores,essores predicta omnia firmiter	essores predicta omnia,essores predicta pe,essores omnia predicta,essores predicta li,essores predicta instru,essores predicta ho,essores prescripta,cessores predicta omnia,essores predicta omni,essores predicta le	essores omnia omni

Table 1: Some examples of LaBERTa's output using different decoding modes. Predictions are separated by comma.

Decoding Mode	Model	CER	Overlap	HR@1
Iterative Boundary	LaBERTa	$\mathbf{55.40}_{\pm 1.09}$	$\mathbf{56.08}_{\pm 0.53}$	$\mathbf{35.47}_{\pm 1.04}$
	PhilBERTa	$63.25_{\pm 2.63}$	$48.70_{\pm 2.71}$	$28.56_{\pm 2.80}$
Iterative	LaBERTa	$\mathbf{72.30}_{\pm 0.88}$	$\mathbf{26.67}_{\pm 0.71}$	$\mathbf{1.87}_{\pm 0.31}$
	PhilBERTa	$75.34_{\pm 1.13}$	$23.12_{\pm 1.31}$	$1.32_{\pm 0.21}$
Simple	LaBERTa	$\mathbf{72.38}_{\pm 0.86}$	$\mathbf{26.31}_{\pm 0.65}$	$\mathbf{1.80}_{\pm 0.24}$
	PhilBERTa	$75.83_{\pm 1.15}$	$22.48_{\pm 1.24}$	$1.15_{\pm 0.20}$

Table 2: Overall Performance, span length 5-10 characters (Mean $\pm$ 95% CI, HR stands for hit-rate)

Text Type	Model	CER	Overlap	HR@1	HR@3	HR@5	HR@10
Protocol	LaBERTa	$\mathbf{46.93}_{\pm 2.71}$	$\mathbf{61.84}_{\pm 1.31}$	$\mathbf{41.03}_{\pm 2.10}$	$\mathbf{51.70}_{\pm 4.42}$	$\mathbf{56.39}_{\pm 4.82}$	$\mathbf{62.99}_{\pm 3.37}$
	PhilBERTa	$48.18_{\pm 4.91}$	$59.66_{\pm 4.17}$	$40.21_{\pm 6.14}$	$48.00_{\pm 4.43}$	$53.30_{\pm 5.39}$	$59.46_{\pm 5.18}$
Text Body	LaBERTa	$\mathbf{49.84}_{\pm 1.63}$	$\mathbf{60.91}_{\pm 1.29}$	$\mathbf{38.98}_{\pm 2.09}$	$\mathbf{51.21}_{\pm 1.37}$	$\mathbf{54.90}_{\pm 1.23}$	$\mathbf{58.81}_{\pm 0.84}$
	PhilBERTa	$63.36_{\pm 2.99}$	$49.23_{\pm 2.90}$	$28.49_{\pm 3.37}$	$41.48_{\pm 4.76}$	$44.80_{\pm 5.11}$	$48.53_{\pm 4.26}$
Eschatocol	LaBERTa	$\mathbf{62.38}_{\pm 3.04}$	$\mathbf{50.48}_{\pm 1.44}$	$\mathbf{32.81}_{\pm 1.92}$	$\mathbf{46.66}_{\pm 2.19}$	$\mathbf{50.05}_{\pm 0.84}$	$\mathbf{53.34}_{\pm 1.85}$
	PhilBERTa	$66.11_{\pm 5.36}$	$46.49_{\pm 4.15}$	$27.62_{\pm 3.81}$	$40.21_{\pm 3.24}$	$43.37_{\pm 3.28}$	$48.20_{\pm 4.51}$
Rogationes	LaBERTa	$76.77_{\pm 4.34}$	$38.09_{\pm 1.78}$	$16.27_{\pm 2.46}$	$\mathbf{29.57}_{\pm 3.82}$	$33.33_{\pm 3.72}$	$\mathbf{37.42}_{\pm 3.42}$
	PhilBERTa	$\mathbf{73.22}_{\pm 6.27}$	$\mathbf{38.22}_{\pm 2.52}$	$\mathbf{16.80}_{\pm 2.51}$	$28.72_{\pm 2.76}$	$\mathbf{33.79}_{\pm 4.00}$	$36.95_{\pm 4.25}$
Overall	LaBERTa	$\mathbf{55.40}_{\pm 1.09}$	$\mathbf{56.08}_{\pm 0.53}$	$\mathbf{35.47}_{\pm 1.04}$	$\mathbf{48.27}_{\pm 1.01}$	$\mathbf{52.01}_{\pm 0.93}$	$\mathbf{56.09}_{\pm 0.80}$
	PhilBERTa	$63.25_{\pm 2.63}$	$48.70_{\pm 2.71}$	$28.56_{\pm 2.80}$	$40.88_{\pm 3.62}$	$44.55_{\pm 3.76}$	$48.83_{\pm 3.54}$

Table 3: Stratified Iterative Boundary Results, span length 5-10 characters(Mean $\pm$ 95% CI, HR stands for hit-rate)

Decoding Mode	Model	CER	Overlap	HR@1
Iterative Boundary	LaBERTa	$\mathbf{55.22}_{\pm 0.95}$	$\mathbf{41.18}_{\pm 0.61}$	$\mathbf{18.75}_{\pm 1.32}$
	PhilBERTa	$62.40_{\pm 3.65}$	$33.52_{\pm 2.99}$	$12.97_{\pm 2.69}$
Iterative	LaBERTa	$\mathbf{63.98}_{\pm 0.60}$	$\mathbf{27.66}_{\pm 0.47}$	$\mathbf{1.15}_{\pm 0.17}$
	PhilBERTa	$67.53_{\pm 1.57}$	$23.50_{\pm 1.15}$	$1.10_{\pm 0.29}$
Simple	LaBERTa	$\mathbf{65.03}_{\pm 0.54}$	$\mathbf{25.39}_{\pm 0.58}$	$\mathbf{0.82}_{\pm 0.12}$
	PhilBERTa	$68.92_{\pm 1.64}$	$21.50_{\pm 1.02}$	$0.79_{\pm 0.08}$

Table 4: Overall Performance, span length 10-20 characters (Mean $\pm$ 95% CI, HR stands for hit-rate)

Text Type	Model	CER	Overlap	HR@1	HR@3	HR@5	HR@10
Protocol	LaBERTa	$\mathbf{39.51}_{\pm 3.63}$	$\mathbf{52.47}_{\pm 1.97}$	$\mathbf{27.84}_{\pm 3.73}$	$\mathbf{36.91}_{\pm 5.02}$	$\mathbf{43.90}_{\pm 5.82}$	$\mathbf{49.14}_{\pm 5.43}$
	PhilBERTa	$42.94_{\pm 4.66}$	$49.14_{\pm 3.52}$	$24.74_{\pm 3.73}$	$35.07_{\pm 4.05}$	$39.22_{\pm 4.78}$	$44.44_{\pm 5.08}$
Text Body	LaBERTa	$\mathbf{50.93}_{\pm 1.73}$	$\mathbf{45.56}_{\pm 0.93}$	$\mathbf{21.62}_{\pm 1.21}$	$\mathbf{31.00}_{\pm 1.85}$	$\mathbf{34.38}_{\pm 1.92}$	$\mathbf{37.51}_{\pm 2.27}$
	PhilBERTa	$62.20_{\pm 4.37}$	$34.07_{\pm 3.67}$	$13.13_{\pm 3.13}$	$21.14_{\pm 3.08}$	$23.58_{\pm 3.89}$	$26.45_{\pm 3.67}$
Eschatocol	LaBERTa	$\mathbf{63.70}_{\pm 0.92}$	$\mathbf{34.21}_{\pm 1.23}$	$\mathbf{14.55}_{\pm 1.55}$	$\mathbf{21.62}_{\pm 1.83}$	$\mathbf{23.95}_{\pm 2.70}$	$\mathbf{26.17}_{\pm 3.28}$
	PhilBERTa	$67.45_{\pm 4.76}$	$29.46_{\pm 3.52}$	$10.65_{\pm 3.52}$	$17.73_{\pm 3.96}$	$20.47_{\pm 5.14}$	$23.60_{\pm 3.94}$
Rogationes	LaBERTa	$74.11_{\pm 1.18}$	$\mathbf{22.83}_{\pm 1.43}$	$3.47_{\pm 1.48}$	$\mathbf{7.81}_{\pm 3.45}$	$\mathbf{11.70}_{\pm 6.16}$	$\mathbf{13.64}_{\pm 4.48}$
	PhilBERTa	$\mathbf{73.23}_{\pm 2.11}$	$22.63_{\pm 1.43}$	$\mathbf{4.00}_{\pm 2.62}$	$6.52_{\pm 3.46}$	$9.03_{\pm 3.60}$	$11.55_{\pm 3.13}$
Overall	LaBERTa	$\mathbf{55.22}_{\pm 0.95}$	$\mathbf{41.18}_{\pm 0.61}$	$\mathbf{18.75}_{\pm 1.32}$	$\mathbf{27.22}_{\pm 2.02}$	$\mathbf{30.78}_{\pm 2.05}$	$\mathbf{33.81}_{\pm 2.11}$
	PhilBERTa	$62.40_{\pm 3.65}$	$33.52_{\pm 2.99}$	$12.97_{\pm 2.69}$	$20.65_{\pm 2.94}$	$23.39_{\pm 3.94}$	$26.58_{\pm 3.08}$

Table 5: Stratified Iterative Boundary Results, span length 10-20 characters(Mean $\pm$ 95% CI, HR stands for hit-rate))

Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8440–8451.

Giovanni Feo, editor. 2001. *Le carte bolognesi del secolo XI*. Number 53 in Fonti per la storia d'Italia medievale. Istituto storico Italiano per il Medio Evo, Roma.

Thomas Laurs. 2024. Towards a readability formula for latin. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)@ LREC-COLING-2024*, pages 170–175.

Koren Lazar, Benny Saret, Asaf Yehudai, Wayne Horowitz, Nathan Wasserman, and Gabriel Stanovsky. 2021. Filling the gaps in ancient akkadian texts: a masked language modelling approach. *arXiv preprint arXiv:2109.04513*.

Maddalena Modesti. 2012. *Studi per l'edizione delle carte bolognesi del secolo XII: prosopografia dei notai ed edizione critica di due cartulari notarili*. Pàtron Editore, Bologna.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Frederick Riemenschneider and Anette Frank. 2023. Exploring large language models for classical philology. *arXiv preprint arXiv:2305.13698*.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

Tag	Type-Token Ratio	Hapax-Token Ratio
Protocol	0.137	6.8%
Eschatocol	0.223	14.8%
Text Body	0.129	6.7%
Rogationes	0.310	19.7%

Table 6: Lexical richness and Hapax Legomena-to-token ratio across notary document sections.

9. Language Resource References

Evaluating Hierarchical Aggregation and LLM-Based Matching for Synset Selection in Ancient Greek

Luca Brigada Villa[1], Marco Passarotti[2], Chiara Zanchi[1], Riccardo Ginevra[2], Erica Fratellini[2], Eleonora Litta[2]
[1]University of Pavia
Piazza del Lino, 2, 27100 Pavia, Italy
{luca.brigadavilla, chiara.zanchi}@unipv.it
[2]Università Cattolica del Sacro Cuore
Largo A. Gemelli 1, 20123 Milan, Italy
{marco.passarotti, riccardo.ginevra, erica.fratellini, eleonoramaria.litta}@unicatt.it

Abstract

This paper presents a structured framework for WordNet synset selection applied to Ancient Greek lexical material. Starting from synonym definitions extracted from the Liddell–Scott–Jones (LSJ) lexicon, we compare two strategies: hierarchy-driven aggregation via bounded hypernym trees and LLM-based definitional matching with pairwise ranking. Graded human evaluation shows that structure-aware methods provide a robust baseline, particularly for nouns and verbs, while LLM-based reranking does not consistently improve performance, especially for highly ploysemous groups of synonyms. Beyond supporting the development of an Ancient Greek WordNet, the study highlights the methodological portability of the framework to other languages and lexical resources.

Keywords: WordNet, synset selection, Ancient Greek, LLMs

1. Introduction

Lexical-semantic resources such as WordNet (Miller, 1992; Fellbaum, 1998), FrameNet (Baker et al., 1998), and multilingual initiatives such as BabelNet (Navigli and Ponzetto, 2010) and Concepticon (List et al., 2025) provide structured inventories of senses organized into semantic networks or hierarchical taxonomies. These resources are widely used in linguistic research and natural language processing, where tasks often require linking lexical material to an appropriate sense. Despite the apparent simplicity of this objective, selecting a representative sense for a set of semantically related lexical items remains challenging, especially when the items reflect subtle sense distinctions or span multiple regions of the semantic hierarchy. The present work addresses this problem in the context of synonym groups derived from Ancient Greek lexical material, using definitions extracted from the *Liddell–Scott–Jones* (LSJ) lexicon as the basis for alignment to WordNet synsets.

A central difficulty arises from polysemy and semantic dispersion. Groups of near-synonymous words may correspond to closely related but distinct synsets, or they may cluster around different abstraction levels within the WordNet hierarchy. As a result, identifying a single synset that best captures the intended sense involves balancing structural coherence within the taxonomy and semantic similarity at the level of glosses.

In this work, we investigate two complementary strategies for synset selection. The former leverages the hierarchical organization of WordNet by constructing bounded hypernym trees and ranking candidate roots according to the portion of the retrieved semantic space they organize. The latter relies on semantic comparison between candidate glosses and a derived metadefinition, using large language models (LLMs) to perform pairwise preference judgments aggregated through an Elo ranking scheme.

We evaluate these strategies across multiple experimental conditions, including different parts of speech and both monosemous and polysemous synonym groups. Using graded human annotation and position-weighted scoring, we compare the effectiveness of structure-based and LLM-based approaches.

The paper is structured as follows. Section 2 situates the study within existing approaches to lexical-semantic alignment and synset selection. Section 3 presents the methodological framework, including the construction of synonym groups, definition extraction, Bag of Definitions (BoD) representation, synset retrieval, hypernym-tree structuring, and the two candidate selection strategies, together with the human evaluation protocol. Section 4 reports quantitative results, and Section 5 discusses their implications. Section 6 concludes and outlines directions for future research.

2. Related Work

Research on synset selection and lexical alignment has developed at the intersection of lexical semantics and natural language processing.

Proceedings of LT4HALA 2026 @ LREC 2026, pages 368–379.
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

In linguistic theory, questions of sense granularity, polysemy, and semantic field structure have long shaped debates about how lexical meaning should be partitioned and organized (Geeraerts, 2001, 2007; Taylor, 2003). In computational settings, these issues surface in tasks such as word sense disambiguation, gloss matching, and taxonomy alignment, where the goal is to associate lexical material with entries in structured semantic inventories such as WordNet. Selecting an appropriate synset for a group of near-synonymous forms requires reconciling fine-grained lexicographic distinctions with higher-level taxonomic abstraction.

Within natural language processing, this problem has been studied primarily under the umbrella of word sense disambiguation and gloss-based matching. Early approaches relied on lexical overlap between context and glosses, most notably variants of the Lesk algorithm (Lesk, 1986; Banerjee and Pedersen, 2002). Subsequent work incorporated distributional semantics, enabling similarity comparison between contextual or definitional representations in vector space (Mikolov et al., 2013; Pennington et al., 2014). More recently, contextualized encoders have been employed to model interactions between glosses and lexical contexts (Huang et al., 2019), further refining sense selection mechanisms.

Beyond local similarity, several studies have emphasized the importance of taxonomic structure in guiding sense choice. The hierarchical organization of WordNet has been used to constrain disambiguation, promote semantic coherence, and support alignment across resources (Navigli, 2009; Ponzetto and Navigli, 2009). From a linguistic standpoint, hierarchical relations provide a principled means of balancing abstraction and specificity, though care must be taken to avoid systematic preference for overly general synsets.

Cross-lingual lexicoraphic alignment, especially when involving historical languages, introduces additional complexity. Mapping definitions from classical or non-English lexica to English-based resources such as WordNet requires mediating between distinct lexicographic traditions and potentially divergent sense inventories. Prior work on multilingual WordNets and lexical alignment (Bond and Foster, 2013) highlights both the potential and the challenges of projecting sense distinctions across languages.

More recently, large language models (LLMs) have been applied to definitional comparison and semantic matching tasks, including gloss-based word sense disambiguation and definition ranking. Work on contextualized sense embeddings has shown that transformer models can align lexical items with WordNet glosses, inducing sense-specific representations through similarity between distributional and definitional embeddings (Loureiro et al., 2022; Scarlini et al., 2020). In these approaches, synset selection is typically framed as direct similarity scoring or comparative ranking of glosses. While such methods effectively capture semantic relatedness at the level of definitional content, they generally abstract away from the explicit hierarchical organization of the lexical taxonomy. The interaction between hierarchical structure and LLM-driven semantic comparison remains comparatively underexplored.

The present study contributes to this line of research by explicitly contrasting structure-based aggregation within the WordNet hierarchy with LLM-based semantic matching, evaluated through graded human annotation in the context of Ancient Greek lexical material aligned to WordNet.

3. Methodology

Our methodology addresses the problem of mapping synonym groups to WordNet synsets by relying on definitional evidence rather than on lexical items alone. Starting from dictionary definitions associated with the lemmas in each group, we construct an intermediate representation that captures the semantic content of the group as a whole. This representation is then used to identify and organize candidate synsets from the target inventory. The following subsections describe the construction of the data representation, the procedures used to retrieve and structure candidate synsets, and the methods adopted to select and evaluate final candidates.

3.1. Synonym groups

We start from 16 synonym groups previously extracted by Marchesi et al. (2025)[1]. The dataset (Appendix A) is balanced by part of speech, comprising four groups for each of the following categories: verbs, nouns, adjectives, and adverbs. The 16 groups are further balanced with respect to lexical ambiguity, with eight groups populated by predominantly monosemous lemmas and eight by predominantly polysemous lemmas. Each group contains a set of Ancient Greek lemmas intended to be near-synonymous in at least one sense. Throughout the paper, the unit of analysis is the *group* rather than individual lemmas, since the final target is a synset-level mapping for the group as a whole. Groups are constructed based on shared definitional content rather than strict synonymy. They should therefore be interpreted as

[1]The dataset was created by extracting data from back-translation dictionaries and it is available at https://github.com/unipv-larl/llms-ag/tree/main/Data%20for%20fine%20tuning.

semantically coherent clusters reflecting overlapping regions of meaning, rather than sets of fully interchangeable lexical items. This perspective aligns with a relatively shallow notion of synonymy, whereby lexical items are grouped based on partial substitutability or shared semantic content rather than full equivalence, as is also the case in WordNet synsets (Fellbaum, 1998).

3.2. Definition extraction from LSJ

For each lemma in a synonym group, we extract all definitional material available in the corresponding entry of the online version of the LSJ (Liddell et al., 1996). LSJ entries are provided in a structured XML format, where definitions are organized into `senses` (and potentially nested `subsenses`). For each lemma, we collect the list of English definition fragments associated with every sense and subsense, without attempting prior lemma-level sense disambiguation. This choice is deliberate: LSJ definitions are often highly granular and fragmentary, and synonym groups frequently conflate multiple closely-related senses, making pre-disambiguation both costly and error-prone at this stage.

The output of this step is a multiset of short English definition strings per lemma. These are subsequently merged at the group level (Section 3.3).

3.3. Bag of Definitions representation

We represent each synonym group via a *Bag of Definitions* (BoD), defined as the multiset union of the LSJ definitions extracted for all lemmas in the group:

$$\text{BoD} = \{d_1, d_2, \ldots, d_n\}$$

where each d_i is an English LSJ definition. Intuitively, the BoD acts as a semantic fingerprint for the synonym group, aggregating definitional evidence across lemmas while remaining agnostic to within-lemma sense boundaries.

3.4. Synset retrieval via BoD–gloss matching

The mapping task is formulated as a retrieval problem: given a BoD query representing a synonym group, we rank all candidate synsets in Open English WordNet (OEWN) by semantic similarity between the BoD text and a synset textual representation.

Synset representations. For each OEWN synset, we build a short document consisting of its gloss. This choice aligns the target representation with the definitional nature of the BoD.

POS filtering. Since synonym groups are POS-homogeneous, we restrict retrieval to synsets with the corresponding POS (verb, noun, adjective, adverb). This reduces the search space and mitigates systematic errors due to cross-POS ambiguity.

Bi-encoder models. We compute BoD–synset similarity using four sentence embedding models:

- `all-mpnet-base-v2` (Song et al., 2020; Reimers and Gurevych, 2019),
- `e5-base-v2` (Wang et al., 2022),
- `bge-base-en-v1.5` (Xiao et al., 2023),
- `all-MiniLM-L6-v2` (Wang et al., 2020; Reimers and Gurevych, 2019).

Each model independently encodes the BoD query and the synset document into a fixed-dimensional vector; synsets are ranked by cosine similarity. For each group we obtain a complete ranking over the POS-filtered synset inventory.

3.5. Hypernym-tree structuring of top-ranked synsets

Retrieval produces a ranked list, but the highest-ranked region may still contain semantically heterogeneous candidates, especially when BoDs mix multiple related senses. To introduce structure, we compute hypernym trees for synsets in the top region of the ranking and use these trees as an abstraction mechanism.

All WordNet access and hierarchical traversals were implemented using the `wn` Python library (Goodman and Bond, 2021), which provides programmatic access to synsets and lexical relations.

Hypernym trees can be constructed only for nouns and verbs, since in WordNet these parts of speech are organized into well-formed hierarchical taxonomies. Adjectives and adverbs, by contrast, lack a comparable hypernym–hyponym structure in WordNet and are instead connected primarily through non-taxonomic relations (e.g. similar-to, pertains-to). For this reason, the procedure described in this section applies exclusively to nominal and verbal synsets.

For the purposes of this work, a hypernym tree is defined as the subgraph induced by taking a synset as the root and iteratively following the hyponym relation in the target inventory. Starting from an individual synset, its hypernym tree therefore consists of the synset itself and the set of more specific synsets that fall under it in the WordNet hierarchy, organized as a rooted structure in which lower nodes correspond to increasingly specific semantic categories. In this sense, hypernym trees

capture the semantic scope covered by a synset, rather than its ancestry, and provide a way to characterize how broadly or narrowly a candidate synset organizes the retrieved semantic space.

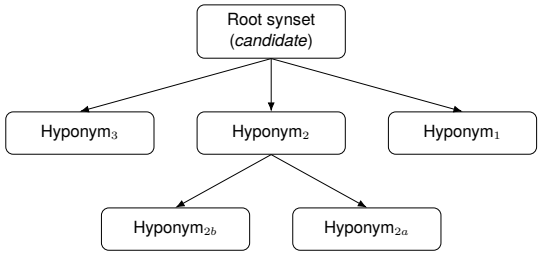


Figure 1: Schematic hypernym tree used as an abstraction device. A candidate synset is treated as a root, and the tree expands downward by iteratively following hyponym relations, capturing the semantic scope organized by the root.

By projecting retrieved synsets onto the hypernym trees rooted at candidate synsets, we move from a flat ranked list to a structured representation in which candidates can be compared in terms of the portion of the hierarchy they subsume. This abstraction makes it possible to reason about candidate synsets not only in terms of individual similarity scores, but also in terms of their capacity to organize related candidates within the lexical hierarchy. To avoid selecting candidates with excessively broad semantic scope, we restrict the procedure to hypernym trees rooted at synsets that have at least one hypernym; synsets without hypernyms are therefore excluded from consideration as roots. Concretely, given the top-100 synsets retrieved for a synonym group, we construct all possible three-level hypernym trees in which these candidates participate; that is, for each synset we consider (i) the tree rooted at the synset itself, (ii) the tree rooted at its direct hypernym, and (iii) the tree rooted at the hypernym of that hypernym. With reference to the schematic representation in Figure 1, synsets from the top-100 retrieved set may occupy different positions within the constructed trees: they can serve as roots, appear at intermediate levels, or occur among the lower-level hyponyms.

Because trees are constructed independently for each eligible root, partial structural overlap between trees may arise (e.g., when a synset and its hypernym both generate trees). We deliberately do not merge such overlapping subtrees into deeper composite structures. Merging would systematically favor higher-level synsets in the hierarchy, as they would inherit coverage from multiple lower-level trees and thus accumulate artificially inflated scope. By keeping trees distinct and limiting their depth, we ensure that each candidate is evaluated on the basis of how much of the top-100 retrieved set is locally organized within at most two hyponymic levels, rather than along arbitrarily long hypernym chains.

3.6. Candidate selection strategies

We compare two strategies for selecting final candidate synsets from the retrieved set. The applicability of these strategies depends on the part of speech, reflecting structural differences in the organization of WordNet. In particular, the hypernym-based strategy 3.6.1 applies only to nouns and verbs, for which hypernym–hyponym hierarchies are available.

3.6.1. Strategy A: hypernym-tree coverage

Strategy A is applied exclusively to nominal and verbal synsets. Rather than operating directly on the initially retrieved synsets, this strategy derives candidate synsets from the roots of the hypernym trees constructed over the retrieved set.

Starting from the top-ranked retrieved synsets, we construct a forest of eligible hypernym trees (cf. Section 3.5). Each distinct root of these trees – whether or not it appears in the original retrieved list – is treated as a candidate synset. Candidate synsets are ranked by the cardinality of the intersection between the original top-100 retrieved synsets and the set of nodes in their corresponding hypernym trees. Roots whose trees cover more retrieved synsets are preferred, as they provide broader yet structured abstraction over the retrieved semantic space.

3.6.2. Strategy B: LLM-based metadefinition matching

Strategy B selects and ranks candidate synsets by directly comparing their glosses to a metadefinition derived from the corresponding BoD, using pairwise judgments produced by LLMs.

Metadefinition construction. For each synonym group, the associated BoD is first transformed into a single metadefinition. This step is carried out by prompting three different LLMs, namely `Qwen2.5-3B-Instruct-GGUF` (Yang et al., 2024, 2025), `gemma-2b-it-GGUF` (Mesnard et al., 2024), and `Mistral-7B-Instruct-v0.2-GGUF` (Jiang et al., 2023) with identical instructions, in order to reduce model-specific biases and increase robustness (see Appendix B for the prompt text and parameters). The resulting metadefinitions provide an abstract semantic target against which candidate synsets are evaluated.

Candidate set. The set of candidate synsets depends on the part of speech. For nouns and verbs, candidates consist of all synsets that serve as roots of hypernym trees covering at least two synsets from the original top-100 cosine-similarity retrieval set (cf. Sections 3.4 and 3.5). This constraint ensures that only roots with a minimal degree of semantic support in the retrieved set are considered. For adjectives and adverbs, which lack a hypernym–hyponym hierarchy in WordNet, candidates are defined as the top-50 synsets ranked by cosine similarity.

Pairwise comparison. Candidate synsets are ranked via a sequence of pairwise comparisons. At each step, two candidates are sampled at random from the candidate set. The LLM is prompted with the metadefinition and the glosses of the two synsets, and asked to determine which gloss is semantically closer to the metadefinition (see Appendix C for the prompt text and parameters). The model is required to output a binary preference, without justification.

Elo-based ranking. The outcomes of the pairwise comparisons are aggregated using an Elo rating system (Elo, 1978). For each synonym group, the number of pairwise comparisons is set to five times the number of candidate synsets, resulting in a total of $5 \times |C|$ Elo matches, where $|C|$ denotes the size of the candidate set. Each candidate synset i is assigned an initial Elo score R_i, which is updated after each comparison. Given two candidates i and j, the expected score of i is computed as

$$E_i = \frac{1}{1 + 10^{(R_j - R_i)/400}}.$$

After observing the outcome of the comparison, the rating of i is updated according to

$$R'_i = R_i + K(S_i - E_i),$$

where $S_i \in \{0, 1\}$ indicates whether i is preferred by the LLM, and K is a scaling factor, set to $K = 24$ in our experiments. The rating of j is updated analogously. Over successive comparisons, synsets that are consistently preferred with respect to the metadefinition accrue higher Elo scores, resulting in a global ranking of candidates. An Elo-based ranking is preferred over direct similarity scoring because it aggregates relative judgments across many local comparisons, reducing sensitivity to score calibration and allowing a stable global ordering to emerge from noisy pairwise preferences.

3.7. Human annotation for evaluation

The outputs of all candidate selection strategies – hypernym-tree coverage for nouns and verbs (Strategy A), output of the bi-encoder models for adjectives and adverbs and LLM-based matching (Strategy B) – were evaluated through manual annotation.

For each embedding model and for each synonym group, we extracted the top-5 synsets produced by each strategy. Two independent annotators assessed the degree to which each proposed synset adheres to the semantic content of the corresponding synonym group. Annotations were assigned on a four-level ordinal scale reflecting increasing degrees of semantic relatedness: strong adherence (3), partial adherence (2), marginal adherence (1), and no adherence (0).

In addition, a third independent annotator selected one synset per synonym group that best represents the intended meaning of the group, without access to the experimental results. This independently chosen synset serves as a reference point for evaluating the alignment between the automatically ranked candidates and a human-identified representative sense.

4. Results

This section reports the empirical evaluation of the proposed strategies. We first assess inter-annotator agreement, then compare performance using both hit-based and graded score measures across embedding models, parts of speech, and monosemous versus polysemous groups. Given the limited size of the dataset (16 groups), results should be interpreted as indicative trends rather than as statistically significant comparisons.

4.1. Inter-annotator agreement

We assessed annotation reliability using Cohen's κ. Agreement was computed in two ways: (i) a binary version collapsing the scale into non-adherence (0) versus adherence (1–3), and (ii) a quadratic weighted κ over the full four-level ordinal scale (0–3), thereby accounting for the graded nature of the judgments.

Table 1 reports agreement per group. Across groups, binary κ yielded a macro-average of 0.506 (sd = 0.253), with values ranging from 0.109 to 0.820 and a pooled estimate of 0.520. Ordinal κ produced a slightly higher macro-average of 0.529 (sd = 0.311), with values ranging from 0.019 to 0.899 and a pooled estimate of 0.611.

Overall, agreement can be characterized as moderate, with substantial variability across groups. In general, monosemous groups tend to exhibit higher agreement than polysemous ones, suggesting that semantic dispersion within polysemous sets increases annotation difficulty. The higher pooled value observed for ordinal κ

		polysemous								monosemous							
	group	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
κ	binary	0.654	0.259	0.707	0.205	0.622	0.534	0.182	0.354	0.820	0.790	0.762	0.724	0.146	0.790	0.109	0.434
	ordinal	0.655	0.269	0.889	0.249	0.694	0.770	0.082	0.198	0.744	0.899	0.776	0.810	0.158	0.867	0.019	0.377

Table 1: Inter-annotator agreement (Cohen's κ) per evaluation group. Binary κ refers to the 0 vs. 1–3 distinction; ordinal κ corresponds to quadratic weighted κ computed over the full 0–3 scale.

indicates that, even where exact level matches were inconsistent, annotators often agreed on the relative degree of adherence.

4.2. Hit-based evaluation

We first evaluate performance in terms of hit rate. A hit is recorded for a given embedding–strategy combination when at least one synset among the top-5 candidates is annotated by a human evaluator as showing at least marginal adherence (score ≥ 1) to the synonym group. For each combination, we report the proportion of groups for which a hit was observed. Hit rates are computed separately for each annotator and then averaged, ensuring that evaluation units remain independent at the group level.

Table 2 summarizes results across monosemous and polysemous groups as well as across parts of speech.

Overall, the non-LLM baseline ("none"), corresponding to hypernym-tree coverage for nouns and verbs and direct cosine top-5 selection for adjectives and adverbs, consistently achieves the highest hit rates across embeddings. In particular, the `bge-base-en-v1.5` model reaches near-ceiling performance for verbs and nouns, with hit rates of 1.00 in the two categories.

By contrast, LLM-based reranking generally reduces hit rates across embeddings. Among the three LLMs, Qwen tends to preserve performance more effectively than Mistral and Gemma, though results vary depending on embedding and part of speech. The drop in performance is particularly noticeable for polysemous groups, suggesting that reranking may introduce additional instability in semantically dispersed contexts.

Across parts of speech, nouns and verbs show higher hit rates overall, reflecting the structural advantage provided by hypernym-based aggregation. Adjectives and adverbs exhibit greater variability, especially under LLM reranking, likely due to the absence of hierarchical structure and stronger reliance on surface semantic matching.

To further assess alignment quality, Table 3 reports the proportion of groups for which the independently annotated gold synset – or one of its hypernyms (up to two levels up) or hyponyms (up to two levels down) – was retrieved among the top-5 candidates. This metric is substantially stricter, as it requires proximity to a specific reference synset rather than general semantic acceptability.

Under this criterion, performance decreases markedly across all embeddings and strategies. Even in the strongest configurations (e.g., `e5-base` without reranking), retrieval rates remain moderate and vary considerably across parts of speech. LLM-based reranking does not consistently improve gold retrieval and in several cases reduces it, particularly for polysemous groups. These results indicate that while multiple semantically acceptable synsets may be retrieved (as reflected in the hit-based evaluation), exact or near-exact alignment with a single gold synset is considerably more demanding.

Taken together, the hit-based and gold-based analyses suggest that structural aggregation supports the retrieval of semantically plausible candidates, but that convergence toward a specific target synset—especially in polysemous settings—remains challenging.

4.3. Graded score evaluation

While hit-based evaluation captures whether at least one acceptable synset is retrieved, it does not reflect how well relevant synsets are ranked within the top-5. We therefore compute a graded relevance score based on the four-level annotation scheme.

For each synonym group, each synset in the top-5 receives a relevance value in the range 0–3. Scores are aggregated using a normalized position-weighted scheme, assigning higher weight to higher-ranked positions and normalizing the resulting sum to ensure comparability across models and strategies.

Formally, let $r_k \in \{0, 1, 2, 3\}$ denote the relevance score assigned to the synset at rank k, for $k = 1, \ldots, K$ (with $K = 5$ in our setting). We adopt a logarithmic discounting scheme inspired by Discounted Cumulative Gain (DCG), assigning weight

$$w_k = \frac{1}{\log_2(k+1)}.$$

The normalized position-weighted score is then computed as:

$$S = \frac{\sum_{k=1}^{K} \frac{r_k}{\log_2(k+1)}}{\sum_{k=1}^{K} \frac{3}{\log_2(k+1)}}.$$

Embedding	+ LLM	poly	mono	verbs	nouns	adjs	advs
all-MiniLM-L6-v2	none	0.6875	0.8125	0.75	0.75	0.875	0.625
	+ mistral	0.4375	0.5625	0.5	0.625	0.375	0.5
	+ qwen	0.6875	0.625	0.75	0.875	0.625	0.375
	+ gemma	0.4375	0.5	0.5	0.625	0.375	0.375
all-mpnet-base-v2	none	0.6875	0.6875	0.625	0.5	0.875	0.75
	+ mistral	0.25	0.5	0.375	0.375	0.375	0.375
	+ qwen	0.5625	0.625	0.625	0.375	0.625	0.75
	+ gemma	0.3125	0.5	0.5	0.375	0.375	0.375
bge-base-en-v1.5	none	0.8125	0.875	1	1	0.625	0.75
	+ mistral	0.5	0.4375	0.625	0.375	0.375	0.5
	+ qwen	0.8125	0.6875	0.875	0.75	1	0.375
	+ gemma	0.5	0.4375	0.625	0.375	0.375	0.5
e5-base-v2	none	0.875	0.8125	0.75	1	0.75	0.875
	+ mistral	0.375	0.4375	0.25	0.625	0.375	0.375
	+ qwen	0.6875	0.625	0.75	0.625	0.625	0.625
	+ gemma	0.375	0.4375	0.25	0.625	0.375	0.375

Table 2: Proportion groups for which at least one valid synset was identified (per category and part of speech) across embedding models and reranking strategies. “None” denotes the absence of LLM reranking, corresponding either to the hypernym-based strategy (when applicable) or to direct top-5 selection by cosine similarity.

Embedding	+ LLM	poly	mono	verbs	nouns	adjs	advs
all-MiniLM-L6-v2	none	0.125	0.25	0.5	0	0	0.25
	+ mistral	0	0	0	0	0	0
	+ qwen	0.125	0.25	0.25	0.5	0	0
	+ gemma	0	0	0	0	0	0
all-mpnet-base-v2	none	0	0.25	0.25	0.25	0	0
	+ mistral	0	0.25	0	0.25	0	0.25
	+ qwen	0.125	0.125	0.25	0	0.25	0
	+ gemma	0	0.25	0	0.25	0	0.25
bge-base-en-v1.5	none	0.25	0.375	0.5	0.25	0	0.5
	+ mistral	0.25	0	0.25	0.25	0	0
	+ qwen	0.125	0	0	0	0.25	0
	+ gemma	0.25	0	0.25	0.25	0	0
e5-base-v2	none	0.25	0.5	0.5	0.5	0	0.5
	+ mistral	0.125	0.125	0	0.5	0	0
	+ qwen	0.25	0.5	0.5	0.25	0.25	0.5
	+ gemma	0.125	0.125	0	0.5	0	0

Table 3: Proportion of groups for which the gold synset – or one of its hypernyms (up to two levels up) or hyponyms (up to two levels down) – was retrieved among the top-5 candidates (polysemous vs. monosemous targets and by part of speech) across embedding models and reranking strategies. “None” denotes the absence of LLM reranking, corresponding either to the hypernym-based strategy (when applicable) or to direct top-5 selection by cosine similarity.

The denominator corresponds to the maximum attainable score (i.e., when all $r_k = 3$), ensuring that $S \in [0, 1]$. This formulation prioritizes correct ranking of highly relevant synsets at top positions, which is crucial in a manual selection scenario where only a small number of candidates is retained.

Table 4 reports normalized scores across embeddings, distinguishing between polysemous and monosemous groups as well as parts of speech.

Across all embeddings, the non-LLM baseline (“none”) consistently achieves the highest graded scores. This pattern mirrors the hit-based results but is even more pronounced, indicating that LLM-based reranking not only reduces the likelihood of retrieving at least one acceptable synset, but also tends to lower the overall concentration of semantically appropriate candidates at higher ranks.

Among reranking strategies, Qwen generally performs better than Mistral and Gemma, though it rarely surpasses the baseline. In some

Embedding	+ LLM	poly	mono	verbs	nouns	adjectives	adverbs
all-MiniLM-L6-v2	none	0.2018	0.1493	0.1699	0.1164	0.3010	0.114
	+ mistral	0.0394	0.0795	0.0510	0.0981	0.0513	0.0372
	+ qwen	0.1125	0.1378	0.0774	0.2697	0.1141	0.0395
	+ gemma	0.0424	0.0785	0.0510	0.0988	0.0513	0.0407
all-mpnet-base-v2	none	0.1464	0.1476	0.1115	0.1189	0.2865	0.0712
	+ mistral	0.0462	0.0779	0.0211	0.0994	0.0392	0.0885
	+ qwen	0.0552	0.1016	0.0925	0.0626	0.1037	0.0549
	+ gemma	0.0498	0.0779	0.0282	0.0994	0.0392	0.0885
bge-base-en-v1.5	none	0.2288	0.1226	0.1631	0.1746	0.1833	0.1818
	+ mistral	0.0598	0.0576	0.0688	0.0485	0.0267	0.0909
	+ qwen	0.1702	0.1111	0.1862	0.0943	0.2402	0.0419
	+ gemma	0.0598	0.0576	0.0634	0.0485	0.0321	0.0909
e5-base-v2	none	0.1674	0.1599	0.0931	0.1716	0.2157	0.1740
	+ mistral	0.0775	0.0752	0.0283	0.1344	0.0925	0.0503
	+ qwen	0.1182	0.1073	0.0556	0.1113	0.1442	0.1397
	+ gemma	0.0704	0.0697	0.0283	0.1233	0.0784	0.0503

Table 4: Normalized position-weighted scores (polysemous vs. monosemous and by part-of-speech) across embedding models and reranking strategies. "None" denotes the absence of LLM reranking, corresponding either to the hypernym-based strategy (when applicable) or to direct top-5 selection by cosine similarity.

cases, QWEN narrows the gap (e.g., nouns with `all-MiniLM-L6-v2` and adjectives with `bge-base-en-v1.5`), but improvements over the non-LLM strategy are limited and inconsistent.

With respect to monosemy and polysemy, performance differences are relatively small in the baseline condition, whereas reranking tends to amplify variability across groups. Polysemous sets show greater sensitivity to reranking, suggesting that pairwise LLM judgments may struggle when the retrieved candidates span multiple related semantic regions.

Part-of-speech differences remain visible in the graded scores. Nouns and verbs, which benefit from hypernym-based structural aggregation in the baseline condition, generally achieve more stable performance. Adjectives and adverbs exhibit greater fluctuation under reranking, consistent with their reliance on direct semantic comparison rather than hierarchical organization.

Overall, the graded evaluation confirms the robustness of the structurally grounded baseline and indicates that LLM-based pairwise reranking does not consistently improve ranking quality under the present configuration.

5. Discussion

The results highlight a consistent structural advantage for the hypernym-based strategy. Across embeddings and evaluation metrics, the non-LLM baseline – relying on hierarchical aggregation for nouns and verbs and direct cosine similarity for adjectives and adverbs – achieves the highest hit rates and graded scores. For nouns and verbs, these results suggest that organizing retrieved synsets through local hypernym structure provides a robust mechanism for concentrating semantically coherent candidates near the top of the ranking.

By contrast, LLM-based pairwise reranking does not consistently improve performance. Although pairwise judgments offer a flexible way to compare glosses against metadefinitions, they appear sensitive to semantic dispersion within the candidate set, particularly in polysemous groups. The Elo aggregation scheme stabilizes local comparisons, but the overall ranking remains dependent on the quality of individual LLM preferences, which may introduce additional variance.

Differences between monosemous and polysemous groups further reinforce this interpretation. Polysemous sets tend to fragment across multiple regions of the lexical hierarchy, reducing the effectiveness of both structural aggregation and semantic comparison. In such contexts, local hierarchical concentration appears more reliable than iterative pairwise reranking.

Overall, these findings suggest that explicitly leveraging lexical hierarchy remains a strong baseline for synset selection, while LLM-based semantic matching, at least in its current pairwise configuration, does not systematically outperform structure-aware methods. This points toward the potential benefit of hybrid approaches that combine hierarchical constraints with more globally cal-

ibrated semantic scoring.

6. Conclusion and Future Perspectives

This study compared hierarchy-driven aggregation and LLM-based definitional comparison for synset selection using Ancient Greek lexical material aligned to WordNet. The results indicate that structure-aware methods provide a strong and stable baseline, particularly for nouns and verbs, whereas LLM-based reranking does not consistently yield improvements, especially in polysemous contexts.

Beyond its specific contribution to the development of an Ancient Greek WordNet, the main strength of this work is methodological. The proposed framework – combining definitional extraction, hierarchical aggregation, and graded evaluation – is portable and applicable to other languages and lexical resources, offering a principled way to examine the interaction between synonym groupings and lexical hierarchy.

A limitation of the approach is its reliance on predefined synonym groups. Where curated synonym dictionaries are available, the method can be directly applied and may help reveal differences in semantic granularity across resources. In languages lacking such materials, synonym sets can be generated automatically with LLMs; although noisier, they provide a feasible starting point.

Future work will extend the framework cross-linguistically and explore hybrid pipelines in which automatically induced synonym sets are refined through hierarchical validation and human evaluation.

7. Bibliographical References

Satanjeev Banerjee and Ted Pedersen. 2002. An adapted lesk algorithm for word sense disambiguation using wordnet. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 2276 of *Lecture Notes in Computer Science*, pages 136–145. Springer, Berlin, Heidelberg.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362, Sofia, Bulgaria. Association for Computational Linguistics.

Arpad E. Elo. 1978. *The Rating of Chessplayers, Past and Present*. Arco Publishing, New York.

Dirk Geeraerts. 2001. The definitional practice of dictionaries and the cognitive semantic conception of polysemy. *Lexicographica*, 17:6–21.

Dirk Geeraerts. 2007. Lexicography. In Dirk Geeraerts and Hubert Cuyckens, editors, *The Oxford Handbook of Cognitive Linguistics*, pages 1160–1175. Oxford University Press, New York.

Michael Wayne Goodman and Francis Bond. 2021. Intrinsically interlingual: The wn python library for wordnets. In *Proceedings of the 11th Global Wordnet Conference*, pages 100–107, University of South Africa (UNISA). Global Wordnet Association.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for word sense disambiguation with gloss knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3509–3514, Hong Kong, China. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, SIGDOC '86, page 24–26, New York, NY, USA. Association for Computing Machinery.

Daniel Loureiro, Alípio Mário Jorge, and Jose Camacho-Collados. 2022. Lmms reloaded: Transformer-based sense embeddings for disambiguation and beyond. *Artificial Intelligence*, 305:103661.

Beatrice Marchesi, Annachiara Clementelli, Andrea Maurizio Mammarella, Silvia Zampetta, Erica Biagetti, Luca Brigada Villa, Virginia Mastellari, Riccardo Ginevra, Claudia Roberta Combei, and Chiara Zanchi. 2025. Towards the Semi-Automated Population of the Ancient Greek WordNet. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 647–658, Cagliari, Italy. CEUR Workshop Proceedings.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak,

Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2).

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Simone Paolo Ponzetto and Roberto Navigli. 2009. Large-scale taxonomy mapping for restructuring and integrating wikipedia. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, IJCAI'09, page 2083–2088, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020. With more contexts comes better performance: Contextualized sense embeddings for all-round word sense disambiguation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3528–3539, Online. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: masked and permuted pre-training for language understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

John R. Taylor. 2003. *Linguistic Categorization*, 3 edition. Oxford University Press, Oxford.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei

Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

8. Language Resource References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Henry George Liddell, Robert Scott, Henry Stuart Jones, and Roderick McKenzie. 1996. A greek-english lexicon. `http://www.perseus.tufts.edu`. Digital edition, Perseus Digital Library.

Johann Mattis List, Annika Tjuka, Frederic Blum, Alžběta Kučerová, Carlos Barrientos Ugarte, Christoph Rzymski, Simon Greenhill, and Robert Forkel, editors. 2025. *CLLD Concepticon 3.4.0*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

George A. Miller. 1992. WordNet: A lexical database for English. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden. Association for Computational Linguistics.

A. Groups of synonyms

Group 1: ἀποκείρω, ἐπικρύπτω, κατέχω, καταπαύω, καταλύω, διαλύω, παύω, καταστρέφω, κλέπτω, κρύπτω, κατεργάζομαι, ὑποστέλλω.
Semantic domain: concealment, suppression, removal, end

Group 2: τρύχω, βασανίζω, στρεβλόω, παρακρούω, σπάω, ἑλκόω, ἀρταμέω, κνάπτω, διασπάω, διασπαράσσω, σπαράσσω, δάκνω, λυπέω, ἀλγύνω.
Semantic domain: physical or psychological affliction, damage

Group 3: τύπος, περιγραφή, χαρακτήρ, χάραγμα, ὑπογραφή, στίβος.
Semantic domain: mark, representation, inscription

Group 4: αἰθήρ, πόλος, ὄροφος, ὀροφή, οὐρανός, κέραμος, εὐδία, τέγος, ἀήρ, ῥυμός, Διοπετής, εὐφημία.
Semantic domain: upper space, sky, atmosphere, roof

Group 5: ἀλίγκιος, ἐμφερής, παραπλήσιος, ἴσος, προσφερής, ὅμοιος, προσόμοιος.
Semantic domain: similarity, equivalence, likeness

Group 6: χθόνιος, ὑπόνομος, κατάγειος, νέρτερος, ὕπουλος
Semantic domain: subterranean, hidden, lower realm

Group 7: πόθεν, πῃ, ποῖ, τάχα, ἄν.
Semantic domain: indefiniteness, uncertainty

Group 8: τέως, νῦν, τρίς, ἀεί, ἄλλοτε, ἐνίοτε, πηνίκα, καίριος, πολλάκις, ὁπηνίκα.
Semantic domain: (indefinite) temporal reference

Group 9: ῥυπάω, αὐχμέω.
Semantic domain: filth, neglect, squalor

Group 10: προοιμιάζομαι, ὑπαγορεύω.
Semantic domain: discourse initiation

Group 11: εὐθηνία, εὐπραγία.
Semantic domain: prosperity, success, well-being

Group 12: ὑποθήκη, φραδή, ὑπόνοια.
Semantic domain: suggestion, advice

Group 13: ἄκοπος, ἀκάματος.
Semantic domain: absence of fatigue, tirelessness

Group 14: ἀνήλιος, ἀναύγητος.
Semantic domain: absence of sun, darkness

Group 15: κάτωθεν, ἐκεῖσε, ἐκεῖ.
Semantic domain: direction, location

Group 16: ἑτέρωθι, ἀλλαχοῦ, ἄλλοθι, ἄλλος, ἑτέρωθεν, ἑτέρωσε, ἀλλαχόθεν.
Semantic domain: elsewhere

B. Prompt to generate the metadefinition from the BoD

Prompt:

```
You are a lexicographer.

Write exactly ONE definition (ONE
sentence) for the dominant sense in
the BoD.
Requirements:
- Start immediately with the
definition.
- No lists, no repetition, no
semicolons.
- KEEP IT SHORT!
- Max 25 words.
- Use ONLY information from the BoD.
- JUST ONE SENTENCE.
- STOP after the definition.

BoD:
{bod_block}

Definition:
```

Parameters: `max_tokens=80, temperature=0.2, top_p=0.9, repeat_penalty=1.15`

C. Prompt to select between two candidates

Prompt:

```
Metadefinition:
{metadef}

Gloss A:
{gloss_a}

Gloss B:
{gloss_b}

Which gloss matches the
metadefinition better?
Answer ONLY with ``A'' or ``B''.
```

Parameters: `max_tokens=2, temperature=0.0`

Miktub: A Manuscript Dataset of Historical Maltese for Handwritten Text Recognition

Thomas Koppens, Claudia Borg
Department of Artificial Intelligence, University of Malta
{thomas.koppens.22, claudia.borg}@um.edu.mt

Abstract

The digitisation of handwritten historical material is essential for preserving cultural heritage and enabling search and computational analysis. For Maltese, historical handwritten resources are scarce, and, to the best of current knowledge, no public handwritten text recognition (HTR) dataset for historical Maltese exists. We introduce a Manuscript Dataset of Historical Maltese (Miktub), collected from the University of Malta Library Special Collections: 35 scanned pages transcribed by specialists and converted into a line-level HTR dataset. A key challenge was robust line extraction from heterogeneous pages; fully automatic line segmentation was insufficient, so we developed a semi-automatic pipeline combining horizontal projection profiling with lightweight post-processing and manual refinement to maximise line fidelity. We provide two annotation variants, including a corrected/standardised version (Miktub-COR) designed to improve consistency, accessibility, and downstream learning stability. We benchmark two strong public HTR models, HTR-VT and VAN, and report the best test performance of 4.68% character error rate (CER) and 13.59% word error rate (WER) on Miktub-COR with VAN. We will release Miktub publicly upon acceptance, along with scripts and splits, to support historical Maltese-language technology research.

Keywords: Handwriting Recognition, Maltese Historical Manuscripts

1. Introduction

Large-scale digitisation initiatives increasingly rely on Handwritten Text Recognition (HTR) to reduce the cost of transcribing handwritten documents, enabling search and computational study of collections that are otherwise locked in image form. Yet HTR performance depends heavily on the availability of labelled data and careful preparation of line-level training material. HTR is often seen as an extension of Optical Character Recognition (OCR), in which scanned or photographed documents are converted into machine-readable form through stages such as text-region detection, line/character segmentation, and recognition. Whilst all these processes are present, HTR must cope with substantial intra- and inter-writer variation, inconsistent character shapes, and irregular spacing, which makes it a considerably more complex recognition problem and motivates the use of more expressive modelling approaches.

This study addresses HTR for Historical Maltese Manuscripts. While HTR is used in several applied domains (e.g., medical documents and cheques), historical manuscripts introduce additional challenges such as physical degradation and ink bleed-through, which can distort character appearance and line structure. In practice, digitising historical collections still often relies on manual line segmentation and transcription—procedures that are repetitive, time-consuming, and require trained personnel. HTR can substantially accelerate this workflow by automatically segmenting and transcribing pages, leaving humans primarily with post-hoc correction; correspondingly, reducing recognition errors directly decreases the manual effort required to clean transcriptions.

Maltese is a Semitic language, with influence from Italian and English (Borg and Gatt, 2017). It is written in the Latin script, with a small set of orthographic diacritics and language-specific letters (notably ċ, ġ, ħ, ż). For HTR, the recogniser must still capture fine-grained distinctions introduced by diacritics, which are particularly vulnerable to noise (e.g., fading, bleed-through) in historical scans. The problem offers an opportunity to leverage techniques and architectures developed for other Latin-script settings, while still confronting the low-resource and historical-domain constraints specific to Maltese collections. The motivation for this work is rooted in cultural preservation and access. Many historically significant Maltese documents remain available only in handwritten form, and the scale of unprocessed archival material far exceeds the human resources typically available for full manual transcription.

Recent advances in neural HTR have made large gains in automation, but these systems are predominantly optimised for high-resource languages and domains; consequently, their effectiveness is limited when applied to historical Maltese texts. This project is therefore driven by the need to bridge that gap by developing and evaluating an HTR pipeline for the accurate transcription of historical handwritten Maltese, supporting preservation, scholarly study, and the broader dissemination of Malta's written heritage.

Concretely, the work pursues a model that min-

Proceedings of LT4HALA 2026 @ LREC 2026, pages 380–388.
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

imises transcription error, especially character error rate (CER), to reduce manual intervention in downstream transcription workflows. Accordingly, this work makes three primary contributions:

1. Miktub, a line-level dataset derived from 35 historical handwritten pages curated for diversity in authorship and difficulty (e.g., faded/struck-through content), enabling research in low-resource historical HTR.

2. A practical semi-automatic line segmentation workflow that bridges the gap between unreliable full automation and expensive manual segmentation, designed for heterogeneous scans

3. Strong baseline results with two state-of-the-art public architectures (HTR-VT and VAN), including analysis of how transcription consistency affects training and error rates.

For Maltese, especially historical Maltese, resources are limited; as far as we are aware, no historical Maltese HTR dataset currently exists, motivating the creation and release of Miktub.[1]

2. Literature Review

Early automated handwriting recognition relied on template matching and manually engineered features, inspired by online HTR methods that reconstruct stroke trajectories (Plamondon and Privitera, 1999). Systems used Hidden Markov Models (HMMs) and handcrafted descriptors (Plamondon and Srihari, 2000). These pipelines typically required explicit segmentation into characters or words and manual design choices for aligning feature sequences to HMM states.

Deep learning shifted the field toward end-to-end training. A key step was the introduction of Connectionist Temporal Classification (CTC), which enables learning alignments between long input sequences and shorter transcriptions without requiring pre-segmented character boundaries (Graves et al., 2006). Coupled with Multidimensional LSTMs (MDLSTMs), this provided robustness to two-dimensional distortions (e.g., slant and local shifts) while allowing line-level transcription without explicit segmentation (Graves and Schmidhuber, 2008). The standard CTC mechanism—including an explicit blank symbol to separate repeated characters—remains influential due to its suitability for HTR's alignment problem. Puigcerver (2017); Bluche and Messina (2017) found that MDLSTMs often do not substantially outperform faster alternatives based on a CNN feature extractor + BLSTM + CTC, especially when combined with data augmentation; these 1D recurrent models achieve strong accuracy while offering improved computational efficiency.

Beyond purely recurrent/CTC pipelines, attention-based encoder–decoder models became prominent: convolutional backbones produce feature maps, while attentive decoders learn soft alignments over image regions, improving robustness to irregular spacing, cursive joins, and local degradations (Bluche et al., 2017; Wu et al., 2019).

Hybrid objectives combining attention-based cross-entropy with CTC have also been used to stabilise training and improve convergence in low-data regimes. Transformer-based approaches unify visual encoding and sequence generation. TrOCR (Li et al., 2023) exemplifies this direction by pairing a Vision Transformer (ViT) encoder with a Transformer decoder: the encoder represents a line image as a sequence of patch embeddings and applies self-attention to capture long-range dependencies, while the decoder generates text autoregressively via cross-attention. Pre-training on large printed/handwritten corpora is central to this paradigm, enabling strong transfer when fine-tuned on line-level benchmarks.

2.1. HTR-VT and VAN

Two strong public architectures align well with the constraints of historical, low-resource HTR. HTR-VT (Li et al., 2025) adapts the ViT idea to handwriting by replacing patch embedding with a lightweight CNN that yields dense, high-resolution token embeddings aligned with stroke patterns. It introduces span feature masking (masking contiguous token spans during training) to encourage contextual reasoning rather than overfitting to local stroke cues, and it combines this with Sharpness-Aware Minimization (SAM) to improve data efficiency. Decoding is performed with CTC, preserving the benefits of alignment-free training. HTR-VT is particularly relevant here because it achieves strong results on the LAM dataset (2.8% CER / 7.4% WER), suggesting it may transfer well to similar historical letter collections.

The second architecture, VAN (Coquenet et al., 2023), is motivated by a major practical bottleneck in historical collections: the fragility of explicit line segmentation. Instead, VAN processes a paragraph image using a lightweight convolutional backbone and applies a recurrent vertical attention mechanism that scans top-to-bottom, implicitly isolating lines without an external segmentation stage. It iterates attention steps until all lines are covered, enabling end-to-end paragraph transcription, and has reported strong performance on historical datasets such as READ 2016 and IAM with relatively few parameters. Importantly, VAN's para-

[1] Code and data available here: https://github.com/MLRS/Miktub/

graph decoding still depends on a line recogniser, so line-level training (and thus line-aligned supervision) remains necessary, and robustness to unseen handwriting typically requires further fine-tuning.

2.2. Post-processing for HTR

Many HTR systems treat recognition as a two-stage problem: an optical model produces raw text, followed by a language-informed post-processing module that corrects likely errors (Li et al., 2023; Bluche and Messina, 2017). Traditional post-processing commonly integrates lexicons and n-gram language models into decoding, thereby reducing out-of-vocabulary errors and improving plausibility (Povey et al., 2011). However, two limitations are particularly acute in low-resource settings: these methods require carefully constructed corpora/dictionaries for LM training, and n-grams capture only short-range dependencies, limiting context-sensitive correction (Chen and Goodman, 1999; Bengio et al., 2003).

Neural post-processing addresses these issues by learning a mapping from noisy HTR output to corrected text using Seq2Seq architectures with attention, borrowing from machine translation and grammatical error correction (Neto et al., 2020). Because these models are data-driven, they can adapt to the characteristic error patterns of a given optical model without relying on handcrafted lexicons. Neto et al. (2020) conducts a comparative study evaluating statistical and neural correction approaches across multiple HTR datasets and reports that neural methods substantially outperform statistical baselines; notably, a Luong-attention Seq2Seq model corrected a markedly larger share of erroneous sentences than a widely used HMM-LM approach. This motivates prioritising neural correction when sufficient paired data exists, while still retaining lightweight statistical correctors as practical baselines.

2.3. HTR datasets

High-performance HTR models can be pre-trained on large public datasets before fine-tuning on a target domain to improve error rates (Parres and Paredes, 2023). Widely used datasets such as IAM (Marti and Bunke, 2002) and RIMES (Grosicki and El-Abed, 2011) consist of more recent documents with contemporary handwriting styles and show few signs of wear, suggesting they may not share many characteristics with our data. On the other hand, many benchmark datasets and other popular historical datasets were considered for the purpose of pre-training models before being fine-tuned on our dataset, including READ 2016 (Sánchez et al., 2016), Bentham (Sánchez et al., 2015), LAM (Cascianelli et al., 2022).

The target Maltese collection is characterised as early 20th-century cursive with occasional block-letter titles, exhibiting stains and bleed-through and spanning multiple authors over decades—placing it between older historical corpora and modern handwriting datasets. On this basis, READ 2016 is treated as a weaker match due to its highly specialised script/language, and Bentham is excluded primarily because it is single-author and thus offers limited stylistic variability for robust pre-training. In contrast, LAM is argued to be the closest overall match: it is large, historically degraded, spans decades (capturing handwriting drift), and resembles a letter-style manuscript domain.

3. Methodology

3.1. Data Source and Scope

We construct the Miktub dataset, named after the Maltese word for 'written', from historical material held by the University of Malta Library Special Collections Department. They provided 35 scanned page images together with full textual transcriptions, with explicit selection criteria aimed at maximising diversity and practical difficulty: multiple authors over multiple decades (hence multiple handwriting styles as shown in Fig. 1) and the inclusion of phenomena common in archival documents, such as struck-through text, fading, and other degradations. The intent is to support HTR development in a realistic setting and to promote robustness when models are deployed on unseen authors and pages.

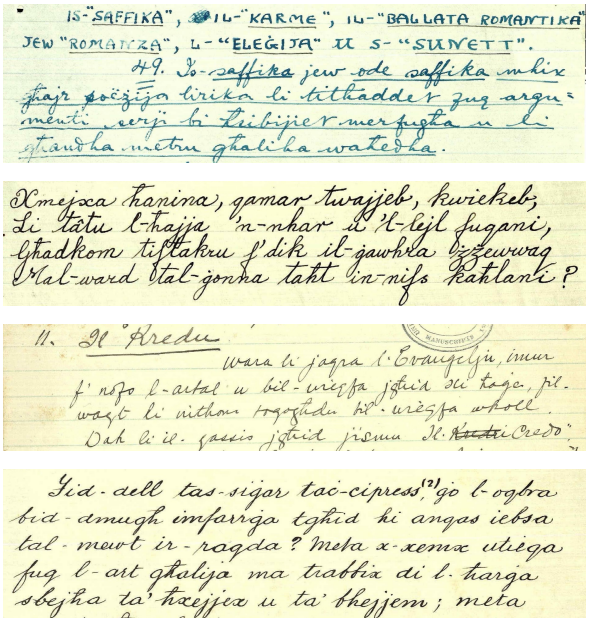


Figure 1: Four samples from different pages of the Miktub dataset illustrating the variety in conditions and handwriting.

3.2. Line-level dataset construction

Most contemporary HTR systems assume line-level inputs. Consequently, each page image and its transcription are converted into aligned line instances. Text splitting is straightforward: each page transcript is programmatically split into individual lines, stored as a single text file per line, grouped by page. Line image segmentation is substantially harder for heterogeneous historical pages. An initial approach was based on Horizontal Profile Projection (HPP) combined with A* pathfinding to refine line boundaries (Shernobyl, 2024; Muthu, 2022). However, this approach proved unsuitable: it frequently merged multiple lines, missed portions of lines, and was slow (minutes per page), which impeded iterative development. Multiple enhancements were explored, including peak clustering, line-height constraints, dilation strength, peak-detection thresholds, and cropping, yet the A* component remained both costly and unreliable, so it was removed.

The segmentation effort then focused on detecting robust line boundaries from HPP peaks. Two clustering strategies were tested (DBSCAN and adaptive gap thresholding), but both still produced closely spaced spurious boundaries and required sensitive tuning across pages. A simpler post-processing rule—peak filtering, i.e., rejecting any peak occurring within a fixed pixel radius of the previous accepted peak—provided a practical compromise: it enforces at most one boundary per line region and generalises across pages without per-page parameter search, at the cost of occasionally keeping the less accurate of two nearby candidate boundaries.

3.3. Semi-automatic segmentation workflow

A key empirical finding is that pages in the Miktub dataset are non-standardised, so parameters that work for one page often fail on others; some pages are not fully segmentable under a purely automatic regime, making full automation infeasible for building a clean training corpus. The adopted solution is a semi-automatic workflow: a fast automatic stage proposes approximate boundaries, followed by a manual adjustment stage that ensures each crop preserves the full vertical context of the handwriting (ascenders/descenders) while minimising overlap with neighbouring lines.

The automatic stage retains HPP but adds (i) optional auto-cropping to remove irrelevant margins (especially for small documents scanned on an A4 canvas) and (ii) the peak-filtering step. Most parameters are fixed globally (e.g., dilation and peak threshold), while cropping is enabled by default and disabled when it would clip text. This streamlined procedure runs in seconds and typically places boundaries slightly below baselines; however, because it effectively "slices" the page rather than explicitly isolating each line, it can cut off descenders and occasionally ascenders. Manual adjustment, therefore, remains necessary for high-quality line images.

Manual refinement is implemented as border editing (top/bottom and, optionally, left/right), rather than editing a single separating line, allowing a single line crop to be expanded without altering adjacent crops. This step is particularly important for interlinear corrections, where an author inserts text between two main lines (sometimes with arrows indicating intended reading order). Depending on length and layout, such corrections are either merged into the corrected line (when short and readable in-place) or treated as standalone lines (when written as a distinct intermediate line). Extreme cases may still force overlap between crops due to tight vertical spacing.

3.4. Transcriptions and dataset variants

During alignment, two transcription issues emerged. First, line-order mismatches occurred because the exported transcriptions supplied by the Data Provider are derived from line-bounding polygons; the textual line order may not correspond to the vertical order of line images. This problem is especially common when corrections are present, because a single correction line can be split into multiple annotated segments. Second, the transcriptions contained systematic inconsistencies and errors, including irregular spacing around punctuation, spacing near clitic articles, accent/diacritic substitutions or omissions, and case errors; additionally, some words were misspelt or omitted.

To handle these issues, we define two dataset variants. Miktub-ORG preserves the original transcriptions "as provided"; consequently, pages with annotation/image mismatches are excluded, reducing coverage. In contrast, Miktub-COR applies a standardisation pass with explicit guidelines for punctuation and stylistic consistency, correcting transcription errors and ensuring one-to-one alignment between each line image and its text. A log of edits is maintained for traceability. Miktub-COR standardisation prioritises consistent formatting to improve model training stability and usability for search and assistive technologies, while largely maintaining a diplomatic stance (transcribing what is written rather than "correcting" author misspellings). An exception is made when marks are clearly displaced (e.g., a delayed cross or dot shifting position), where the intended character may be transcribed to avoid systematic ambiguity.

As a result, Miktub-COR includes all 35 pages

and totals 1076 line images, with an 80/10/10 split into 860/107/109 lines for train/validation/test. Miktub-ORG is smaller: 21 pages, 668 lines total, split into 507/63/64. Agreement between transcription variants, measured at the page level on the full dataset to avoid alignment mismatches, was 3.76% CER and 16.67% WER.

For subsequent paragraph-level experiments, bounding-box metadata for line locations in the original page images is required; because this was not produced during initial dataset creation, it is generated retrospectively via template matching to recover line bounding boxes.

3.5. Out-of-distribution (OOD) test set

To assess generalisation to unseen handwriting, the Data Provider later provided additional pages from the letters collection of P.P. Saydon and Ġużè Aquilina (1932–1946). The two pages were sampled, segmented using the same semi-automatic colour workflow as Miktub-COR, and transcribed from scratch. The resulting Saydon–Ġużè Letters (SGL) dataset contains 55 line images, used exclusively as an OOD test set.

3.6. Experimental protocol and model training

We evaluate two open HTR architectures: HTR-VT and VAN. HTR-VT is used first because its published results on LAM enable replication and because it reports state-of-the-art performance on relevant benchmarks. During training, HTR-VT stores checkpoints for the best validation CER and WER. VAN supports selection by best CER or WER and also retains the latest checkpoint for resumability; following common HTR practice, we prioritise CER when selecting the best model.

Pre-training Before training on Miktub, HTR-VT is trained on the LAM dataset to reproduce reported performance and to obtain a strong initialisation for fine-tuning. Although an online link for pre-trained weights exists, only a READ2016 checkpoint is available. Thus, we pre-train HTR-VT on the LAM dataset and retain the same hyperparameters as in Li et al. (2025). The resulting model (HTR-VT-LAM) is used for further fine-tuning in the next stage.

HTR-VT on Miktub-ORG The initial Miktub experiments use a binary line image variant (Miktub-ORG-BINARY) with 668 lines, training both from scratch and via fine-tuning from HTR-VT-LAM. To support HTR-VT, the dataset is reformatted from a page-folder structure into a flat index of line images and annotations; train/val/test splits are stored as lists of line IDs (80/10/10). The character set size is adjusted for the dataset (92 classes for Miktub-ORG). Training instability near convergence motivates scrutiny of data quality; binary crops are found to be narrow and sometimes poorly binarised, reducing legibility. A colour variant (Miktub-ORG-COLOUR) is therefore created using the semi-automatic segmentation pipeline while retaining the original transcriptions.

Fine-tuning implementation and freezing Fine-tuning is implemented by loading pre-trained weights and freezing a configurable prefix of named layers; the classifier head is re-initialised when class counts differ. An ablation over freezing depth shows that freezing beyond the second embedding layer limits domain adaptation; the best setting freezes up to the second embedding layer and is used thereafter.

Pre-training dataset comparison To test whether LAM's domain similarity matters, a comparative fine-tuning experiment uses READ2016 as an alternative pre-training source. READ2016 is a large historical benchmark (10,527 pages; Early Modern German) with similar page degradations but markedly different script; available weights allow skipping full pre-training, enabling direct comparison under the same fine-tuning pipeline.

HTR-VT and VAN on Miktub-COR We consider two training modes for HTR-VT. The first is training HTR-VT from scratch on the Miktub-COR dataset. The second is to use HTR-VT-LAM (trained on LAM) and fine-tune this model on Miktub-COR. Because Miktub-COR expands coverage (all 35 pages) and modifies transcriptions, it also increases the character set to 97 classes. Empirical comparisons indicate that Miktub-COR is preferable to Miktub-ORG, and subsequent experiments focus on Miktub-COR. The best HTR-VT Miktub-COR model is additionally evaluated on the SGL OOD set.

VAN experiments start with line-level training on Miktub-COR to enable direct comparison with HTR-VT, and because VAN's paragraph-level pipeline typically relies on line-level weights. Dataset conversion is performed from the HTR-VT format into VAN's expected structure (split folders plus a pickle file containing annotations and charset). VAN is trained with default hyperparameters, selecting the best checkpoint based on validation CER.

4. Evaluation

We evaluate the raw outputs of the HTR systems (prior to any post-processing) on (i) the in-distribution test sets of each dataset variant and (ii) an out-of-distribution (OOD) test set written by

Experiment	CER (%)	WER (%)
ORG-BINARY	14.37	34.24
ORG-BINARY+LAM ≤ 3	12.42	29.86
ORG-BINARY+LAM ≤ 2	12.42	29.86
ORG-BINARY READ2016	14.95	37.86
ORG-COLOUR	10.86	28.43
ORG-COLOUR+LAM	10.23	24.12
COR	6.02	15.57
COR+LAM	**5.21**	**13.93**

Table 1: **Test Set Results of HTR-VT experiments** Experiments which freeze N layers during fine-tuning are shown as $\leq N$.

an unseen author. Reported metrics are CER and WER; for post-processing experiments, we also report SER (sequence/line error rate).

4.1. In-distribution HTR results

In-distribution results for all dataset variants using HTR-VT are summarised in Table 1.

Miktub-ORG-BINARY HTR-VT The aim of this initial experiment is to understand the limits of small/noisy data and the impact of freezing embedding layers. This Maltese baseline is trained directly on Miktub-ORG-BINARY and yields 14.37% CER / 34.24% WER, indicating that the small, noisy subset does not, by itself, support low error rates. Several fine-tuning strategies were then tested. Freezing a large portion of the network (up to the third embedding block) was strongly detrimental (26.44% CER / 56.26% WER), suggesting insufficient capacity remained to adapt to the target domain. In contrast, freezing only the first two embedding layers improved performance to 12.42% CER / 29.86% WER, showing that pre-training can help provided most layers remain trainable. Finally, replacing LAM with READ2016 pre-training did not help (14.95% CER / 37.86% WER), reinforcing that transfer effectiveness depends on the relevance of the source domain.

Miktub-ORG-COLOUR HTR-VT To address poor legibility caused by tight cropping and imperfect binarisation, line images were regenerated with the improved semi-automatic segmentation procedure and kept in colour (Miktub-ORG-COLOUR). This change alone reduced the HTR-VT baseline to 10.86% CER / 28.43% WER, a substantial absolute gain over the binary variant. Fine-tuning from the LAM checkpoint produced a small CER improvement but a larger WER improvement, reaching 10.23% CER / 24.12% WER. The evaluation

Experiment	CER (%)	WER (%)
COR	**4.68**	**13.59**

Table 2: **Test Set Results of VAN experiment**.

attributes this pattern to a ceiling effect at the character level once crops become clearer, while pre-training still provides useful priors that reduce word-level confusions.

Miktub-COR HTR-VT The final dataset iteration, Miktub-COR, uses improved line images alongside revised transcriptions, leading to a large jump in baseline accuracy: 6.02% CER / 15.57% WER without external pre-training. Interestingly, applying the earlier fine-tuning approach (with freezing) worsened results (7.10% CER / 18.40% WER), motivating a different strategy. Following findings that Transformer-based HTR benefits from adapting all layers, the model was fine-tuned from LAM with no freezing, yielding the best HTR-VT performance on Miktub-COR: 5.21% CER / 13.93% WER.

Overall, these results underline that (i) transcription quality and consistency can be as impactful as architectural choice, and (ii) for Transformer HTR, partial freezing can hinder domain adaptation when the target distribution differs materially.

Miktub-COR VAN Because Miktub-COR was the most reliable dataset variant, it was the only one used to train VAN. The resulting line-CTC baseline achieved 4.68% CER / 13.59% WER, outperforming the strongest HTR-VT configuration in this work and constituting the best "pure HTR" result reported in (Table 2).

4.2. Out-of-distribution evaluation (SGL)

To probe generalisation, the best HTR-VT models from each dataset variant were evaluated on the SGL OOD test set (two pages from an unseen author). Table 3 summarises how performance drops substantially across the board: 20.37% CER / 53.04% WER for Miktub-ORG-BINARY+LAM, 19.48% / 51.45% for Miktub-ORG-COLOUR+LAM, and 15.93% / 41.01% for the Miktub-COR model fine-tuned on all layers. Notably, the gap between binary and colour variants narrows in OOD, suggesting that once handwriting deviates significantly from training, gains from improved cropping alone are limited, and remaining dataset inaccuracies/inconsistencies become the dominant constraint. Although OOD error rates remain high, the best Miktub-COR model outputs are a useful starting point for assisted manual transcription, enabling a bootstrapping cycle in which new OOD labels can

be incorporated to improve robustness in subsequent iterations.

Experiment	CER (%)	WER (%)
ORG-BINARY+LAM ≤ 2	20.37	53.04
ORG-COLOUR+LAM	19.48	51.45
COR+LAM	15.93	41.01

Table 3: **Test Set Results HTR-VT experiments on the SGL OOD dataset** Experiments which freeze N layers during fine-tuning are shown as $\leq N$.

5. Conclusion

This work targeted accurate recognition of early 20th-century Maltese manuscripts and shows that—given careful data preparation—modern HTR systems can reach low error rates even on challenging historical material. The key driver was an iterative loop that jointly improved (i) line segmentation quality and (ii) transcription accuracy/consistency, complemented by the strategic use of a closely related pre-training dataset.

A central outcome is the construction of a progressively improved line-level dataset. Starting from the initially segmented Miktub-ORG-BINARY, the pipeline was refined to produce a higher-quality colour variant (Miktub-ORG-COLOUR) with better segmentation boundaries, and finally Miktub-COR, which incorporates updated manual transcriptions to improve alignment, accuracy, and consistency. Each iteration improved the legibility and reliability of training data and translated into stronger model performance.

On the modelling side, multiple experiments establish practical guidance for transfer to historical Maltese. For HTR-VT, the best-performing approach is to fine-tune all layers of a model pre-trained on LAM using the corrected Miktub-COR data, rather than relying on partial freezing strategies. In parallel, a single experiment with VAN achieved the best overall results reported in the work, reaching 4.7% CER and 13.6% WER, and thereby outperforming HTR-VT on the in-distribution evaluation and, by extension, on the OOD SGL setting reported in the study.

These findings jointly underscore (a) the value of domain-relevant pre-training and (b) the importance of transcription consistency when training high-capacity recognisers on low-resource historical data.

Despite meeting the primary objectives, limitations remain. The current dataset size (notably the 35-page Miktub-COR collection) constrains generalisation to unseen authors; both HTR-VT and VAN struggle with block letters and sequences of capital letters, and broader coverage (both in- and out-of-distribution) is expected to improve robustness.

Future work should therefore prioritise (i) expanding Miktub-COR with more transcribed pages and more handwriting-varied styles, (ii) incorporating expert-reviewed transcriptions to reduce annotation noise, and (iii) exploring larger or more recent pre-trained architectures (e.g., TrOCR) while recognising that data scale and quality are likely to remain the dominant bottlenecks. Post-processing remains open: one promising direction is fine-tuning language models for Maltese HTR correction or integrating them more tightly into decoding.

Overall, the project demonstrates that, even under limited resources, substantial accuracy is attainable for historical Maltese HTR, and that domain-specific datasets plus iterative refinement are foundational for progress in low-resource handwritten document understanding.

6. Acknowledgements

We acknowledge the collaboration and support of the University of Malta Library Special Collections, in particular Antida Mizza and Luke Joseph Brincat, for supplying the manuscripts, transcripts, annotations and feedback on this work.

7. Bibliographical References

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Théodore Bluche, Jérôome Louradour, and Ronaldo Messina. 2017. Scan, attend and read: End-to-end handwritten paragraph recognition with mdlstm attention. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 01, pages 1050–1055.

Théodore Bluche and Ronaldo Messina. 2017. Gated convolutional recurrent neural networks for multilingual handwriting recognition. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 01, pages 646–651.

Claudia Borg and Albert Gatt. 2017. Morphological Analysis for the Maltese Language: The challenges of a hybrid system. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 25–34, Valencia, Spain. Association for Computational Linguistics.

Silvia Cascianelli, Vittorio Pippi, Maarand Martin, Marcella Cornia, Lorenzo Baraldi, Kermorvant Christopher, and Rita Cucchiara. 2022. The lam dataset: A novel benchmark for line-level handwritten text recognition. In *International Conference on Pattern Recognition*.

Stanley F Chen and Joshua Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–394.

Denis Coquenet, Clement Chatelain, and Thierry Paquet. 2023. End-to-End Handwritten Paragraph Text Recognition Using a Vertical Attention Network . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 45(01):508–524.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 369–376, New York, NY, USA. Association for Computing Machinery.

Alex Graves and Jürgen Schmidhuber. 2008. Offline handwriting recognition with multidimensional recurrent neural networks. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, NIPS'08, page 545–552, Red Hook, NY, USA. Curran Associates Inc.

Emmanuele Grosicki and Haikal El-Abed. 2011. Icdar 2011 - french handwriting recognition competition. In *2011 International Conference on Document Analysis and Recognition*, pages 1459–1463.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2023. Trocr: Transformer-based optical character recognition with pre-trained models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):13094–13102.

Yuting Li, Dexiong Chen, Tinglong Tang, and Xi Shen. 2025. Htr-vt: Handwritten text recognition with vision transformer. *Pattern Recognition*, 158:110967.

U.-V. Marti and H. Bunke. 2002. The iam-database: an english sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, 5(1):39–46.

S. P. Muthu. 2022. Line segmentation in handwritten documents. https://github.com/muthuspark/line-segmentation-handwritten-doc. Last accessed: April 2026.

Arthur Flor de Sousa Neto, Byron Leite Dantas Bezerra, and Alejandro Héctor Toselli. 2020. Towards the natural language processing as spelling correction for offline handwritten text recognition systems. *Applied Sciences*, 10(21):7711.

Daniel Parres and Roberto Paredes. 2023. Finetuning vision encoder–decoder transformers for handwriting text recognition on historical documents. In *Document Analysis and Recognition - ICDAR 2023*, pages 253–268, Cham. Springer Nature Switzerland.

R. Plamondon and C.M. Privitera. 1999. The segmentation of cursive handwriting: an approach based on off-line recovery of the motor-temporal information. *IEEE Transactions on Image Processing*, 8(1):80–91.

R. Plamondon and S.N. Srihari. 2000. Online and off-line handwriting recognition: a comprehensive survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):63–84.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Nagendra Goel, Mirko Hannemann, Yanmin Qian, Petr Schwarz, and Georg Stemmer. 2011. The kaldi speech recognition toolkit, workshop on automatic speech recognition and understanding. *US IEEE Signal Processing Society, Hilton Waikoloa Village, Big Island, Hawaii*.

Joan Puigcerver. 2017. Are multidimensional recurrent layers really necessary for handwritten text recognition? In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 01, pages 67–72.

Shernobyl. 2024. Writer identification system. https://github.com/Shernobyl/Writer-Identification-System. Last accessed: April 2026.

Joan Andreu Sánchez, Verónica Romero, Alejandro H. Toselli, and Enrique Vidal. 2016. Icfhr2016 competition on handwritten text recognition on the read dataset. In *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 630–635.

Joan Andreu Sánchez, Alejandro H. Toselli, Verónica Romero, and Enrique Vidal. 2015. Icdar 2015 competition htrts: Handwritten text recognition on the transcriptorium dataset. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1166–1170.

Long Wu, Ta Li, Li Wang, and Yonghong Yan. 2019. Improving hybrid ctc/attention architecture with time-restricted self-attention ctc for end-to-end speech recognition. *Applied Sciences*, 9(21):4639.

Smelling the Past: Investigating Historical Models for Olfactory Event Extraction

Teresa Paccosi Marijn Koolen
KNAW Humanities Cluster, DHLab, Amsterdam
teresa.paccosi@dh.huc.knaw.nl, marijn.koolen@di.huc.knaw.nl

Abstract

In this paper, we present a series of experiments using historical language models to investigate the impact of pretraining on data that more closely resembles the task domain, focusing on the case study of automatic olfactory event extraction. We tested five monolingual historical and contemporary pretrained models on the task of extracting olfactory events using a benchmark spanning several centuries. The languages investigated are English, French, Italian, German, and Dutch. The aim of our research is not only to assess whether historical models can improve performance on this diachronically oriented task, but also to gain deeper insight into the factors influencing model performance through a detailed analysis of performance patterns. We examine potential sources of variation and previously proposed hypotheses to account for lower performance observed in this task, thereby offering a more comprehensive understanding of model behavior in this context.

Keywords: historical smell, historical models, model evaluation

1. Introduction

Within the traditional hierarchy of the five senses, olfaction has long been relegated to a marginal role, particularly in Western societies where vision has held a dominant position (Winter et al., 2018). This neglect, coupled with the widespread assumption that smell lacks a well-defined semantic field (Sperber, 1975), has contributed to the limited number of studies devoted to this sensory domain. Only in recent years olfaction has begun to attract renewed scholarly attention (Howes, 2006), and its cultural significance has started to be acknowledged, including its value as a historical example of intangible cultural heritage (Tullett et al., 2022). In the fields of Natural Language Processing (NLP) and Digital Humanities (DH), engagement with the domain of olfaction is likewise relatively recent. Precisely because smell has received comparatively little systematic attention, it constitutes a promising area for computational investigation. Nevertheless, the structured analysis of olfactory terminology and the automatic extraction of smell-related information remain underexplored within NLP research, especially from a diachronic point of view. This paper presents an experimental investigation of language models pre-trained on historical texts, applied to the specific case study of olfactory language. Previous work on this case study had relied solely on models trained on contemporary corpora. Using a similar setting to the one presented in Menini (2024), we compare models pre-trained on contemporary texts against those pre-trained on historical texts. We combine multiple evaluation perspectives aiming at assessing the impact of using historical models for this task, through (a) standard metrics such as precision, recall, and F_1 score, (b) a comparison of strict versus lenient F_1 scores to account for span boundary choices (previously identified as major source of problems for some entity types), and (c) detailed error analysis including overlap assessment and manual inspection. By situating our work within the broader line of research on historical model performance (Plank, 2016; Ehrmann et al., 2023), especially the influence of pre-training data similarity to target tasks (Gururangan et al., 2020; Gladstone et al., 2025), we focus on a specific case study that has not previously been explored from this perspective.

Our contributions are twofold: (a) the release of fine-tuned historical models for extracting olfactory events, and (b) a systematic analysis highlighting concrete factors that can lead to performance improvements.[1]

2. Related Work

Most research on sensory language has primarily focused on developing structured resources to model perceptual domains. In this vein, Tekiroğlu et al. (2014) introduce *Sensicon*, a large-scale sensory lexicon that automatically associates English words with the five senses, while other studies target domain-specific descriptive vocabularies, such as those used by whiskey and wine reviewers, as in Hamilton et al. (2023) and Lefever et al. (2018). Focusing more specifically on smell-related events, Brate et al. (2020) introduced an annotation framework and semi-supervised methods for capturing olfactory experiences, while McGregor and McGillivray (2018) employed distributional

[1] Code and data are available at this GitHub repository.

Proceedings of LT4HALA 2026 @ LREC 2026, pages 389–399.
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

semantics to isolate smell descriptions. More recently, Tonelli and Menini (2021) built a FrameNet-inspired annotation scheme and a benchmark for olfactory events in historical texts (Menini et al., 2022), which Menini (2024) employed to conduct experiments on the same task we are investigating in this paper. The author used BERT models pretrained on contemporary language, which have also previously shown strong performance on the extraction of sensory-related information (Stojanov et al., 2021). However, as computational methods are increasingly applied in historical research, it becomes clear that the use of appropriately designed systems and models is essential. Simply increasing the size of a language model or modifying its architecture does not guarantee improved performance (Alajrami and Aletras, 2022; Sun et al., 2022). Instead, the effectiveness depends heavily on the alignment of data distributions between pretraining and fine-tuning phases (Li et al., 2022). Supporting this point, Verkijk et al. (2025) show that a model pretrained exclusively on in-domain historical data outperforms models trained on much larger amounts of data of a more diverse nature.

3. Extracting Olfactory Events

In this section, we outline the theoretical background of our study, focusing on the principles of Frame Semantics on which the olfactory event extraction task is based. We briefly discuss how Frame Semantics is applied to model olfactory events in text, and then introduce the models used in this paper along with the details of our experimental setup.

3.1. Theoretical Background

Our work builds on the olfactory benchmark presented in Menini et al. (2022)[2] and the work of Menini (2024), who finetuned various BERT models on this benchmark for the task of extracting smell-related concepts, specifically targeting the olfactory frame. The original benchmark covers six languages, English, French, Italian, German, Dutch, and Slovenian. We excluded Slovenian, as we lack the necessary linguistic expertise to perform a reliable error analysis for this language, and, to our knowledge, no monolingual historical model is currently available for it. The benchmark is based on Frame Semantics theory (Fillmore et al., 2006), modeling olfactory events as a distinct frame with its Lexical Units (LUs) and Frame Elements (FEs). LUs are words that evoke the olfactory frame, defined as "pairing[s] of a word with a meaning" (Ruppenhofer et al., 2016), and in this case include terms describing odours, such as *smell*, *perfume*, or *stench* (referred to here as *Smell_Word*). Each frame also defines a set of FEs, representing the semantic roles or participants associated with the LU. For example, in "The cook carefully smelled the soup", *smell* is the LU, while *the cook* and *the soup* serve as the *Perceiver* and *Smell_Source* FEs, respectively. In Table **??**, we report a list of all the FEs in the benchmark, along with brief descriptions, which constitute the set of tags targeted in the olfactory frame extraction task.

3.2. Models

Since our experiments focus mainly on historical language models, and to our knowledge, the only existing multilingual historical model, hmBERT (Schweter et al., 2022), covers German, English, French, Swedish, and Finnish, but not Italian or Dutch, we opted for a strictly monolingual setting for each language. In this setup, both the training data and the models are language-specific, and we include contemporary models for comparison alongside the historical ones.

The language-specific historical models we use for the olfactory extraction task are the following:

- **English**: MacBERTh[3] , (Manjavacas and Fonteyn, 2021)
- **Italian**: BERToldo[4] (Aprosio et al., 2022),
- **French**: D'AlemBERT[5] (Gabay et al., 2022),
- **Dutch**: GysBERT[6] (Manjavacas Arevalo and Fonteyn, 2022),
- **German**: Historical German BERT[7]

The language-specific contemporary models we used are instead:

- **English**: bert-base-cased[8] (Devlin et al., 2019),
- **Italian**: bert-base-italian-cased[9] (Schweter, 2020),

[2]Detailed information about the benchmark is provided in Menini et al. (2022) and the corresponding github repository: https://github.com/Odeuropa/benchmarks_and_corpora.

[3]https://huggingface.co/emanjavacas/MacBERTh

[4]https://github.com/dhfbk/historical-bert?tab=readme-ov-file

[5]https://huggingface.co/pjox/dalembert

[6]https://huggingface.co/emanjavacas/GysBERT

[7]https://huggingface.co/redewiedergabe/bert-base-historical-german-rw-cased

[8]https://huggingface.co/google-bert/bert-base-cased

[9]https://huggingface.co/dbmdz/bert-base-italian-cased

Frame Element	Example Sentence
Smell Source	The person, object or place that has a specific smell. *The odour [of tar] and [pitch] was so strong.*
Odour Carrier	The carrier of an odour, either an object (e.g. handkerchief) or atmospheric elements (wind, air) *The unpleasant smell [of the vapour] of linseed oil extended for a considerable distance.*
Quality	A quality associated with a smell and used to describe it. *Earth has a [strong], [aromatic] odour.*
Perceiver	The being that perceives an odour, who has a perceptual experience, not necessarily on purpose. *The scent is described by [Dr. Muller] as delicious.*
Evoked Odorant	The object, place or similar that is evoked by the odour, even if it is not in the scene. *In offensive perspiration of the feet [a peculiar cabbage-like] stench is given off.*
Location	The location where the smell event takes place. *And, particularly, [at the foot of the garden], where he felt so very offensive a smell that has sickened him.*
Time	An expression describing when the smelling event occurred. *Galeopsis smells fetid [at first handling], [afterwards] aromatic.*
Circumstances	The state of the world under which the smell event takes place. *[When stale] the lobster has a rank stench.*
Effect	An effect or reaction caused by the smell. *An ill smell [gives a nauseousness].*

Table 1: Overview of the Frame Elements (FEs) as reported in (Menini et al., 2022). Lexical Units (i.e., *Smell_Words* are underlined and the FE of interest is in square brackets. The same definitions hold for all languages included in the benchmark.

- **Dutch**: BERTje[10] (De Vries et al., 2019),
- **French**: FlauBERT[11] (Le et al., 2020),
- **German**: bert-base-german-cased[12] (Chan et al., 2020).

For the contemporary setting, we re-ran all experiments using the same models investigated in Menini (2024) to create a fully controlled experimental setup with consistent test data and predictions across languages. Beyond simply reproducing previous results, this setup enabled us to explore more deeply the factors underlying performance differences. Specifically, we examined how predictions overlap with gold annotations, analysed the nature of errors made by different models in span selection, and compared the behaviour of their tokenisers, with the goal of gaining a deeper understanding of how historical data influences model performance.

3.3. Multitask Approach for Olfactory Frame Extraction

Given that previous results on this task showed that the multitask setting significantly outperforms the single-task setting, we decided to focus our analysis on the former. In this setting, we employ multitask learning (Caruana, 1997), an approach in which the model learns multiple tasks simultaneously while sharing a common representation, enabling knowledge transfer across tasks. We model the classification of the LU and each FE class as separate tasks, following previous work on this task that suggested that detecting LUs, typically expressed as single tokens, can serve as an auxiliary task, supporting the identification of more complex FEs through shared representations. For this setup, we use MaChAmp (Van Der Goot et al., 2021), a toolkit for fine-tuning contextualised embeddings in multitask NLP scenarios. MaChAmp initialises from a pre-trained encoder and fine-tunes it using an inverse square root learning rate schedule with linear warm-up (Howard and Ruder, 2018). Each task is assigned a dedicated decoder, and prediction is performed using Conditional Random Fields (Lafferty et al., 2001), which are well suited for sequence labeling. We configure MaChAmp to run 10 BIO-based sequence labeling tasks, each corresponding to a LU or FE. Evaluation is conducted using span-level strict and lenient precision (P), recall (R), and F_1. We experiment with the same setting in all languages using uniform weighting at 1 for all tasks, a learning rate of $1e-4$, a batch size of 32, and 30 epochs.

4. Results

In Table 3, we show the results of the experiments for both contemporary and historical models, in terms of strict and lenient precision (P), recall (R) and F_1. From a general perspective, it is noticeable that using historical models has a positive impact on F_1 scores for almost all labels in En-

[10] https://huggingface.co/GroNLP/bert-base-dutch-cased

[11] https://huggingface.co/flaubert/flaubert_base_cased

[12] https://huggingface.co/google-bert/bert-base-german-cased

glish and Dutch, and to a lesser extent in German, whereas for Italian and French contemporary models consistently perform better. A more detailed discussion of these findings is provided in Section 5.1. Examining the results more closely, we observe some cross-linguistic and cross-label patterns. Although the majority of labels do not achieve particularly high scores with either contemporary or historical models, certain labels consistently perform well: notably, *Smell_Word*, and *Quality*, and, compared to the other labels, also *Smell_Source*, and *Evoked_Odorant*. The strong performance of *Smell_Word* is not surprising, as it often corresponds to lexically consistent single words, making its identification relatively straightforward. The other labels, however, probably benefit from regular syntactic realisations: *Quality* frequently appears as an adjective preceding the *Smell_Word*, *Smell_Source* often functions as a direct object, and *Evoked_Odorant* typically occurs in predictable prepositional constructions such as "like" or "as". Still, both *Smell_Source* and *Evoked_Odorant* present a variability of expression that likely do not help model generalisation. More complex or longer entities, such as *Circumstances*, and *Effect*, consistently show low performances and a lower recall, reflecting the difficulty of correctly identifying multi-token spans with variable realisation. Comparing strict and lenient evaluations, we observe notable differences for labels with longer spans, especially in English, German and Dutch. For example, in *Circumstances*, where recall shows notable improvement in languages such as English, German, and Italian, supporting the idea that incorrect span boundaries are one of the main sources of errors for this entity type. We provide a more detailed discussion of the differences between strict and lenient evaluation results in Section 5.3.

Model	tokens	data	coverage
BERToldo$_all$ (it)	∼10M	970 MB	1200-1900
D'AlemBERT (fr)	∼185M	–	1500-1790
MacBERTh (en)	∼3.9B	–	1450-1950
GysBERT (nl)	∼7.1B	–	1618-1999
Hist BERT (de)	–	–	1840-1920

Table 2: Size and temporal coverage of pretraining data for each historical model.

5. Discussion and Analysis

In this section, we provide a more detailed discussion of the results presented in Section 4. This includes an analysis of the difference in performance under lenient versus strict evaluation, a comparison of Subword Fertility Rates across models, an examination of the gain or loss in F1 when comparing historical and contemporary models, previously suggested as a possible factor influencing model performance (Rust et al., 2021), and an analysis of the most common errors made by the models in terms of span selection.

5.1. Contemporary Versus Historical Models

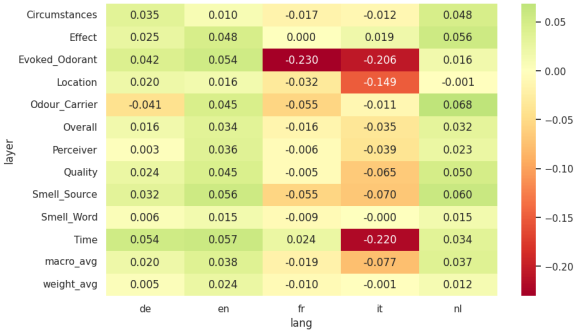


Figure 1: F1 differences (historical − contemporary) across layers and languages.

Comparing the potential gain from using historical models for this task, a clear pattern emerges: Italian and French achieve better performance with contemporary models, whereas English and Dutch, and, to a lesser extent, German, benefit more from the historical setting. These differences may be partly related to the size and temporal coverage of the pretraining data for each historical model, as summarised in Table 4. French and Italian historical models appear to perform worse, likely due to the comparatively limited historical data available (approximately 185M and 10M tokens, respectively), which may constrain their ability to capture complex or less frequent historical patterns. Moreover, the French historical model was primarily pretrained on 17th-century texts, whereas the majority of texts in the French benchmark come from the late 18th century. This suggests that performance may not only be influenced by the size of the pretraining data but also by the similarity of the pretraining domain, a factor previously observed in Dutch by Verkijk et al. (2025) as potentially impacting model effectiveness.

In contrast, the English and Dutch historical models achieve comparatively stronger results, plausibly benefiting from larger pretraining corpora spanning a broader temporal range, which may enable them to model both frequent and rare phenomena more effectively. German shows only a marginal improvement: its historical pretraining corpus spans 1840–1920, which is only slightly earlier than the texts used for the contemporary model. As a result, the model gains some benefit from exposure to older linguistic patterns, but the over-

		Strict						Lenient					
		Contemporary			Historical			Contemporary			Historical		
Entity type	Lang.	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Circumstances	de	0.240	0.112	0.142	**0.262**	**0.143**	**0.177**	0.532	0.243	0.312	**0.540**	**0.301**	**0.372**
	en	**0.363**	0.283	0.311	0.328	**0.318**	**0.321**	0.554	0.449	0.486	**0.572**	**0.562**	**0.564**
	fr	**0.040**	**0.057**	0.041	0.028	0.017	0.024	**0.494**	**0.428**	**0.368**	0.156	0.057	0.093
	it	0.203	**0.178**	**0.177**	**0.278**	0.123	0.164	0.505	**0.495**	**0.470**	**0.534**	0.222	0.297
	nl	0.196	0.076	0.107	**0.223**	**0.119**	**0.155**	0.314	0.122	0.173	**0.343**	**0.185**	**0.240**
Effect	de	0.207	0.066	0.100	**0.241**	**0.088**	**0.125**	**0.472**	0.151	0.228	0.452	**0.165**	**0.235**
	en	0.160	0.153	0.155	**0.212**	**0.207**	**0.204**	0.358	0.342	0.348	**0.382**	**0.345**	**0.352**
	fr	0.000	0.000	0.000	0.000	0.000	0.000	**0.250**	**0.200**	**0.250**	0.000	0.000	0.000
	it	0.075	0.040	0.060	**0.133**	**0.045**	**0.079**	**0.375**	**0.135**	**0.222**	0.287	0.089	0.159
	nl	0.221	0.097	0.122	**0.241**	**0.145**	**0.178**	**0.473**	0.209	0.272	0.468	**0.294**	**0.356**
Evoked Odorant	de	0.350	0.297	0.317	**0.401**	**0.330**	**0.359**	**0.706**	**0.607**	**0.644**	0.670	0.546	0.596
	en	0.326	0.343	0.312	**0.374**	**0.385**	**0.367**	0.594	**0.608**	0.574	**0.601**	0.596	**0.577**
	fr	**0.423**	**0.381**	**0.397**	0.375	0.086	0.167	**0.451**	**0.410**	**0.426**	0.375	0.086	0.167
	it	**0.356**	**0.353**	**0.329**	0.124	0.125	0.123	**0.501**	**0.552**	**0.483**	0.328	0.302	0.300
	nl	0.362	0.281	0.312	**0.337**	**0.323**	**0.328**	0.561	0.440	0.487	**0.562**	**0.540**	**0.548**
Location	de	0.228	0.089	0.126	**0.253**	**0.106**	**0.146**	**0.330**	0.132	0.185	0.323	**0.140**	**0.192**
	en	0.424	0.424	0.423	**0.432**	**0.462**	**0.439**	**0.569**	0.568	**0.567**	0.542	**0.588**	0.555
	fr	0.206	**0.166**	**0.174**	**0.225**	0.106	0.143	0.436	**0.317**	**0.338**	**0.504**	0.238	0.321
	it	**0.312**	**0.178**	**0.219**	0.197	0.050	0.070	0.498	**0.300**	**0.365**	**0.500**	0.176	0.222
	nl	**0.350**	0.156	**0.214**	0.273	**0.177**	0.213	**0.517**	0.226	0.312	**0.436**	**0.285**	**0.343**
Odour Carrier	de	**0.313**	**0.130**	**0.160**	0.202	0.093	0.118	**0.313**	**0.130**	**0.160**	0.309	0.113	0.152
	en	0.355	0.320	0.329	**0.435**	**0.352**	**0.375**	0.468	0.423	0.435	**0.550**	**0.448**	**0.474**
	fr	0.412	**0.159**	**0.253**	**0.452**	0.130	0.197	**0.457**	**0.174**	**0.278**	0.452	0.130	0.197
	it	0.334	**0.151**	**0.194**	**0.374**	0.125	0.183	0.470	**0.215**	**0.273**	**0.512**	0.162	0.241
	nl	0.247	0.106	0.146	**0.319**	**0.163**	**0.215**	0.303	0.127	0.176	**0.381**	**0.193**	**0.254**
Perceiver	de	**0.393**	**0.236**	0.286	0.378	**0.236**	**0.289**	**0.436**	0.264	0.319	**0.436**	**0.273**	**0.334**
	en	0.395	0.438	0.412	**0.431**	**0.468**	**0.448**	**0.531**	**0.581**	**0.551**	0.529	0.576	0.550
	fr	0.157	**0.058**	**0.084**	**0.188**	0.050	0.078	**0.417**	**0.146**	**0.215**	0.216	0.060	0.092
	it	0.297	**0.204**	**0.231**	**0.427**	0.150	0.192	0.433	**0.301**	**0.338**	**0.474**	0.183	0.224
	nl	**0.377**	0.210	0.268	0.358	**0.248**	**0.291**	**0.426**	0.237	0.303	0.421	**0.291**	**0.342**
Quality	de	0.553	0.565	0.556	**0.596**	**0.569**	**0.579**	0.758	**0.773**	0.761	**0.794**	0.757	**0.771**
	en	0.674	0.696	0.684	**0.716**	**0.742**	**0.729**	0.824	0.852	0.837	**0.840**	**0.871**	**0.855**
	fr	0.390	**0.451**	**0.415**	**0.419**	0.403	0.410	**0.640**	**0.740**	**0.682**	0.601	0.578	0.589
	it	**0.732**	**0.739**	**0.734**	0.682	0.661	0.670	**0.807**	**0.815**	**0.810**	0.784	0.759	0.769
	nl	0.633	0.599	0.614	**0.671**	**0.661**	**0.664**	0.741	0.700	0.718	**0.757**	**0.749**	**0.751**
Smell Source	de	0.395	0.312	0.346	**0.415**	**0.352**	**0.378**	0.587	0.462	0.514	**0.597**	**0.508**	**0.545**
	en	0.475	0.456	0.465	**0.505**	**0.541**	**0.520**	**0.636**	0.610	0.621	0.631	**0.677**	**0.651**
	fr	**0.359**	**0.336**	**0.329**	0.300	0.262	0.274	**0.581**	**0.557**	**0.534**	0.446	0.392	0.408
	it	**0.400**	**0.418**	**0.405**	**0.400**	0.289	0.335	0.562	**0.588**	**0.569**	**0.588**	0.426	0.493
	nl	0.351	0.283	0.312	**0.401**	**0.348**	**0.372**	0.549	0.442	0.487	**0.571**	**0.497**	**0.530**
Smell Word	de	0.801	**0.808**	0.800	**0.816**	0.807	**0.806**	0.815	**0.822**	0.814	**0.831**	**0.822**	**0.821**
	en	0.889	0.855	0.871	**0.900**	**0.873**	**0.886**	0.898	0.864	0.880	**0.911**	**0.884**	**0.897**
	fr	**0.796**	0.817	**0.801**	0.780	**0.819**	0.792	**0.845**	**0.870**	**0.852**	0.802	0.841	0.814
	it	0.856	**0.905**	**0.877**	**0.897**	0.859	0.876	0.870	**0.920**	**0.891**	**0.910**	0.872	0.889
	nl	0.740	0.784	0.761	**0.746**	**0.808**	**0.776**	0.754	0.799	0.775	**0.763**	**0.826**	**0.793**
Time	de	0.234	0.087	0.122	**0.376**	**0.137**	**0.176**	0.420	0.172	0.232	**0.524**	**0.229**	**0.286**
	en	0.471	0.387	0.416	**0.521**	**0.441**	**0.473**	0.624	0.507	0.549	**0.633**	**0.531**	**0.573**
	fr	0.244	**0.086**	0.120	**0.452**	0.077	**0.144**	**0.646**	**0.286**	**0.347**	0.524	0.121	0.207
	it	**0.490**	**0.425**	**0.453**	0.262	0.213	0.233	**0.768**	**0.654**	**0.703**	0.700	0.581	0.629
	nl	**0.296**	0.140	0.181	0.291	**0.182**	**0.216**	**0.445**	0.240	**0.297**	0.346	0.227	0.265
macro avg.	de	0.371	0.270	0.295	**0.394**	**0.286**	**0.316**	0.537	0.376	0.417	**0.548**	**0.385**	**0.430**
	en	0.453	0.435	0.438	**0.485**	**0.479**	**0.476**	0.606	0.580	0.585	**0.619**	**0.608**	**0.605**
	fr	0.302	**0.251**	**0.263**	**0.344**	0.195	0.244	**0.529**	**0.413**	**0.432**	0.437	0.250	0.316
	it	**0.412**	**0.359**	**0.375**	0.384	0.264	0.297	**0.583**	**0.497**	**0.518**	0.570	0.377	0.428
	nl	0.377	0.273	0.304	**0.386**	**0.317**	**0.341**	**0.508**	0.354	0.400	0.505	**0.409**	**0.442**
weighted avg.	de	0.633	0.611	0.615	**0.644**	**0.608**	**0.620**	0.719	**0.674**	0.686	**0.731**	0.673	**0.693**
	en	0.681	0.667	0.672	**0.697**	**0.699**	**0.696**	0.756	0.742	0.746	**0.762**	**0.770**	**0.763**
	fr	0.606	**0.629**	**0.608**	**0.608**	0.609	0.598	**0.703**	**0.743**	**0.705**	0.653	0.637	0.634
	it	0.708	**0.735**	**0.717**	**0.747**	0.697	0.717	0.754	**0.778**	**0.759**	**0.790**	0.720	0.746
	nl	0.601	0.590	0.591	**0.604**	**0.608**	**0.603**	0.661	0.623	0.635	**0.664**	**0.651**	**0.653**

Table 3: Evaluation scores per entity type and overall, for contemporary and historical models for all five languages, using both strict and lenient interpretations.

all similarity to contemporary language limits the potential advantage of historical specialisation.

Overall, these observations suggest that the quantity and relevance of training data may play a key role in affecting the model performance on this particular task.

5.2. Subword Fertility Rate

One reason for creating historical models from scratch is that the tokeniser is also trained from scratch, on the same historical texts as the pre-training data. The underlying idea is that the resulting vocabulary may better capture the spelling and

Model	tokens	data	coverage
BERToldo$_all$ (it)	∼10M	970 MB	1200-1900
D'AlemBERT (fr)	∼185M	–	1500-1790
MacBERTh (en)	∼3.9B	–	1450-1950
GysBERT (nl)	∼7.1B	–	1618-1999
Hist BERT (de)	–	–	1840-1920

Table 4: Size and temporal coverage of pretraining data for each historical model.

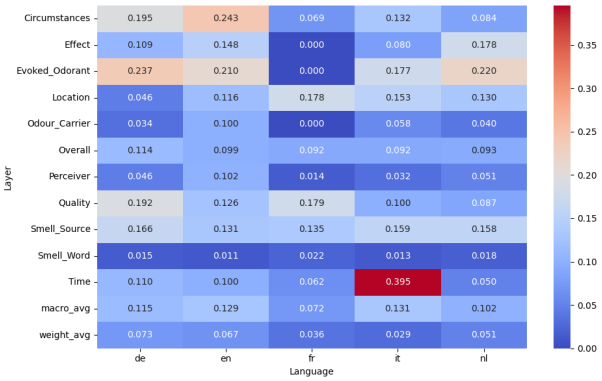
word usage of the historical period. If the spelling of many words in a language has changed strongly between the period of the task data (e.g. the 17th century) and the period of the pretraining data of contemporary models (e.g. 20th and 21st century), then the contemporary tokeniser may split many tokens in historical texts into multiple subtokens, while a tokeniser trained on the same historical texts would use fewer subtokens and for most common words would need only a single subtoken. To test whether the difference in performance could be related to the difference in tokenisers, we look at the Subword Fertility Rate (SFR) (Rust et al., 2021). This is the mean number of subtokens per token for a representative samples of texts. A higher SFR may correspond to a decrease in performance, although Ali et al. (2024) found that this not a reliable measure, at least not for English. We compute the SFR for the tokenisers of contemporary and historical models of each of the five languages using the ground truth datasets, with SFR scores for all tokens in the dataset, and for tokens in the spans of each entity type.

For overall SFR, there is little difference between historical and contemporary models for English (SFR of 1.13 vs. 1.23 respectively), French (1.27 vs. 1.21) and Italian (1.15 vs. 1.25). The differences are bigger for Dutch (1.23 vs. 1.48). Interestingly, for German, the historical tokeniser has a substantially higher SFR than the contemporary model (1.84 vs. 1.29). Based on these numbers, there seems to be no clear relation between differences in SFR and differences in performance. This is further demonstrated when we zoom in on specific entity types. Most models perform better on *Smell_Words* than on other entities, partly because *Smell_Words* are among the most frequent entity types (no Frame Element is annotated in their absence), and partly because they are typically single words, whereas most other entity types span longer sequences, as discussed in Section 4. But for most languages the tokens of *Smell_Words* have the highest SFR of all entity types, and for entity types with the lowest SFR, such as *Time* and *Circumstances* the performance is much lower. When we compare contemporary and historical models on *Smell_Words*, we find that differences in performance in terms of F_1 are minimal (see Table 3), while the SFR differences are big (and in different directions) for English (2.14 for contemporary vs. 1.54 for historical), German (2.18 vs. 2.63) and Italian (1.89 vs. 1.55). Only for Dutch (2.01 vs. 1.94) and French (1.58 vs 1.70) are the SFR differences small. In other words, there is no direct relationship between SFR and performance, although it may be confounded with other factors, such as the amount of orthographic shift in spelling between historical and contemporary texts in a language, the amount of pretraining data that a base model has seen and the amount of task-specific training data per language and entity type. Of course, SFR is not the only aspect that matters in comparing tokenisers of contemporary and historic models, but we include this analysis because SFR can be analysed without recreating the models from scratch.

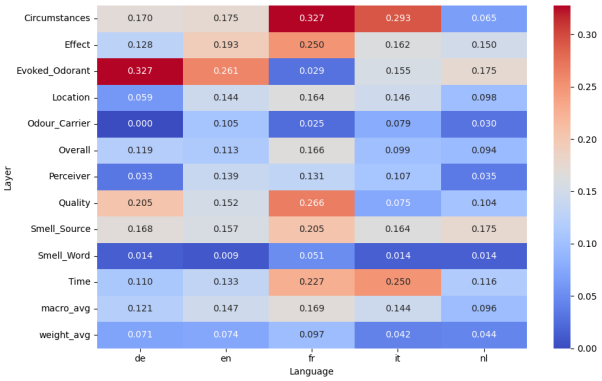
5.3. Exact Versus lenient

In the work of Menini (2024), it was hypothesised that the low performance on certain labels could be partly due to issues with span boundary selection. To test this intuition, we compute lenient F_1 scores, where a predicted tag is considered correct if it partially overlaps with the ground truth tag and has the same entity type. This is useful because, with exact F_1, we risk discarding cases where the assigned tag is actually correct, but one or more words in the expression were not perfectly aligned between the gold and predicted annotations. This issue is particularly relevant for labels that often include entire phrases or sentences, such as *Circumstances*, *Evoked_Odorant* and *Effect*, but it can also provide insights for other labels.

The results reported in Table 3 were compared in terms of loss and gain on exact F_1 in Fig. 2. This comparison clearly shows that while using lenient evaluation, gains are particularly pronounced in the contemporary case. In the case of the contemporary models, a possible correlation emerges between the span length of a label and the gain observed under the lenient evaluation. In particular, *Circumstances*, *Effect* and *Evoked_Odorant*, which tend to correspond to longer spans, show the largest improvements in almost all languages. By contrast, in the historical models, gains are minimal for almost all labels across all languages, with the notable exception of *Time* in Italian. The marked improvement for *Time* is also visible in the contemporary setting. This pattern may be explained by the fact that Italian is the language with the highest average number of tokens per span for this label, followed by French, but further investigation is needed to confirm this hypothesis. These findings support the previously proposed intuition regarding the low performance observed for specific labels. At the same time, the improvements



(a) Lenient F1 gain for historical models



(b) Lenient F1 gain for contemporary model

Figure 2: Comparison of exact and lenient F1 scores across languages and labels for historical (a) and contemporary (b) models.

remain relatively limited for the vast majority of labels, indicating that the primary sources of error are likely to lie elsewhere.

5.4. Span Analysis

Based on random samples of 50-70 partial matches per language, we find that longer ground truth spans and predicted spans often differ by a one or two words, where the added or missing words are arguably not necessary for a meaningful tag. In the following examples we report some recurring span boundary mismatches between the ground truth (GT) annotations and the predicted spans (PRED) in English.[13]

Separate vs. merged span instances:

- *Smell_Word* [nl]
  - GT: “reuk en geur” (smell and scent) as single entity
  - PRED: “reuk” (smell) and "geur" (scent) as two entities

Inclusion or omission of prepositions:

- *Smell_Source* [nl]
  - GT: “hare bloemen” (her flowers)
  - PRED: “bloemen” (flowers)
- *Smell_Source* [en]
  - GT: “fine incense”
  - PRED: “with fine incense”

Unclear span boundaries in longer expressions:

- *Effect* [it]
  - GT: “che serva per generare il vento” (which serves to generate the wind)
  - PRED: “per generare il vento” (to generate the wind)
- *Circumstances* [de]
  - GT: “besonder, wenn sie getrocknet” (particularly, when they are dried)
  - PRED: “wenn sie getrocknet” (when they are dried)

Partial specifications or reduced mention:

- *Location* [fr]
  - GT: “dans tout l’Orient” (in all of the Orient)
  - PRED: “l’Orient” (the Orient).
- *Evoked_Odorant* [it]
  - GT: “leggermente aromatico” (slightly aromatic)
  - PRED: “aromatico” (aromatic)

We argue that the words added in or missing from the predicted spans are often not necessary for the tag to be useful, so lenient F_1 may be a more useful measure. Also, the examples point at an inconsistency in the ground truth labels, as the preposition is sometimes included and sometimes not. Updating the ground truth could therefore also result in clearer and more consistent training targets. To further investigate where models tend to make the most errors, we also analysed how often they incorrectly identified spans by examining the positions of the predicted start and end tokens, aiming at uncovering systematic tendencies in span predictions and identify the most common types of errors. The results (reported in Fig. 3) show that late start and early end predictions were the most frequent, meaning that models often missed the beginning of spans and closed them too early. Another notable pattern is that predicted spans rarely extended beyond the true end, indicating a tendency to under-predict span length. These findings further support that one of the main sources

[13] For a comprehensive overview of the types of errors, refer to this notebook in the GitHub repository.

	side	Start			End		
	location	early	exact	late	early	exact	late
lang	model_type						
de	contemporary	0.30	0.29	0.41	0.38	0.49	0.13
	historical	0.33	0.26	0.41	0.35	0.54	0.10
en	contemporary	0.21	0.37	0.42	0.46	0.47	0.07
	historical	0.26	0.35	0.38	0.42	0.48	0.10
fr	contemporary	0.17	0.26	0.57	0.52	0.34	0.14
	historical	0.17	0.30	0.53	0.55	0.36	0.09
it	contemporary	0.24	0.31	0.45	0.42	0.48	0.10
	historical	0.23	0.41	0.36	0.50	0.41	0.09
nl	contemporary	0.22	0.30	0.48	0.41	0.50	0.09
	historical	0.22	0.29	0.49	0.34	0.55	0.11

Figure 3: Late and early start and end in model predictions across all five languages, comparing contemporary and historical models.

Lang.	#	Add	Skip	One vs. multi GT_1, P_m	One vs. multi P_1, GT_m
DE	49	0.31	0.51	0.12	0.08
EN	43	0.49	0.30	0.19	0.05
FR	44	0.00	0.82	0.18	0.00
IT	51	0.18	0.69	0.16	0.00
NL	51	0.25	0.63	0.16	0.00

Table 5: Types of errors per language, whether the predicted has added or skipped tokens w.r.t. the GT, or whether one GT span covers multiple predicted spans (GT_1, P_m) or vice versa (P_1, GT_m).

of difficulty lies in correctly identifying span boundaries.

Zooming in on specific types of errors, we distinguish between the model adding (Add) or skipping (Skip) tokens with respect to the GT, or whether a single GT span corresponds to multiple predicted spans (GT_1, P_m) or vice versa (P_1, GT_m). The percentage of errors per type is shown in Table 5. There are clear differences between languages. For English, the model is more likely to add additional tokens than to skip tokens, whereas for all the other languages it is more likely to skip tokens. For French, skipped tokens represent 81% of the boundary errors. A less common error is that a single GT span is tagged as multiple spans by the model. The reverse is very rare.

Finally, we look at linguistic error patterns in terms of POS tags. For French and Italian, adjectives are the most commonly skipped (30% and 31% of errors respectively). Other notable error groups are skipped determiners for German, Italian and Dutch (8%, 14% and 18% of errors respectively) and added or skipped prepositions for English (12% and 9%), and noun phrases for English, French, Italian and Dutch (9%, 16%, 10% and 8%). We note that far fewer mistakes are made with adverbs and verbs (5% and 4% across all five languages respectively). As mentioned above, these boundary issues could be due to inconsistent ground truth span selection, with adjectives and prepositions more often than not being excluded in the ground truth spans, leading models learning to mostly skip them too.

6. Conclusions

In this work, we conducted a series of experiments across five languages to investigate the impact of historical models on the extraction of olfactory frames, a task previously explored only in contemporary contexts. We fine-tuned multiple historical models and compared their performance with contemporary models in terms of strict and lenient precision, recall, and F_1, while also considering the potential influence of subtoken fertility rate and performing a detailed error analysis. Our results show that historical models improve performance in English, Dutch, and to a lesser extent German, whereas for Italian and French they lead to a decrease in performance. This effect is likely related to the amount and temporal specificity of the pretraining data. These findings raise interesting questions about the notion of a "historical domain". Our results suggest that assuming all historical texts constitute a single domain may obscure the real factors driving model performance. In particular, historical shifts language use (spelling, vocabulary and word usage) is only on aspect affecting the similarity between pretraining corpora and the benchmark data. But there may be many other factors, such as text type, genre and topic that, along with corpus size and coverage, play a key role. Future work should disentangle these factors to better understand the various dimensions of language difference and change and how such characteristics affect model learning. The analysis of lenient scores further highlights that historical models yield gains across nearly all labels and languages, suggesting that the low performance of many labels is at least partially due to span boundary issues, as previously hypothesised by Menini (2024), and confirmed by our span analysis. At the same time, the lenient evaluation also reveals some inconsistencies in the ground truth annotations, indicating areas for potential improvement in the dataset. Overall, our study demonstrates that the use of historical models can provide benefits for olfactory extraction tasks in older texts, while also pointing to important directions for understanding domain specificity, annotation quality, and label-level challenges.

7. Limitations

This study relies on a single experimental setting. While our analyses provide valuable insights into span prediction errors and model behavior, further experiments exploring alternative settings, parameters, and training strategies would be necessary to assess the generalisability of our findings for the task of olfactory event extraction.

8. Acknowledgments

This research was funded by the European Union under grant agreement 101088548 - TRIFECTA. Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. This work used the Dutch national e-infrastructure with the support of the SURF Cooperative using grant nos. EINF-13810 and EINF-16838. Both authors contributed equally to this work.

9. Bibliographical References

Ahmed Alajrami and Nikolaos Aletras. 2022. How does the pre-training objective affect what large language models learn about linguistic properties? *arXiv preprint arXiv:2203.10415*.

Mehdi Ali, Michael Fromm, Klaudia Thellmann, Richard Rutmann, Max Lübbering, Johannes Leveling, Katrin Klug, Jan Ebert, Niclas Doll, Jasper Buschhoff, et al. 2024. Tokenizer choice for llm training: Negligible or crucial? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3907–3924.

Alessio Palmero Aprosio, Stefano Menini, and Sara Tonelli. 2022. Bertoldo, the historical bert for italian. In *Proceedings of the second workshop on language technologies for historical and ancient languages*, pages 68–72.

Ryan Brate, Paul Groth, and Marieke van Erp. 2020. Towards olfactory information extraction from text: A case study on detecting smell experiences in novels. In *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 147–155, Online. International Committee on Computational Linguistics.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Wietse De Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. Bertje: A dutch bert model. *arXiv preprint arXiv:1912.09582*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named entity recognition and classification in historical documents: A survey. *ACM Computing Surveys*, 56(2):1–47.

Charles J Fillmore et al. 2006. Frame semantics. *Cognitive linguistics: Basic readings*, 34:373–400.

Simon Gabay, Pedro Ortiz Suarez, Alexandre Bartz, Alix Chagué, Rachel Bawden, Philippe Gambette, and Benoît Sagot. 2022. From FreEM to d'AlemBERT: a large corpus and a language model for early Modern French. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3367–3374, Marseille, France. European Language Resources Association.

Clovis Gladstone, Zhao Fang, and Spencer Dean Stewart. 2025. Ground truth generation for multilingual historical nlp using llms. *arXiv preprint arXiv:2511.14688*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Leah Hamilton, Caitlin Neill, and Jacob Lahne. 2023. Flavor language in expert reviews versus consumer preferences: an application to expensive american whiskeys. *Food Quality and Preference*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

David Howes. 2006. Charting the sensorial revolution.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001)*, pages 282–289.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

E. Lefever, I. Hendrickx, I. Croijmans, A. Van den Bosch, and A. Majid. 2018. Discovering the language of wine reviews: A text mining account. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Junyi Li, Tianyi Tang, Zheng Gong, Lixin Yang, Zhuohao Yu, Zhipeng Chen, Jingyuan Wang, Xin Zhao, and Ji-Rong Wen. 2022. ElitePLM: An empirical study on general language ability evaluation of pretrained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3519–3539, Seattle, United States. Association for Computational Linguistics.

Enrique Manjavacas and Lauren Fonteyn. 2021. Macberth: Development and evaluation of a historically pre-trained language model for english (1450-1950). In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, pages 23–36.

Enrique Manjavacas Arevalo and Lauren Fonteyn. 2022. Non-parametric word sense disambiguation for historical languages. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 123–134, Taipei, Taiwan. Association for Computational Linguistics.

Steven McGregor and Barbara McGillivray. 2018. A distributional semantic methodology for enhanced search in historical records: A case study on smell. In *14th Conference on Natural Language Processing*.

Stefano Menini. 2024. Semantic frame extraction in multilingual olfactory events. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)*, pages 14622–14627.

Stefano Menini, Teresa Paccosi, Sara Tonelli, Marieke Van Erp, Inger Leemans, Pasquale Lisena, Raphael Troncy, William Tullett, Ali Hürriyetoğlu, Ger Dijkstra, et al. 2022. A multilingual benchmark to capture olfactory situations over time. In *Proceedings of the 3rd workshop on computational approaches to historical language change*, pages 1–10.

Barbara Plank. 2016. What to do about non-standard (or non-canonical) language in nlp. *arXiv preprint arXiv:1608.07836*.

Josef Ruppenhofer, Michael Ellsworth, Miriam RL Petruck, Christopher R Johnson, and Jan Scheffczyk. 2016. Framenet ii: Extended theory and practice.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135.

Stefan Schweter. 2020. Italian bert and electra models. *Zenodo*.

Stefan Schweter, Luisa März, Katharina Schmid, and Erion Çano. 2022. hmbert: Historical multilingual language models for named entity recognition. *arXiv preprint arXiv:2205.15575*.

Dan Sperber. 1975. *Rethinking symbolism*. 11. CUP Archive.

Riste Stojanov, Gorjan Popovski, Gjorgjina Cenikj, Barbara Koroušić Seljak, and Tome Eftimov. 2021. A fine-tuned bidirectional encoder representations from transformers model for food named-entity recognition: Algorithm development and validation. *Journal of medical Internet research*, 23(8):e28229.

Simeng Sun, Brian Dillon, and Mohit Iyyer. 2022. How much do modifications to transformer language models affect their ability to learn linguistic knowledge? In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 46–53, Dublin, Ireland. Association for Computational Linguistics.

Serra Sinem Tekiroğlu, Gözde Özbal, and Carlo Strapparava. 2014. Sensicon: An automatically

constructed sensorial lexicon. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1511–1521, Doha, Qatar. Association for Computational Linguistics.

Sara Tonelli and Stefano Menini. 2021. Framenet-like annotation of olfactory information in texts. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 11–20.

William Tullett, Inger Leemans, Hsuan Hsu, Stephanie Weismann, Cecilia Bembibre, Melanie A Kiechle, Duane Jethro, Anna Chen, Xuelei Huang, Jorge Otero-Pailos, et al. 2022. Smell, history, and heritage. *The American Historical Review*, 127(1):261–309.

Rob Van Der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (machamp): A toolkit for multi-task learning in nlp. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197.

Stella Verkijk, Piek Vossen, and Pia Sommerauer. 2025. Language models lack temporal generalization and bigger is not better. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20629–20637.

Bodo Winter, Marcus Perlman, and Asifa Majid. 2018. Vision dominates in perceptual language: English sensory vocabulary is optimized for usage. *Cognition*, 179:213–220.

Contemporizing 20-th Century Estonian

Heiki-Jaan Kaalep
University of Tartu
Tartu, Estonia
heiki-jaan.kaalep@ut.ee

Abstract

The paper describes a contemporization effort of a 1.9 million word corpus of Estonian parliament minutes from 100 years ago. The paper describes the corpus of Asutaw Kogu (the Constitutional Assembly) and the main differences of language that require one to contemporize it for modern researchers. The effort is implemented as a work flow that combines a freely available speller lexicon, hand-crafted transformation rules and various corpus-based word lists into finite state transducers. Evaluation on a 53,000 token subset of the corpus showed that 0.02% of text tokens ended up with an incorrect contemporary form, corresponding to 0.05% of the corpus vocabulary. However, if we count only the tokens that actually need changing in the contemporization process, we see that 0.12% end up being incorrect, corresponding to 0.15% of the corpus vocabulary. An additional experiment with generative AI showed that using it as a contemporization tool results in a content-preserving, but more formal version of the original minutes.

Keywords: Estonian, finite-state transducer, historical corpus

1. Introduction

Digitization and text recognition (OCR or typing in) creates possibilities for a researcher to "1) perform a full-text search in their digitized corpora, or 2) to perform distant reading: using Natural Language Processing (NLP) tools, automatically process documents in order to make analyses over large corpora. One major obstacle to this research design is language change. The older the historical texts are, the more they diachronically deviate from the current standard language." (Ehrmanntraut, 2024)

The current paper is about contemporizing the language used by the founders of the Republic of Estonia a hundred years ago – language used in the Asutaw Kogu 'the Constituent Assembly' (AK).

A contemporary researcher who would like to find passages in the transcripts that might be relevant for the research question, is confronted with the question: what linguistic expression(s) should one use for querying? Formulating a query requires an hypothesis about how speakers might have used language to express the notions that the researcher is interested in.

Choosing between the linguistic options requires some intuition about the language – its orthographic norms, vocabulary, inflectional and derivational morphology. It is unrealistic to expect a contemporary researcher to have good intuition about a past language. So, it would be nice to have some oracle providing guidance, e.g. providing modern orthographic variants and lemmas of historical words. The oracle could be seen to provide modern paraphrases of the original text.

By way of terminology one should note that it would be inaccurate to use the term "normalize" for describing the changes necessary for making near-past Estonian similar to contemporary one, because it was already a standard normative language, albeit one that is different from the one used today.

2. Related Work

Approaches to historical text normalization can be divided into type- or lexicon-based and token- or text-based ones. A lexicon-based approach tries to utilize the observation that orthographies tend to have more or less regular sound-letter mapping conventions. It has been common for this approach to write rules that correspond to historical changes and variations in these mappings, notably using weighted finite state transducers (FST) (Beesley and Karttunen, 2003), e.g. (Porta et al., 2013) for historical Spanish, (Etxeberria et al., 2016) for historical Basque, (Koskenniemi and Kuutti, 2017) for historical Finnish. A text-based approach views the task as a special type of machine translation. This has evolved to using LLMs, e.g. (Bawden et al., 2022) for early French, (Bracke, 2025) for historical German.

Makarov and Clematide (2020) experimented with historical German, English, Spanish, Icelandic, Portuguese, Hungarian, Slovene and Swedish, representing periods between 14th and 19th centuries. They found that the upper bound on accuracy for a non-contextual oracle that selected the most frequent normalization for each historical word was 97.0 on average, indicating that lexical normalization is a very reasonable strategy.

Proceedings of LT4HALA 2026 @ LREC 2026, pages 400–406.
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

3. Corpus of Asutaw Kogu

The Constituent Assembly was the first parliament of Estonia. It acted as the national representative body and the legislative power of Estonia from April 23rd 1919 until December 20th 1920. Its task was to lay the foundations for the Estonian statehood, to adopt the Constitution and the Land Act. The elections for the Constituent Assembly were held from 5th to 7th April 1919.

The Constituent Assembly held five sessions in 170 sittings, in which 88 laws were passed. During their term of office, the members of the Constituent Assembly also sent greetings to foreign states, discussed the problems connected with research and culture, followed the events in foreign policy, and expressed their opinion about them.[1]

In 2009, students at the Graduate School of Humanities and Sociology, University of Tokyo, under the guidance of Prof. Kazuto Matsumura, created a corpus of Asutaw Kogu by typing in the published minutes of the 170 sittings (Kogu, 1919-1920) (Asutawa Kogu protokollid I-V), a total of 1.93 million words. The corpus is freely available.[2]

Before the 1920-ies, three local languages – Estonian, German and Russian – were widely spoken and understood. Some members of Asutaw Kogu used their constitutional right to give speeches in their native German or Russian in Asutaw Kogu. No translation for these speeches was provided because everyone understood the languages (Raag, 2008, p. 167). These speeches are not included in the corpus.

Corpus text is divided into paragraphs <p> and sentences <s>. The sentences are numbered. Speakers, agenda items, and passages in foreign language are tagged. The original publication presented text in a two-column format. The corpus contains tags <col> indicating the start of a new column; columns are numbered.

4. Deviations from Contemporary Estonian in AK

Over the course of the last 100 years, Estonian has changed. The changes involve pronunciation, morphology, vocabulary and syntax, all having an effect on spelling. A speller of modern Estonian from Filosoft[3] or Giellatekno[4] is able to recognize 57,000 inflectional word forms out of the total 85,000 that make up the word stock of AK, i.e. 67%; this corresponds to 1,600,000 tokens out of the total 1,930,000 corpus tokens, i.e. 83%.

- The most notable change has been purely orthographic – using *v* instead of *w*. This change started in 1905 and ended in the 1930-ies, mainly happening during 1920-ies. Ignoring *v/w* differences, a modern speller would recognize 76,000 or 89% of the vocabulary; 1,870,000 or 97% of the corpus tokens. In other words, it would miss 9000 (11%) of vocabulary items, corresponding to 60,000 (3%) corpus tokens.
- Another purely orthographic change is using *š* and *ž* instead of *sh*.
- Spelling change corresponding to change in pronunciation has been: using *f* instead of *w*, e.g. *wormaalsus* → *formaalsus* ‘formality’, using a double letter instead of a single one (to indicate a long sound), or, vice versa, a single one instead of a double, e.g. *bilans* → *bilanss* ‘balance’, *mehanism* → *mehhanism* ‘mechanism’, *inseneer* → *insener* ‘engineer’, *kommisjon* → *komisjon* ‘commission’.
- Most frequent changes in suffixes have been the following:
  - *ismus* → *ism*, e.g. *radikalismus* → *radikalism* ‘radicality’
  - *iker* → *ik*, e.g. *tehniker* → *tehnik* ‘technician’, *poliitiker* → *poliitik* ‘polititian’
  - *lik*, *line* → *lik*, *line*, *ne*, e.g. *sõjalik* → *sõjaline* ‘military’, *sümpaatlik* → *sümpaatne* ‘likeable’, *loodusline* → *looduslik* ‘natural’, *üleilmline* → *üleilmne* ‘world-wide’
  - *iteet* → *sus*, e.g. *agressiviteet* → *agressiivsus* ‘agressiveness’, *stabiliteet* → *stabiilsus* ‘stability’
  - *mata* → *matu*, e.g. *kõlbmata* → *kõlbmatu* ‘useless
- Some inflectional classes have changed, meaning that all the words that belong to certain class are nowadays inflected (partly) differently than 100 years ago.
  - Plural partitive allomorph of suffix *-kond* has changed from *e* to *i*, e.g. *osakonde* → *osakondi* ‘departments’, plural partitive allomorph *a* has been substituted by *e*, e.g. *kasulikka* → *kasulikke* ‘useful’, *punkta* → *punkte* ‘points’
  - Plurality allomorph of words ending with *-ik* has changed from *ui* to *e*, e.g. *kodanikuile* → *kodanikele* ‘to citizens’

[1] https://www.riigikogu.ee/en/introduction-and-history/history-riigikogu/constituent-assembly/

[2] https://www.cl.ut.ee/korpused/baaskorpus/akp/

[3] https://github.com/Filosoft/vabamorf

[4] https://github.com/giellalt/lang-est-x-utee/

  - Plurality allomorphs *i*, *u*, and *e* have been substituted by *de* as the default allomorph for several inflectional classes, e.g. *rühmis* → *rühmades* 'in groups'
  - Tens of individual words have moved from one inflectional class into another, e.g. *liige* 'member' plural comitative case form has changed from *liigetega* to *liikmetega* 'with members'.

- Haplology has been lost in impersonal forms of *da/ta*-ending verbs, e.g. from *parandakse* to *parandatakse* 'is being repaired'.
- Words have shortened, in particular losing *-eer-*, e.g. *kontroleeri* → *kontrolli* 'control', *analüseeri* → *analüüsi* 'analyze' and *-ne-*, e.g. *tutvunema* → *tutvuma* 'get acquainted', *valminema* → *valmima* 'ripen'.
- Tens of high-frequency words have changed, e.g. a productive first component of compound words has changed from *nõnda* to *nii* 'so', e.g. *nõndapalju* → *niipalju* 'so much'; *siiamaalne* changed to *senine* 'until now', *otstarb* to *otstarve* 'purpose', *ütelus* to *ütlus* 'saying', *kaebtus* to *kaebus* 'complaint'.

5. Operationalization with FST

Mapping historical variants to modern ones is the task of mapping strings, and as such similar to mapping a spelling variant to a norm or an erroneous word to a correct one. Ordering string manipulation rules according to their likelihood, allowing/prohibiting certain rule configurations, and checking the correctness of resulting strings would be a complex task for a programmer; an example of such an approach to process non-standard Estonian is presented by (Kaalep and Muischnek, 2011). Managing this combinatorial complexity becomes much easier with FSTs, because they offer a natural way of avoiding the need to explicitly deal with all of the combinatorics. The implementation described in the current article is available at TartuNLP repository[5]

It is natural to treat a string mapping FST as a composition of two transducers $E \circ V$, where E is an edit transducer and V is a vocabulary that acts as a filter that restricts the set of output strings, e.g. the vocabulary of contemporary Estonian. E is essentially a collection of letter and string pairs representing various ways in which one could modify the input string. The smaller the weight attached to a letter or string pair, the more plausible this change is. In the notation of the HFST toolkit (Lindén et al., 2011) used in this work, an example excerpt from E would be: `[w (->) v::10] .o. [ri (->) n::300]`. The example is a composition `.o.` of two optional mappings `(->)` with vastly different weights `10` and `300`. Composing the mappings means that they are all applied, one by one, from left to right. Given an input string *wabariema* containing substrings *w* and *ri*, this example transducer would yield four output variants, with weights 0, 10, 300 and 310. Composition of this example with V representing modern Estonian vocabulary would result in a FST that filters out all the rest but the last one – *vabanema* 'get free' with weight 310 – as a valid modern Estonian word.

The task of converting the vocabulary of AK to modern Estonian can be formulated as a task of building a transducer that maps every word in AK corpus to its modern counterpart: $A \circ E \circ V$, with A being the vocabulary of AK. Its simplest final form would be a two-column table with every vocabulary item as one row; it is trivial to convert this to the LEXC format (e.g. `wabariema:vabanema # ;`) which in turn is the preferred format for (morphological) lexicons to be compiled into FSTs.

We know A, i.e. the left or input side of the resulting transducer – it is the vocabulary of AK, but we do not have complete knowledge of V, i.e. contemporary vocabulary with lexical coverage that would include modern equivalents for all vocabulary items of AK.

The challenge here is to achieve the necessary lexical coverage, in addition to building the edit transducer E. Fortunately, one can download open source Estonian morphology files from the Giellalt repository[6]), build the morphological analyzer FST, extract its input (i.e. surface) side (discarding the output, i.e. lexical side containing lemma and grammatical category info), and thus get a weighted transducer S that accepts all the word forms a speller is able to recognize.

The weights of S have the following explanation. A surface word form may be ambiguous as to its grammatical categories, and the morphological analyzer FST attaches a weight to every possible reading. The weight is calculated by morphological properties: the weight of the lexeme (based on its rank in a frequency dictionary), plus weights of the inflectional categories (singular, plural, nominative, genitive etc.), plus a weight for every derivational affix, plus a penalty weight for any component stem in case the word is a compound word. This complex calculation scheme gives a context-ignorant likelihood for every morphological analysis variant of a word form. This is what an FST morphological transducer outputs as a default. Having discarded the lemma and grammatical info, we are not inter-

[5] https://github.com/TartuNLP/ak_kaasaegseks_LT4HALA/

[6] https://github.com/giellalt/lang-est-x-utee/

ested in all the readings a word form may have, and choose the variant with the smallest weight, saying that this is the weight of the contemporary word form itself.

6. Correcting errors in the AK

Re-typing in has resulted in errors because of the poor quality of the print, and/or the fact that the typists were students of Estonian as a foreign language, not native speakers.

In addition to keyboarding errors (substitutions, omissions, insertions, transpositions), the corpus contains errors similar to those one encounters with OCR, because the typists have misread typographically confusable letters like *l*-*I*-*i*-*t*, or *rn*-*m*. Accented letters *ä*, *ö*, *õ* and *ü* have been often confused with their un-accented counterparts.

Prior to contemporizing, an attempt to correct the corpus semi-automatically was made. How do you know that a token is actually a spelling error (and thus you cannot transform it into a contemporary word)?

The following assumptions seem justified: 1. Spelling errors are rare; the same error does not happen twice in the same word form. 2. A word form that has an error is unique exactly because of this error; the correct form is not unique, i.e. has a frequency greater than one.

However, being unique does not mean that the word form is erroneous.

So, a unique word form might either contain an error or be a genuinely rare word, and one needs to find out which alternative to choose.

It is actually very difficult to decide whether a word form is part of the vocabulary of a language, and thus could possibly be also part of the vocabulary of a given corpus. If a word is common, it is usually listed in some dictionary, and once one has access to this dictionary, it is easy to look it up. However, no dictionary contains all the words of a language.

Approximately half of the vocabulary of a text corpus normally consists of words that occur only once, i.e. hapax legomena; the exact proportion is dependent on the corpus size (Baayen, 2001). This number of unique words in a text corpus is also approximately equal to the number of unknown text tokens, i.e. words that are missing from any lexicon built from some other corpus, e.g. a speller's word list.

To sum it up: taken together, rare words form a large part of the vocabulary of a corpus. One should be cautious when suggesting that some rare word form is actually an error. Not wanting to introduce any new errors in the corpus (by wrongly assuming that a rare word form should be replaced by a more frequent one), one needs to assess what is more likely: is this a rare word in that corpus, or a mis-spelled form of some other word.

Notice that I rely only on type frequencies and context-independent transformations of a single word form.

For assessment, I create two transducers $E \circ S$ and $E \circ A2$ by composing an edit FST E with S, an FST representing the vocabulary of a speller, and with $A2$, an FST containing the set of non-unique word forms of the AK corpus (with their weights equal to zero). Those two transducers will yield different outputs with weights to the same input string. The candidate with the smaller weight wins.

I allow maximum two changes per word that is at least 4 characters long, and one change per shorter words.

The list of hapax legomena comprises 42,000 words. 1070 of them can be converted to some non-hapax word from, by changing up to 2 letters, i.e. can be recognized by $E \circ A2$. However, 470 of them would be recognized by a modern speller in their original form already (i.e. by S), so these are excluded.

Additionally, 160 are recognized by $E \circ S$, while having a smaller weight than those recognized by $E \circ A2$. This means that the unique corpus word is more similar to a modern word than to a more frequent corpus word and thus it is too risky to assume this is an error.

Half of these 160 words are instances where the unique corpus word is an inflectional form that is missing from this corpus, while a more frequent similar form of the same lexeme is present. E.g. for a lexeme *elav* 'living', the corpus contains a unique singular allative case form *elawale*, its contemporary form would be *elavale*, but the most similar corpus non-unique form is plural genitive *elawate*. The other half are rare words which look almost like some other word, e.g. unique word *wälkus* 'flickered', non-unique word *waikus* 'silence'.

Finally, 440 corpus words get from $E \circ A2$ a weight smaller than they get from $E \circ S$, e.g. unique word *riiglkodanikka*, contemporary *riigikodanikke*, corpus non-unique *riigikodanikka* 'citizens'. These 440 are thus good candidates for being substituted by a more frequent corpus word. However, manual inspection reveals that one group of these candidates is susceptible to being erroneous.

For 380 of these, the changes needed for getting a contemporary word are the same that one needs to get a non-unique corpus word, plus possibly a change from *w* to *v*. So the modern speller and the corpus agree on what the correct form of these words should be. There is no error in this group.

For another 30, there is no way to get a contemporary word form, because they are either family names (which are normally not a part of a lexicon), or word forms very different from contempo-

rary language. E.g. unique form *Strändman*, corpus non-unique form *Strandman*, or unique form *kakskümmendvüs*, corpus non-unique form *kaksümmendviis*, the contemporary form containing a space: *kakskümmand viis* 'twenty five'. This group is very error-prone: 6 instances, i.e. 20% of candidates should not be substitued by a similar-looking corpus non-unique word form, because the unique word form is really a rare instance of a correct word, not an error.

The rest 30 are cases where the possible contemporary word contains more changes or less likely ones than the non-unique corpus form, e.g. unique word *Tonis*, contemporary *Topis* 'stuffed', corpus non-unique *Tõnis* (proper name); unique word *rnaterjaali*, contemporary *materjali*, corpus non-unique *materjaali* 'material'. There are no errors in this group.

Why is the second group so much less reliable? Paradoxically, the fact that a word form can be transformed into a correct contemporary word form, lends credibility to the idea that it might actually be just a typing error. Proper names and word forms, which differ greatly from contemporary language, violate the initial assumption that they are just a few key strokes away from a more frequent, correct word.

7. Creating a Conversion FST

Contemporization in our case means that words are substituted by modern ones, and everything else (word order, layout, mark-up) remains unchanged.

As mentioned earlier, the FST for mapping AK words to contemporary ones would be $A \circ E \circ V$. As the first approximation, I take $V = S$.

If one is unable to find a contemporary version of a word, keep the old version. This is more likely to happen with proper names, acronyms and abbreviations, and typographical errors.

The question is thus how to arrive at the FST V containing a contemporary version for every vocabulary item of AK corpus.

The steps would be the following.

1. Create a frequency list of AK corpus words, stripping them of punctuation, and splitting also at hyphens, but without downcasing. The resulting lexicon has 85,000 items.

2.1. Split compound words which are nowadays written with a space, e.g. **kõigeõigem* -> *kõige õigem* 'most correct'.

2.2. Contemporize AK words with $E \circ S$. The final total weight of a word form (after it has passed through this transducer) is a sum of two components: summary of weights of letter changes from E plus weight of the contemporary word form from S. The larger the weight, the less trustworthy the result is (because it is a combination of rarer events).

3. Can I be sure that the output of the transducer $E \circ S$ is really a contemporary version of the original, and not some wildly different word? One might make a lot of changes, to eventually arrive at some acceptable word with a large weight. Obviously, it would be sensible to discard words above some threshold.

However, a word with a large weight might also be a rare inflectional form of a rare compound, the weight from S indicating the unlikely process of its morphological creation, and not the complex way of changing the original by E to arrive at it. It would be nice if I knew that this word is a legitimate word in our corpus; the contemporization weight would then reflect solely the likelihood of changes. So I should include contemporized versions of words that surely exist in AK corpus (S_{AK}) in V, thus $V = S \cup S_{AK}$. I assume that a word shorter than 3 letters, or a contemporized word with a small weight is correct, and thus belong to S_{AK} (with a weight of zero).

4. Proper names should be contemporized, but more cautiously than other words: only the most common orthographical changes (*w* -> *v*, *sh* -> *š*) are allowed (E_{min}). The proper name transducer would be $P_A \circ E_{min}$, where P_A denotes proper names in AK.

When making a list of proper names of AK, I took a position to favor precision over recall. A proper name is a word with an initial capital letter, which occurs sentence-internally, and which has no lowercased version in the corpus. The latter condition is necessary because an orthographic sentence may contain a part starting with a capitalized word, e.g. direct speech, a comment in parenthesis etc. I identified 3100 proper names this way.

5. So the final transducer is $((A - P_A) \circ E \circ (S \cup S_{AK})) \cup (P_A \circ E_{min})$

Using this final transducer revealed that 3000 entries of the 85,000 lexicon were not converted to a plausible contemporary word. When making the final 85,000-entry 2-column table prior to converting it to LEXC format and compiling into a transducer for contemporizing the corpus texts, these 3000 entries were included as simply pairs of identical strings.

They can be grouped in the following way: abbreviations and acronyms, typos (often caused by an inserted or a deleted space character), old inflectional or derivational forms (not covered by mappings of the edit transducer E), loanwords with outdated orthography, proper names which had not been classified as such by the method described above, and some lexemes which contemporized form is missing from the modern speller lexicon.

It should be noted that I have found only one old lexeme that has diverged into more than one modern lexeme – *näitus* 'exhibit' into *näitus* 'exhibition'

and *näide* ‘example’. At first glance, it seems impossible to have a simple 1-to-1 conversion in this case. However, on closer examination, it appears that different inflectional forms of *näitus* in the AK corpus are used in different senses, e.g. *näituseks* ‘for example’ would be contemporized as *näiteks*; *näitusel* ‘at an exhibition’ would be contemporized as *näitusel*.

8. Evaluation

For evaluation, the minutes of 5 sittings (9, 11, 12, 31 and 74) were chosen randomly and checked manually. They contain 53,000 tokens (2.8% of the corpus text volume), with a vocabulary of 10,000 (11.75% of the corpus vocabulary). Both type-in errors and contemporarization errors were checked by comparing with the original PDF-files.

The final transducer contains a few instances where a word with a type-in error gets correctly contemporized, e.g. *wõtrna*, contemporary *võtma* ‘to take’. However, in the checked files, a type-in error always resulted in failure to get a correct contemporary word.

The number of letter type-in errors was found to exceed the number of contemporization errors very slightly. There were 14 word-internal type-in errors – 7 involving a letter, and 7 being failures to delete a hyphenation mark induced by end of line; versus 11 contemporization errors. These 11 errors involved 5 different words, among them three proper names *Westholm*, *Wiedemann* and *Kurs-Olesk* that should not be changed. It is noteworthy that re-typing contributed as much to the overall error rate as did making the words contemporary. In addition, there were 60 type-in errors of changing, adding or deleting a punctuation mark.

The evaluation shows that 0.05% of the resulting vocabulary is populated by incorrectly contemporized items, and the same is true for 0.02% of the text tokens. The large amount of punctuation mark errors came as a surprise. However, this had no impact on the procedure of contemporizing, because the method is word based, and ignores punctuation marks. It will probably have an impact on some follow-up steps of the corpus processing workflow that operates on the contemporized texts, e.g. POS-tagging and lemma disambiguation, but this has not been evaluated.

Remember that 67% of the vocabulary and 83% of the corpus tokens are similar to modern Estonian. This means that only 3300 vocabulary items and 9000 text tokens of the evaluation set are in need of contemporization; thus 0.15% of vocabulary items needing contemporization end up being incorrect, and 0.12% of the corpus tokens.

9. Contemporizing with Office 365 Copilot

A natural idea for transforming one text into another is to use a generative AI model. I made a small experiment with Office 365 Copilot. I gave Copilot two 50-sentence-long extracts. One extract is a part of a continuous speech by a member of AK. The other extract is part of a procedure for adopting a law, where the items are read out loud, commented on briefly, and voted on. I prompted Copilot with "*Alljärgnev tekst on vanas eesti keeles. Teisenda see tänapäevaseks eesti keeleks.*" ‘The following text is in old Estonian. Transform it into modern day Estonian’.

When comparing the AI generated texts with the ones where only words had been contemporized, I found that 1) Copilot preserved the contents of the original, 2) Copilot converted individual words into modern Estonian correctly, 3) Copilot paraphrased a lot, substituting words and expressions with modern ones, reordering text fragments, and changing punctuation and text layout.

As a result, the output by Copilot feels like it has made a modern edited version of the minutes. The text retains the meaning of the original, presenting it in a more formal way, not keeping the original way the speakers expressed themselves.

Several runs with the same texts and prompt were made. The runs were sometimes immediately after each other, sometimes days apart. The resulting texts were very different from each other, but always kept the original meaning and expressed it in the form of modern meeting minutes. It is noteworthy that not a single instance of misrepresenting the original content was found during these runs.

Assuming future AI experiments will confirm the preservation of content when changing the presentational form to a modern one, one could envisage that contemporization would boil down to creating another corpus with AI, parallel to the original AK, and linked to it. This heavily edited new version would be usable for someone who intends to read the original closely but needs help navigating the vast corpus to find the part that needs close reading.

10. Conclusion

A freely available speller lexicon, hand-crafted transformation rules, and various corpus-based word lists were combined into a work flow for contemporizing a corpus of Estonian spoken and written 100 years ago. Evaluation showed that 0.02% of text tokens ended up with an incorrect contemporary form. However, if we count only the tokens that actually need to be changed in the contemporization process, we see that 0.12% end up being

incorrect.

11. Acknowledgements

This work has been supported by the Estonian Research Council grant PRG2006 and by the national program "Estonian Language and Culture in the Digital Age" project EKKD-TA10.

12. Bibliographical References

R Harald Baayen. 2001. *Word frequency distributions*, volume 18. Springer Science & Business Media.

Rachel Bawden, Jonathan Poinhos, Eleni Kogkitsidou, Philippe Gambette, Benoît Sagot, and Simon Gabay. 2022. Automatic Normalisation of Early Modern French. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 3354–3366, Marseille, France. European Language Resources Association.

Kenneth R Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI, Stanford.

Yannic Bracke. 2025. A lexical normalizer for historical spelling variants using a transformer architecture. https://github.com/ybracke/transnormer. GitHub repository.

Anton Ehrmanntraut. 2024. Historical german text normalization using type- and token-based language modeling. *arXiv preprint arXiv:2409.02841*.

Izaskun Etxeberria, Iñaki Alegria, Larraitz Uria, and Mans Hulden. 2016. Evaluating the noisy channel model for the normalization of historical texts: Basque, Spanish and Slovene. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1064–1069, Portorož, Slovenia. European Language Resources Association (ELRA).

Heiki-Jaan Kaalep and Kadri Muischnek. 2011. Morphological analysis of a non-standard language variety. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, pages 130–137.

Asutaw Kogu. 1919-1920. *Asutawa Kogu I-V istungjärk. [Sessions 1 to 5 of the Constitutional Assembly]*. Täht, Tallinn, Estonia.

Kimmo Matti Koskenniemi and Pirkko Kuutti. 2017. Indexing old literary finnish text. In *K + K = 120: Papers dedicated to László Kálmán and András Kornai on the occasion of their 60th birthdays*. Research Institute for Linguistics, Hungarian Academy of Sciences.

Krister Lindén, Erik Axelson, Sam Hardwick, Tommi A. Pirinen, and Miikka Silfverberg. 2011. Hfst—framework for compiling and applying morphologies. In *Systems and Frameworks for Computational Morphology*, pages 67–85, Berlin, Heidelberg. Springer Berlin Heidelberg.

Peter Makarov and Simon Clematide. 2020. Semi-supervised contextual historical text normalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7284–7295, Online. Association for Computational Linguistics.

Jordi Porta, José-Luis Sancho, and Javier Gómez. 2013. Edit transducers for spelling variation in old spanish. In *Proceedings of the workshop on computational historical linguistics at NODALIDA*, pages 22–24.

Raimo Raag. 2008. *Talurahva keelest riigikeeleks [From Peasant Language to State Language]*. Atlex, Tartu, Estonia.

Cost-Aware Pre-Annotation Strategies for Nested NER in Historical Latin Notarial Deeds

Charlene Ellul, Vanessa Buhagiar, Claudia Borg and Charlie Abela
Department of Artificial Intelligence, University of Malta, Malta
{charlene.ellul, vanessa.buhagiar, claudia.borg, charlie.abela}@um.edu.mt

Abstract

Manual annotation for Named Entity Recognition in historical documents remains expensive and time-consuming, particularly for complex nested entity structures in domain-specific texts such as Latin notarial deeds. Active learning frameworks like the Humanities Entity Recognizer (HER) reduce annotation requirements by iteratively selecting informative samples for expert annotation, nevertheless, existing sentence-based sampling strategies create unpredictable annotation costs when sentence lengths vary extensively. We extend the HER to support nested entities through composite BIO label encoding and introduce token-budgeted sample selection to address annotation cost variability. Under token-budgeting, each annotation iteration targets a fixed token budget rather than a fixed sentence count, while Active Curriculum Learning (ACL) ensures diverse sentence length representation in initial samples. Experiments on seventeenth-century Latin notarial deeds from Malta's Notarial Registers Archive demonstrate that token-budgeted sampling achieves comparable macro-span F1 to sentence-based sampling while exhibiting more stable learning trajectories across iterations. Additional experiments examining entity-level performance revealed systematic variation by semantic granularity, with higher-level categorical entities achieving stronger recognition than role-based middle-level entities, which depend on discourse context. Our results demonstrate that controlling sample selection at the token level rather than sentence level provides more predictable annotation planning for active learning in historical document corpora with heavy-tailed sentence length distributions.

Keywords: Latin notarial documents, active learning, nested entities, historical documents

1. Introduction

The Notarial Registers Archive (NRA)[1] in Valletta, Malta preserves an extensive collection of more than 24,000 bound volumes, fragments, and loose folios that span seven centuries of history of the Maltese islands, with the earliest records dating to the fifteenth century. Managed by the Notarial Archives Foundation (NAF)[2], this growing collection encompasses a wide range of legal documents, including marriage contracts, property transactions, and documents relating to commerce and maritime trade. Despite being a fundamental source for the study of Maltese history, this collection is limited in the meaningful access it can provide to the researcher, as the only systematic finding aid is the *Repertorio*, which is an index, typically preserved in the form of a bound manuscript that lists the party names and folio references, offering no insight into the semantic content of the deeds themselves. Another barrier to accessing this collection is that of scribal variability, and it is an issue that is shared by other archival collections in Malta and beyond. The case of the NRA adds a layer of complexity owing to the incorporation of Italian, Sicilian, and Maltese linguistic elements within the main textual corpus that is written in Latin and in later centuries in Italian.

Accurate interpretation of these records requires years of study to build familiarity with the different script types and hands in which the deeds were written, and hinges on a confident understanding of the large compendium of abbreviations and non-standardised historical orthography scribes used in their work. These competencies are increasingly rare, and without them the textual content of the archive remains effectively inaccessible. As a result, a vast documentary heritage that is meticulously preserved risks becoming a silent archive: its knowledge locked within pages that few can decipher.

Recent work in archival and library studies has increasingly argued that archival catalogues and descriptive metadata should not be treated merely as static finding aids, but as structured data capable of supporting large-scale computational analysis and reuse. The adoption of Linked Data and ontology-based models enables archival descriptions to be transformed into interconnected semantic representations that support querying, aggregation, and integration across collections and institutions (Lapôtre, 2017; Bensmann et al., 2017).

This conceptual shift reframes archives as dynamic knowledge infrastructures rather than passive repositories, in which descriptive metadata and document-level content form part of a unified data system (Koch et al., 2023). This underpins initiatives such as the NotaryPedia (Ellul et al., 2019) project that leverages this perspective of a dynamic knowledge infrastructure and seeks to inte-

[1] https://nationalarchives.gov.mt/en/notarial-registers-archive-valletta/
[2] https://nafmalta.org/

Proceedings of LT4HALA 2026 @ LREC 2026, pages 407–417.
11 May 2026. *©ELRA Language Resources Association (ELRA), 2026*

grate archival standards, digitised notarial sources, and language technologies for knowledge extraction into a unified knowledge management platform. Within NotaryPedia, these Linked Data principles are operationalised by enriching archival catalogue descriptions with Named Entity Recognition (NER)-derived entities extracted from the deeds themselves, enabling queryable knowledge structures over deed content.

A major obstacle in developing effective NER models for historical archives is the cost of transcribed datasets and the manual annotations carried out on them. As highlighted in recent surveys (Ehrmann et al., 2023; Keraghel et al., 2024), historical NER typically depends on transcriptions that are often produced or validated with palaeographic expertise, and on entity annotation performed by domain experts, such as historians. Scribal and orthographic variation, abbreviations, domain-specific language, and the need for interpretive judgement make large-scale gold-standard annotation impractical.

In this work, we focus on annotation cost as a planning problem rather than as a direct measurement of annotation time. For historical corpora with heavy-tailed sentence-length distributions, selecting a fixed number of sentences per iteration leads to highly unpredictable annotation workloads. We therefore investigate whether controlling sample selection at the token level can stabilise annotation effort proxies while maintaining learning effectiveness in active learning–based pre-annotation for nested NER in seventeenth-century Latin notarial deeds from the NRA. Preliminary palaeographic assessment of the corpus found that the notarial deeds from the seventeenth century were more consistently legible and less heavily abbreviated than those of the preceding century, thus making this subset suitable for an initial controlled study.

We compare three pre-annotation strategies reflecting common practice and documented limitations: (i) sentence-based sampling as implemented in HER (Erdmann et al. (2019)), (ii) token-budgeted sampling combined with ACL to improve early-stage coverage across sentence lengths and address cost variability noted in historical NER, and (iii) a staggered strategy that incrementally introduces semantic complexity across nested entity levels. The goal is not only to improve model performance but also to develop annotation workflows that are realistic and sustainable for archival institutions.

2. Background And Related Work

2.1. Annotation Cost As A Central Constraint In Historical NER

In a comprehensive survey, Ehrmann et al. (2023) identifies lack of resources as one of the four core challenges of historical NER, alongside document diversity, noisy input, and linguistic variation. Settles et al. (2008) emphasise that annotation effort is not constant per sentence. It varies with sentence length and the complexity of language used. These are properties that are especially pronounced in administrative and legal records, such as notarial deeds.

2.2. Active Learning for NER

Pre-annotation reframes annotation as a human-machine collaboration, in which models generate candidate annotations that are then validated and corrected by language experts. This approach reduces cognitive load and accelerates annotation while preserving expertise.

Pre-annotation can be integrated with active learning, allowing models to select the most informative samples for annotation. A key contribution in this area is the Humanities Entity Recognizer (HER, Erdmann et al., 2019), explicitly designed for humanities data. HER combines sampling seeds to train CRF models iteratively, using only a subset of the manually annotated data.

Active learning is a well-established methodology in NLP. Settles (2009) describes core strategies, including uncertainty sampling, query-by-committee, and density-weighted methods. After being developed and most extensively studied for classification, active learning query strategies were later adapted to structured prediction, including sequence labelling tasks such as NER. Shen et al. (2017) show that standard uncertainty measures for NER operate at the sequence level and that common least-confidence formulations can disproportionately select longer sentences because uncertainty is aggregated across tokens. They propose length-normalised scoring to mitigate this bias. This characteristic is particularly pronounced in historical texts, where long and complex sentences can dominate active learning selection.

2.3. NER in Latin and Historical Documentary Texts

Recent work on NER for historical documentary texts has focused on medieval charters. Chastang et al. (2021) utilise a CRF-based model to handle the shift from single names to complex multi-component denominations and the prevalence of nested entities within legal Latin char-

ters using Burgundian diplomatic documents from the 9th to the 14th century. To show that formulaic legal discourse can support strong cross-regional and cross-chronological generalisation, Aguilar and Stutzmann (2021) train spACy, Flair, and a custom BiLSTM-CRF model on French medieval charters. Integrating NER into Handwritten Text Recognition (HTR) pipelines has also become a priority to reduce cascading errors in noisy archival data. At manuscript level, Boroş et al. (2020) demonstrated that combined HTR+NER approaches showed superior performance over sequential pipelines HTR→NER approaches on multilingual corpora of Latin, German, and Czech charters. Torres Aguilar (2022) uses multi-lingual medieval charters including Latin, with stacked embeddings and BERT-based models to address complex multilingual writing practices, effectively modelling code-switching and bilingual sequences without a performance drop compared to monolingual counterparts. Novotny et al. (2023) presented a bootstrapping pipeline for late medieval charters in Czech, Latin, and German, demonstrating the scalability of these techniques for large-scale historical datasets. Related workflow-oriented research is the research on the late medieval Bolognese Memoriali series (1265–1452) Loss et al. (2025), where they developed a tailored tagging system for facilitating direct metadata extraction from handwritten notarial records into structured databases. Our study shifts attention from medieval charter corpora to seventeenth-century Latin notarial deeds and focuses specifically on nested NER within a cost-aware pre-annotation and active-learning workflow.

2.4. Nested Named Entity Recognition

Nested NER has been recognised as a distinct problem in NLP for over a decade. Finkel and Manning (2009) introduced one of the earliest models for nested entity recognition using constituency parsing, while Lu and Roth (2015) introduce *mention hypergraphs* for joint mention extraction and classification, designed to capture overlapping entity spans. More recent neural approaches achieve strong results on standard nested NER benchmarks, including merge-and-label architectures (Fisher and Vlachos (2019)) and sequence-based learning/decoding methods for extracting inner entities (Shibuya and Hovy (2020)). However, these approaches are typically evaluated on contemporary datasets with abundant annotations and relatively stable languages.

The lack of work explicitly addressing active learning for nested NER in historical contexts represents a significant gap in the literature. Addressing this gap requires rethinking pre-annotation not merely as a speed-up mechanism, but as a structural annotation methodology aligned with semantic granularity and annotation effort.

3. Archival Context And Corpus

The corpus used in this study is selected from the NRA whose main collection consists of registri, which are the official copies of the notarial acts. Buttigieg and Abela (2020) describe a notary's workflow beginning with the *bastardellum* (draft copy) that functioned as a notebook in which the notary recorded the essential details of a deed, such as the date, type of contract and the names of the contracting parties. These notes were eventually expanded into their full legal form in the *minutarium* (original), a volume containing the original version of the act which also carried the highest significance in legal terms. A faithful copy of the original deed was transcribed chronologically in the *registrum* (register). These registers are one of the richest sources for understanding social, economic, and legal life in Malta and the central Mediterranean.

In this work, we use a sample of 102 deeds from a single bound register: the Register of Notary Giovanni Battista Micallef, 1633 – 1634, Volume 1 (R352/1). This register contains notarial deeds primarily written in Latin, although instances occur in which deeds begin in Latin and subsequently continue in Italian/Sicilian. In addition, where no suitable Latin or Italian/Sicilian equivalent exists for a Maltese vernacular term, such words are occasionally rendered in a Latinised form (e.g., *dublectum*[3], skirt). The texts, therefore, exhibit lexical and syntactic influence from Latin and Italian, alongside extensive use of abbreviations typical of notarial practice. The deeds vary in content and length, and also display considerable variation in sentence length, with some clauses spanning several hundred tokens. Owing to the absence of terminal punctuation, which is typical of the Latin language in general, paragraph boundaries are also treated as sentence boundaries.

Figure 1 shows an analysis of sentence length in the corpus, revealing a strongly right-skewed distribution. The majority of sentences contain fewer than 30-40 tokens, while a long tail of very long sentences extends beyond 100 tokens and, in some cases, exceeds 300 tokens. Heavy-tailed distributions of this kind are common across linguistic data. Bentz et al. (2014) have shown that languages with richer inflectional morphology tend to exhibit heavier-tailed patterns in their word-frequency distributions, suggesting that morphological complexity can influence the shape of such distributions.

The corpus is accompanied by expert-curated reference material, including gazetteers derived from

[3]NRA, R494/1/SUB/3, Notary Giacomo Zabbara, f.3r, 11 July 1495

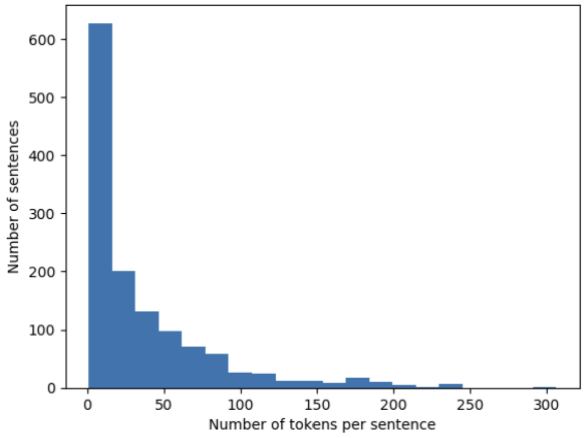


Figure 1: Token distribution per sentence in fullCorpus.

established publications (Wettinger, 2000; Fiorini, 1996).

The transcriptions used in this study were produced as part of the broader NotaryPedia workflow and were normalised for the present experiments by removing inline markup and converting them into a token-per-line format, resulting in 48,197 tokens. [4]

These corpus characteristics motivate the pre-annotation and active learning strategies described in the next section.

4. Annotation Scheme and Entity Hierarchy

4.1. Motivation for Nested Entity Annotation

In our work, nested entities have two purposes. They capture both functional and ontological hierarchies present in notarial deeds. For person entities (PER), nesting reflects both role-based distinctions, such as whether a person is a party or a witness (PER-P/PER-W), and structural components, namely first names and last names (PERFN and PERLN). Similarly, keyword entities can represent part-whole entities, such as a geographical location and these can be embedded within a phrase. For instance, in expressions such as *commrij commendae de Rocchabruna* (the Commander of the Commandery of Roccabruna), the full phrase is annotated as (KEY), while the embedded place name *Rocchabruna* is simultaneously annotated with the geographical entity (GEO), yielding the composite label (explained further in Sec 4.3) `KEY-I+GEOB-B`. Although these nested entities may share token boundaries, they represent distinct semantic layers that are essential for the accurate interpretation of notarial records.

4.2. Entity Levels

Entities are organised into three conceptual layers that reflect different semantic roles and levels of detail. Higher-level entities are the core semantic categories. Middle-level entities are the role-based distinctions, capturing the legal role within a deed. Parties are annotated only when explicitly listed in the marginal notes. Witnesses are usually mentioned at the end of a deed. Lower-level entities are the structural or ontological components, capturing the fine-grained semantic details of a higher-level entity. A definition of all entity labels per level is found in Table 1.

Level	Label	Description
Higher	`DATE`	Temporal expressions
Higher	`DEEDT`	Deed type, typically assigned by the notary, often recorded in the margin
Higher	`GEO`	Geographic entities including place names, territorial references and buildings
Higher	`KEY`	Interesting keyphrases
Higher	`PER`	Person entities
Higher	`INS`	Institution
Middle	`PERP`	Party of a deed
Middle	`PERW`	Witness of a deed
Lower	`DATEAN`	*a nativitate* date format
Lower	`DATEANIN`	*a nativitate* and indiction date format
Lower	`DATEAI`	*ab incarnatione* date format
Lower	`DATEAIIN`	*ab incarnatione* and indiction date format
Lower	`DATEIN`	Indiction date format
Lower	`GEOB`	Buildings
Lower	`GEOPC`	Geographic location where the deed was drawn
Lower	`PERFN`	Person first name
Lower	`PERLN`	Person last name
Lower	`PERNN`	Person nickname

Table 1: Entity labels grouped by semantic level in the annotation scheme.

4.3. Encoding Nested Entities

HER does not inherently support nested entities. To preserve the original architecture while extending its functionality, nested entities were linearised using composite BIO labels (Straková et al., 2019)

[4] The original transcriptions preserves additional inline annotations for palaeographic and documentary phenomena, such as marginal notes, abbreviations, and layout cues. These annotations are not used in the present NER experiments, but were retained because the corpus is intended for reuse in future NotaryPedia modules, including automatic transcription and richer information extraction.

that encode hierarchical structure by concatenating entity tags with a '+' delimiter (e.g. `GEO-B+GEOPC-B` or `PER-B+PERP-B+PERFN-B`). This strategy enables nested entities to be predicted within a single tagging layer. Annotation was conducted in a staggered manner, progressively introducing lower-level entities after higher-level entities had been established.

5. Pre-Annotation Approach

We used the HER (Erdmann et al., 2019) as the baseline active learning framework and describe the extensions required to support nested entity annotation in seventeenth-century Latin notarial deeds. HER was developed for humanities and historical collections, where annotated data are scarce and pre-annotation is intended to support expert correction rather than replace it. Accordingly, our aim is not to compare state-of-the-art NER architectures, but to examine how alternative pre-annotation and sample-selection strategies influence learning behaviour and annotation planning in a low-resource historical setting. We therefore adopt HER's Conditional Random Field (CRF)-based sequence labelling models, which are computationally lightweight and suitable for iterative active learning workflows.

This study does not measure annotation time directly. Instead, annotation effort is approximated using two observable values: (i) the number of annotated tokens per iteration and (ii) the distribution of nested entity depths within each annotation batch. Token count captures gross workload size, while nesting depth reflects semantic and structural complexity. These measures allow us to compare annotation strategies under controlled and reproducible conditions, while acknowledging that actual annotation time may vary with document variation and annotator expertise.

5.1. HER-Based Pre-Annotation Baseline

While Erdmann et al. (2019) report learning curves at fixed token milestones for experimental consistency, our baseline follows the HER repository workflow and uses fixed sentence batches (200 sentences per iteration) to mirror typical tool usage. The individual deeds, distributed across multiple text files, are pre-processed to remove inline annotations and to resolve hyphenated line breaks. The deeds are then concatenated and converted into a one-token-per-line format of the form `<label><tab><token>`, with all labels initialised to `O` (Outside). This constitutes the full unlabelled corpus used throughout the experiments.

The workflow begins with selecting 200 randomly sampled sentences, which are manually annotated as the seed corpus. The remaining sentences constitute the unannotated corpus. Gazetteers are used to perform rule-based pre-annotation, thereby expanding the manual annotation of the seed corpus.

The seed corpus is used to train a CRF tagger. The trained model can be used to pre-annotate the candidate entities of the unannotated corpus. If the performance of the trained CRF model is deemed unsatisfactory, the gazetteers are updated with the newly annotated entities, and a ranking mechanism is applied on the unannotated corpus based on the Pre-Tag DeLex (PTDL) (Erdmann et al. (2019)) strategy to rank them according to how useful they would be to annotate next.

PTDL prioritises sentences likely to contain entities by targeting informative out-of-vocabulary (OOV) tokens. It begins with pre-tagging the unannotated corpus using gazetteer matches extracted from the annotated seed. HER splits it into UNE (sentences containing a pre-tagged entity) and UnoNE (all remaining sentences). It then trains feature-based linear-chain CRF models (CRFsuite; Okazaki (2007)) using only delexicalized (context-based) features and ranks sentences by summing the weighted frequency scores of unique OOV tokens, normalised by sentence length. In our implementation, we ran PTDL for four active-learning rounds. In each round we manually annotated the top-ranked 200 additional sentences, retrained the models, and re-ranked the remaining pool of unannotated corpus. Model training and feature selection are described in Section 5.3.

5.2. Gazetteers Construction

In this study, the initial gazetteers were manually compiled by an expert annotator using established scholarly reference works. These include Place-Names of the Maltese Islands, ca. 1300–1800 (Wettinger, 2000) and the indexes of the Documentary Sources of Maltese History series (Fiorini, 1996). These publications cover personal first names, last names, place names, buildings, institutional entities, and keywords. After each annotation iteration, the gazetteers are expanded using newly annotated data, with additional gazetteers also created for nested entity types.

5.3. Model Training and Feature Selection

For each iteration, a CRF model was trained using three-fold cross-validation on the seed corpus, with two-thirds of the data used for training and one-third for testing in each fold. We trained multiple CRF models using predefined feature-set templates, including word-shape features, character n-grams, contextual token windows (e.g., previous

word/bigram), and gazetteer indicators, and evaluated each configuration by macro span-F1 averaged across folds. The best-performing feature configuration was then used to train the final model on the full seed set for that iteration. Experiments were run on a CPU-only Linux environment (Ubuntu 24.04 LTS, 2 vCPUs, 8 GB RAM).

6. Methodology

6.1. Token-Budgeted Seed Selection Approach

A key contribution of this work is the replacement of sentence-based seeds with token-budgeted seeds. Because the sentence length distribution is skewed, selecting a fixed number of sentences can result in a disproportionate share of annotation effort being allocated to a small number of very long sentences. By introducing a constraint whereby each annotation batch contains approximately 2,000 tokens, this imbalance in annotation cost is substantially reduced. We set 2,000 tokens as a practical compromise: large enough for meaningful retraining, small enough for frequent iterations and to avoid domination by very long sentences.

6.2. Active Curriculum Learning

To address representation imbalance in the initial seed, an ACL strategy is introduced for the first 2,000-token batch. Under this strategy, instead of selecting a random set of sentences from the unlabelled corpus, sentences are first grouped into bins based on token length (e.g. 1–20, 21–50, 51–100, 101–200, >200 tokens). This curriculum-based constraint is applied only in the first iteration to accelerate model learning by exposing the CRF to a representative range of sentence lengths early, following the general principles of ACL described by Jafarpour et al. (2021). From the second iteration onward, the token-budgeted condition followed the same HER/PTDL ranking procedure as the sentence-based baseline, except that sentences were accumulated until the batch reached approximately 2,000 tokens rather than a fixed count of 200 sentences.

6.3. Staggered Training By Entity Level Approach

The HER baseline using the sentence-based approach was also tested without using the nested entity composite label encoding. CRF models were trained separately on Higher-Level, Middle-Level, and Lower-Level entities (refer to Table 1). This staggered training strategy reflects the hierarchical structure of the annotation scheme and the varying semantic complexity associated with each entity group. It also mirrors realistic annotation workflows in archival contexts, where coarse-grained categories are typically established before finer distinctions are introduced.

6.4. Evaluation Metrics

Evaluation focuses on metrics that are meaningful for pre-annotation quality. Token-level accuracy reflects overall tagging correctness but is heavily influenced by non-entity tokens because of class imbalance. Macro-span F1 is computed over exact entity spans, which better captures boundary detection and nested entity correctness. All results are reported using span-level precision, recall, and macro-averaged F1.

7. Results

This section reports the results of three pre-annotation experiments designed to assess (i) the effectiveness of sentence-based pre-annotation as implemented in HER, (ii) the impact of token-budgeted pre-annotation combined with ACL, and (iii) the effect of staggered annotation by entity level.

7.1. Sentence-Based Pre-Annotation (HER Baseline)

We use HER's standard sentence-based active-learning loop as the baseline, adding 200 manually annotated sentences per iteration. Note that this baseline mirrors the HER toolkit's practical annotation loop (fixed-size sentence batches) rather than the token-based batch schedule used in Erdmann et al. (2019) to standardise learning-curve evaluation. The best baseline performance resulted in a macro span-F1 of 0.6675 with feature configuration including (`wordShape`, `charNgrams`, `prevWord`, `prevBiWord`, `prevWordShape`, `contextPosition`, `gazetteers`). As shown in Table 2, macro span-F1 fluctuates between iterations, but improves overall relative to the initial seed, suggesting that additional entity annotations generally increase entity coverage. A decrease in macro-span F1 is observed in the final iteration, despite additional annotation effort. This highlights the limitations of sentence-level sampling under highly skewed sentence-length distributions and motivates the token-budgeted approach introduced next. While sentence-based pre-annotation improves overall entity coverage, annotation costs remain unpredictable due to extreme variation in sentence length.

Iteration	Seed Sentences	Macro span-F1
1	200	0.6288
2	400	0.6090
3	600	0.6675
4	800	0.6412

Table 2: Results of the sentence-based pre-annotation using HER.

7.2. Token-Budgeted Pre-Annotation With Active Curriculum Learning

To address manual annotation cost imbalance, we introduced a token-budgeted condition in which annotation batches contained approximately 2,000 tokens rather than a fixed number of sentences. In the first iteration, an ACL strategy was applied by sampling sentences proportionally across sentence-length bins to ensure representative exposure to both short and long sentences. Figure 2 shows the macro-averaged span-level F1 across iterations for sentence-based pre-annotation and token-budgeted pre-annotation with ACL. Each point in Figure 2 is obtained by three-fold cross-validation on the labelled seed available at that iteration. The figure, therefore, reflects learning progress under increasing labelled data rather than performance on a single fixed held-out test set. Token-based sampling exhibits a smoother learning trajectory across iterations compared to sentence-based sampling. The two conditions are compared as alternative annotation-budget regimes rather than as matched batch sizes: the sentence-based baseline adds 200 sentences per round, which in our runs corresponded to 6,076–13,780 newly annotated tokens, whereas the token-budgeted condition remained close to the intended budget of approximately 2,000 tokens per round.

As shown in Table 3, token-budgeted sampling maintains a consistent annotation workload and reaches macro span-F1 values that are comparable to the sentence-based baseline, while exhibiting a smoother learning trajectory across iterations. Performance peaks at iteration 5 (macro span-F1 = 0.6377) with the feature set (`wordShape`, `charNgrams`, `prevWord`, `prevBiWord`, `gazetteers`). The curriculum-constrained first iteration accelerates early gains relative to random sentence selection, particularly for entities occurring in longer and syntactically complex sentences. Overall, these results demonstrate that controlling annotation cost at the token level leads to smoother learning behaviour while preserving, and in some cases improving, entity coverage.

Full per-entity precision, recall, and F1 scores for the best-performing iteration of each experiment are provided in Appendix 10.

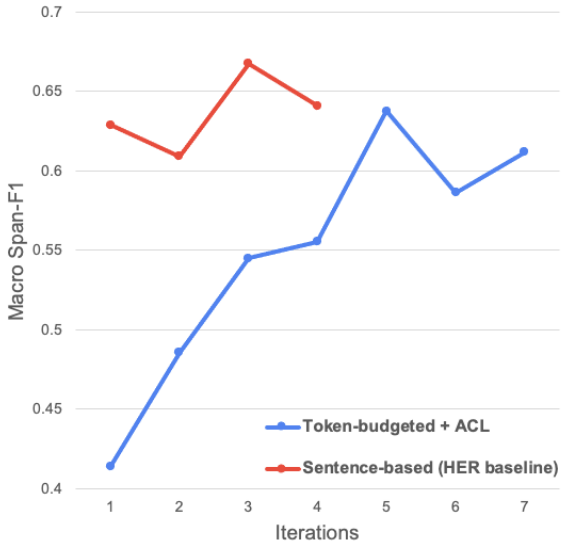


Figure 2: Macro span-F1 across iterations for sentence-based HER and token-budgeted HER with ACL-initialised seed selection. Each point is computed by three-fold cross-validation on the labelled seed available at that iteration.

Iteration	Seed Tokens	Macro span-F1
1	1,996	0.4137
2	4,047	0.4856
3	6,084	0.5449
4	8,231	0.5555
5	10,254	0.6377
6	12,274	0.5864
7	14,365	0.6119

Table 3: Results of token-budgeted pre-annotation with Active Curriculum Learning.

7.3. Annotation Complexity Analysis

To investigate whether token-budgeted sampling implicitly reduces annotation complexity by favouring simpler samples, we analyse the distribution of nested entity depths across annotation strategies. We count the number of annotated entity spans involving single entities, two nested entities, and three nested entities within each annotation batch.

Table 4 reports the average number of annotations by nesting depth across runs for sentence-based sampling and token-budgeted sampling with ACL. While sentence-based sampling results in substantially larger and more variable annotation batches in terms of token count, the relative distribution of nested entity depths is comparable across strategies. These findings suggest that controlling annotation batch size at the token level improves predictability of annotation workload without reducing semantic complexity.

Strategy	Single	2-Nested	3-Nested	Tokens / Iter.
Sentence-based	870 (8.46%)	685.8 (6.67%)	152.8 (1.49%)	10,280.5
Token-based	197.3 (9.78%)	181.1 (8.97%)	35.6 (1.76%)	2,018

Table 4: Average extracted entities and annotation cost per iteration.

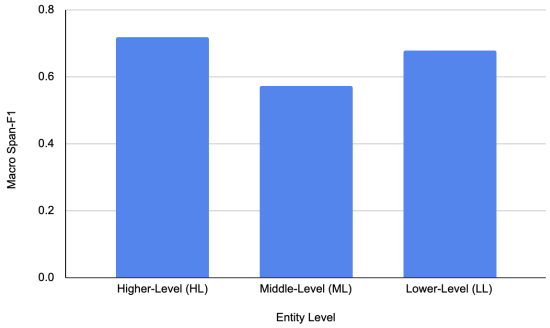


Figure 3: Macro span-F1 by Entity Level (Staggered Annotation)

7.4. Staggered Annotation By Entity Level

To analyse the impact of semantic granularity and nesting complexity, additional experiments were conducted in which models were trained separately on Higher-Level, Middle-Level, and Lower-Level entities. As shown in figure 3, higher-level entities achieve the highest and most stable performance, benefiting from higher frequency and clearer context. Middle-level entities show moderate performance, reflecting the difficulty of distinguishing legal roles such as parties and witnesses. Lower-level entities show some performance relative to middle-level entities, but remain challenging due to their fine-grained semantics, frequent abbreviations, and dependence on nested structures. Performance decreases with increasing semantic granularity, highlighting the challenges posed by fine-grained nested entities in historical notarial documents.

7.5. Entity-Level Error Analysis

The staggered annotation experiments provide the clearest view of the classification of labels by the model. We perform an entity-level error analysis using aggregated confusion matrices across iterations and summarise the distribution of error types in Figure 4. The figure shows that the dominant error mode at all levels is missed entity tokens, indicating that the remaining errors are primarily recall-driven. At the Higher-level, the most frequently misclassified labels are KEY and PER, which are most often missed as O. 10.2% are labels that have their BIO boundary flipped. Errors in rarer categories (e.g., INS and DEEDT) remain consistent with data sparsity. Middle-level entities, PERP

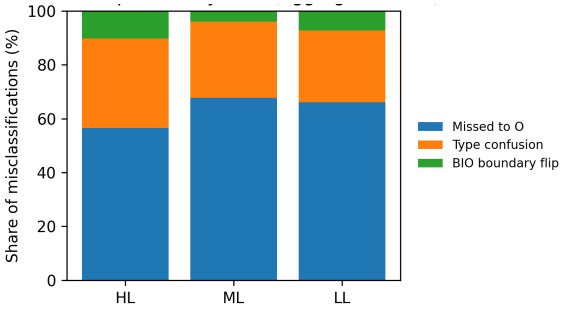


Figure 4: Error composition by level aggregated across iterations

and PERW, consistently obtain the lowest F1 scores and are erroneously classified as O. These entities encode legal roles that are not lexically marked but inferred from their position in the deeds and their formulaic context, making them difficult to capture using local sequence features alone. At the Lower-level, fine-grained components are again predominantly missed (e.g., PERFN−B→O = 522, PERLN−B→O = 336, GEOB−I→O = 201), suggesting that these classes require more annotations and are sensitive to boundary variation. In addition, Lower-level exhibits clear semantic-adjacency confusions, most notably between closely related date subtypes (DATEANIN−I→DATEIN−I = 105 and DATEIN−I→DATEANIN−I = 49). Overall, the staggered setting confirms that increased semantic granularity amplifies recall limitations and subtype ambiguity, with ML role labels (PERP/PERW) emerging as the most consistently misclassified and most likely to benefit from additional targeted annotation and document-level cues.

8. Conclusion

This paper investigated cost-aware pre-annotation strategies for nested NER in seventeenth-century Latin notarial deeds from the Maltese NRA. Extending the HER to support nested entities, we compared sentence-based pre-annotation as a baseline, token-budgeted pre-annotation with ACL, and staggered annotation by entity level. The results show that sentence-based sampling leads to unstable learning behaviour under highly skewed sentence-length distributions, whereas token-budgeted sampling provides more predictable annotation cost and more stable gains in span-level F1.

Staggered annotation experiments reveal differences across semantic levels: Higher-level entities are recognised with the highest accuracy, middle-level role-based entities remain challenging due to their dependence on discourse structure, and lower-level structural entities benefit from regular

patterns and gazetteer support despite nesting. An entity-level error analysis indicates that remaining errors are driven primarily by semantic granularity and document-level dependencies rather than boundary detection failures. Overall, the findings demonstrate that cost-aware pre-annotation can produce high-quality silver-standard annotations and support more efficient annotation planning, laying foundations for scalable knowledge extraction workflows for complex historical archival corpora.

Future work will extend token-budgeted sampling with additional iterations by adding unannotated deeds from a different notary and a different decade of the seventeenth century, testing whether gains in macro span-F1 persist under temporal and scribal shift. We will furthermore track dataset provenance during ranking to quantify which sources are sampled across iterations and how this affects learning stability and label performance. In addition, future work will also investigate encoder-based transformer models for Latin, such as LatinBERT (Bamman and Burns, 2020), to evaluate whether contextual representations can improve nested entity recognition beyond the lightweight CRF-based pre-annotation setting adopted here.

9. Acknowledgments

This research is contributing to the NotaryPedia project, a collaboration between the Notarial Archives Foundation and the Department of Artificial Intelligence at the University of Malta, funded by the Ministry for the National Heritage, the Arts and Local Government. The project aims to bridge traditional archival standards with AI-assisted content analysis, demonstrating how archival collections can be transformed into dynamic, interconnected knowledge infrastructures.

10. Bibliographical References

Sergio Torres Aguilar and Dominique Stutzmann. 2021. Named Entity Recognition for French medieval charters. In *NLP4DH*.

David Bamman and Patrick J. Burns. 2020. Latin BERT: A Contextual Language Model for Classical Philology.

Felix Bensmann, Benjamin Zapilko, and Philipp Mayr. 2017. Interlinking Large-scale Library Data with Authority Records. *Frontiers in Digital Humanities*, Volume 4 - 2017.

Christian Bentz, Douwe Kiela, Felix Hill, and Paula Buttery. 2014. Zipf's law and the grammar of languages: A quantitative study of Old and Modern English parallel texts. *Corpus Linguistics and Linguistic Theory*, 10:175 – 211.

Emanuela Boroş, Verónica Romero, Martin Maarand, Kateřina Zenklová, Jitka Křečková, Enrique Vidal, Dominique Stutzmann, and Christopher Kermorvant. 2020. A comparison of sequential and combined approaches for named entity recognition in a corpus of handwritten medieval charters. In *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 79–84.

Emanuel Buttigieg and Joan Abela. 2020. NAV: a survey of the past, present, and future of the Notarial Archives of Valletta, Malta. *Nuovi Annali della Scuola Speciale per Archivisti e Bibliotecari*, 34:5–26.

Pierre Chastang, Xavier Tannier, and Sergio Aguilar. 2021. A Named Entity Recognition Model for Medieval Latin Charters. *Digital Humanities Quarterly*, 15.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named Entity Recognition and Classification in Historical Documents: A Survey. *ACM Computing Surveys*, 56(2):1–47.

Charlene Ellul, Joel Azzopardi, and Charlie Abela. 2019. NotaryPedia: A Knowledge Graph of Historical Notarial Manuscripts. In *On the Move to Meaningful Internet Systems: OTM 2019 Conferences*, pages 626–645, Cham. Springer International Publishing.

Alexander Erdmann, David Joseph Wrisley, Benjamin Allen, Christopher Brown, Sophie Cohen-Bodénès, Micha Elsner, Yukun Feng, Brian Joseph, Béatrice Joyeux-Prunel, and Marie-Catherine de Marneffe. 2019. Practical, Efficient, and Customizable Active Learning for Named Entity Recognition in the Digital Humanities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2223–2234, Minneapolis, Minnesota. Association for Computational Linguistics.

Jenny Rose Finkel and Christopher D. Manning. 2009. Nested Named Entity Recognition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 141–150, Singapore. Association for Computational Linguistics.

S. Fiorini. 1996. *Documentary Sources of Maltese History Series*. University of Malta Press, Malta.

Joseph Fisher and Andreas Vlachos. 2019. Merge and Label: A Novel Neural Network Architecture for Nested NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5840–5850, Florence, Italy. Association for Computational Linguistics.

Borna Jafarpour, Dawn Sepehr, and Nick Pogrebnyakov. 2021. Active Curriculum Learning. In *Proceedings of the First Workshop on Interactive Learning for Natural Language Processing*, pages 40–45, Online. Association for Computational Linguistics.

Imed Keraghel, Stanislas Morbieu, and Mohamed Nadif. 2024. Recent Advances in Named Entity Recognition: A Comprehensive Survey and Comparative Study.

Inês Koch, Carla Teixeira Lopes, and Cristina Ribeiro. 2023. Moving from ISAD(G) to a CIDOC CRM-based Linked Data Model in the Portuguese Archives. *J. Comput. Cult. Herit.*, 16(4).

Raphaëlle Lapôtre. 2017. Library Metadata on the web: the example of data.bnf.fr. *JLIS.it*, 8(3):58–70.

Edward Loss, Fabiana Guernaccini, and Manuel Carassai. 2025. From Manuscript to Metadata: experiments on Handwritten Text Recognition, Tagging and Importation for the Memoriali series (1265-1452). *JLIS.it*, 16(2):59–85.

Wei Lu and Dan Roth. 2015. Joint Mention Extraction and Classification with Mention Hypergraphs. In *Conference on Empirical Methods in Natural Language Processing*.

Vit Novotny, Kristina Luger, Michal Štefánik, Tereza Vrabcova, and Ales Horak. 2023. People and Places of Historical Europe: Bootstrapping Annotation Pipeline and a New Corpus of Named Entities in Late Medieval Texts. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14104–14113, Toronto, Canada. Association for Computational Linguistics.

Naoaki Okazaki. 2007. CRFsuite: a fast implementation of Conditional Random Fields (CRFs).

Burr Settles. 2009. Active Learning Literature Survey. Technical Report TR1648, University of Wisconsin–Madison Department of Computer Sciences.

Burr Settles, Mark W. Craven, and Lewis A. Friedland. 2008. Active Learning with Real Annotation Costs.

Yanyao Shen, Hyokun Yun, Zachary Lipton, Yakov Kronrod, and Animashree Anandkumar. 2017. Deep Active Learning for Named Entity Recognition. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 252–256, Vancouver, Canada. Association for Computational Linguistics.

Takashi Shibuya and Eduard Hovy. 2020. Nested Named Entity Recognition via Second-best Sequence Learning and Decoding. *Transactions of the Association for Computational Linguistics*, 8:605–620.

Jana Straková, Milan Straka, and Jan Hajic. 2019. Neural Architectures for Nested NER through Linearization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5326–5331, Florence, Italy. Association for Computational Linguistics.

Sergio Octavio Torres Aguilar. 2022. Multilingual Named Entity Recognition for Medieval Charters using Stacked Embeddings and BERT-based Models. In *Second Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA, 2022)*.

G. Wettinger. 2000. *Place-Names of the Maltese Islands, ca. 1300–1800*. Publishers Enterprises Group, Malta.

Appendix

Per-Entity Evaluation Metrics for Main Pre-Annotation Strategies

The following tables report pooled per-entity precision, recall, and F1 scores across the three cross-validation folds for the two main pre-annotation strategies evaluated in this work: the sentence-based HER baseline and the token-budgeted approach with Active Curriculum Learning. These detailed metrics are included for completeness and reproducibility. Results for the HL, ML, and LL staggered models are summarised in Figure 3 in the main text, as these experiments serve as auxiliary analysis rather than primary comparisons.

Entity	Precision	Recall	F1	Support
DATE	0.825	0.786	0.805	84
DATEAN	0.000	0.000	0.000	1
DATEANIN	0.778	0.636	0.700	11
DATEIN	0.779	0.767	0.773	60
DEEDT	0.811	0.577	0.674	52
GEO	0.837	0.788	0.812	421
GEOB	0.274	0.103	0.150	29
GEOPC	0.655	0.776	0.710	49
INS	0.357	0.435	0.392	23
KEY	0.827	0.767	0.796	1676
PER	0.874	0.815	0.843	1224
PERFN	0.923	0.870	0.896	1123
PERLN	0.956	0.880	0.916	785
PERP	0.552	0.408	0.469	103
PERW	0.474	0.394	0.430	137

Table 5: Pooled precision, recall, and F1 across the three folds for the best sentence-based HER baseline iteration (Iteration 3 with macro span-F1 = 0.6675).

Entity	Precision	Recall	F1	Support
DATE	0.781	0.735	0.758	34
DATEAN	0.000	0.000	0.000	1
DATEANIN	0.500	0.625	0.556	8
DATEIN	0.500	0.625	0.555	16
DEEDT	1.000	0.400	0.571	5
GEO	0.864	0.827	0.845	185
GEOB	0.889	0.533	0.667	15
GEOPC	0.586	0.629	0.607	27
INS	0.667	0.222	0.333	9
KEY	0.833	0.682	0.750	547
PER	0.830	0.729	0.776	483
PERFN	0.896	0.798	0.844	445
PERLN	0.957	0.805	0.874	328
PERP	0.000	0.000	0.000	15
PERW	0.547	0.460	0.500	76

Table 6: Pooled precision, recall, and F1 across the three folds for the best token-budgeted + ACL iteration (Iteration 5 with macro span-F1 = 0.6377).

From Lemmatization to Legal Terminology: Assessing an Hybrid Pipeline on Justinian's *Digest*

Paola Marongiu, Eva Sassolini
Istituto di Linguistica Computazionale "A. Zampolli", Consiglio Nazionale delle Ricerche (CNR-ILC)
Via Giuseppe Moruzzi 1, Pisa
{paola.marongiu, eva.sassolini}@ilc.cnr.it

Abstract

This paper evaluates a hybrid NLP pipeline for supporting the extraction of Roman legal terminology from Justinian's Digest. Our goal is not to optimize lemmatization in isolation, but to assess whether integrating a Large Language Model (GPT-4o-mini) as a post-processing component improves lemma quality in ways that are critical for downstream glossary construction. Using LatinPipe as a baseline (F1 = 95.05), we test the integration of GPT-4o-mini under three experimental settings (zero-shot with and without prior lemma information, and few-shot prompting) against a manually annotated gold standard of 3,703 sentences and an expert-validated list of legal Latin technical terms. Results show improvement across all settings, with the best performance achieved in the few-shot configuration. Our analysis shows that the hybrid configuration produces selective improvements, significantly more likely for frequent lemmas and verbs forms, suggesting that the LLM layer primarily assists in resolving morphologically ambiguous inflected forms. Although our experimental conditions may not hold in real-world scenarios, we argue that the main contribution of this work is methodological: demonstrating how evaluation can be aligned with downstream terminological goals, rather than proposing a general-purpose solution to domain-specific lemmatization.

Keywords: Latin, lemmatization, LLM, GPT, Digest, legal text

1. Introduction

Lemmatization is a fundamental task in Natural Language Processing (NLP), which consists in assigning a word form to its canonical citation form, called lemma. Having reliable lemmatization results is crucial for performing more advanced NLP tasks, such as syntactic and semantic parsing, terminology extraction, and in general for building language resources and carrying out different types of linguistic analyses. This paper focuses in particular on the preparatory steps to the construction of a digital bilingual glossary extracted from the parallel corpus of the Justinian's *Digest*. This goal is framed as part of a broader project aimed at producing an Italian translation of the *Digest*, and extracting linguistic resources that can work as translation memories for the translators and, at the same time, provide new data for research in historical linguistics, among others. The glossary will be built using Latin as the pivot language for the term extraction. For this reason, it is essential that the lemmatization of the Latin corpus be of exceptionally high quality, especially concerning Latin legal terms.

The Latin Digest of Emperor Justinian represents a particularly demanding corpus for NLP. Compiled between 529 and 534 CE as part of the Corpus Iuris Civilis (Banchich et al., 2015; Dingledy, 2016), the Digest gathers excerpts from 1,528 writings of 39 Roman jurists, covering roughly eight centuries of legal thought—from the Twelve Tables (449 BCE) to late imperial jurists of the fourth century CE (Dingledy, 2016; Ribary and McGillivray, 2020). The text is divided into fifty books, each subdivided into titles, laws, and paragraphs, and introduced by inscriptions that identify the author and source of each excerpt (Banchich et al., 2015). As the most extensive collection of Roman legal doctrine, the Digest is invaluable not only for legal historians but also for scholars interested in Latin lexicon and morphology within a highly specialized domain. The legal register of the Digest displays terminological density, recurrent formulae, and syntactic constructions that diverge from classical Latin (CL) norms, all of which pose significant challenges for computational processing.

From a linguistic and philological perspective, developing accurate lemmatization for this corpus is a prerequisite for building comprehensive lexical resources, such as domain-specific glossaries or bilingual lexicons of legal Latin. However, manual lemmatization of such an extensive and complex corpus would be prohibitively time-consuming. Automatic or semi-automatic solutions are therefore indispensable for efficiently supporting expert validation and downstream tasks. The nature of the Digest as a Latin domain text brings specific challenges to the task. Latin is a historical language, usually defined as a low-resource

Proceedings of LT4HALA 2026 @ LREC 2026, pages 418–428.
11 May 2026. *©ELRA Language Resources Association (ELRA), 2026*

language,[1] morphologically rich, subject to orthographic variation. Moreover, specialized corpora have an exceptional presence of domain-specific vocabulary, including lemmas that traditional NLP tools might not have seen as much during training.

Building upon recent advances in the field of Artificial Intelligence, this work explores strategies to enhance automatic lemmatization for specialized Latin corpora. By testing different approaches and experimental settings, and validating results against manually annotated samples, we aim to assess the combination of state-of-the-art NLP tools and Large Language Models (LLMs), to apply them to historical and domain-bound texts.

2. Related work

Lemmatization of Latin has been explored through multiple paradigms, ranging from rule- and dictionary-based lexical analyzers to modern neural pipelines. Most recently, the three EvaLatin editions in 2020 (Sprugnoli et al., 2020), 2022 (Sprugnoli et al., 2022) and 2024 (Sprugnoli et al., 2024) have promoted the development of computational resources and tools for Latin NLP. EvaLatin 2020 was specifically devoted to lemmatization and PoS-tagging. The task was divided in three different subtasks, one for each dataset: CL, cross-genre (prose vs. poetry) and cross-time (authors from the 1st century BC to the 1st century CE vs. Thomas Aquinas, an author from the 13th century CE). The best tool across all sub-task was UDPipe (Straka and Straková, 2020), which scored from F1 87.69 (cross-time) to 96.74 (Classical Latin). An evolution of UDPipe, LatinPipe (Straka et al., 2024), was presented to the EvaLatin 2024 campaign, which focused on morphosyntactic analysis and sentiment analysis in Latin. LatinPipe is a graph-based dependency parser born as the evolution of its predecessors UDPipe (Straka and Straková, 2017) and UDPipe 2 (Straka, 2018). This parser leverages pre-trained language models, fine-tuned on five Latin treebanks publicly available in Universal Dependencies (Nivre et al., 2020), version 2.13, and two texts annotated in UD style (see (Straka et al., 2024) for further details):

- ITTB (Passarotti, 2019)
- LLCT (Cecchini et al., 2020a)
- PROIEL in UD 2.13 (Haug and Jøhndal, 2008) and in the UD-style harmonized version (Gamba and Zeman, 2023)
- UDante (Cecchini et al., 2020b)
- Perseus (Bamman and Crane, 2011)
- *De Latinae Linguae Reparatione* by Marcus Antonius Sabellicus (Gamba and Cecchini, 2024)
- the Latin translation of *The Spirals* of Archimedes by Jacopo da San Cassiano, part of the *Archimedes Latinus* UD-style treebank (Fantoli and de Lhoneux, 2022)

Two variants of LatinPipe ranked first and second in both EvaLatin 2024 shared tasks, demonstrating its effectiveness in Latin dependency parsing. The performance of the tools participating in the task was tested against two authors, one for prose (Tacitus' *Germania*), and one for poetry (Seneca's *Hercules Furens*).

The potential of LLMs has been tested on diverse labeling tasks, e.g. event detection and extraction (Chen et al., 2024), discourse functions in spoken corpora (Petukhova and Kochmar, 2025), caused-motion construction (Weissweiler et al., 2025). Most research has focused on modern languages, but a few studies have explored applications on historical languages too. LLMs have been tested on Named Entity Recognition in Historical Urdu (Irfan and Ali, 2025); (Volk et al., 2024) have tested GPT-4 in translation and text summarisation tasks on Late Latin texts; (Farina et al., 2025) tested open and closed-weight models in the detection of semantic features such as spatial relations for Latin motion verbs.

To our best knowledge, little to no research has been specifically devoted to assessing the performance of LLMs on lemmatization, especially for historical languages. Toporkov et al. (2025) recently tested different LLMs in the lemmatization of a selection of modern languages with different levels of morphological complexity: English, Spanish, French, German, Italian, Finnish, Icelandic, Turkish, Swedish, among others. In their work, they compared encoder-based models fine-tuned on language-specific gold data with LLMs in in-context lemmatization for multiple languages, demonstrating that while fine-tuned models still achieve superior performances, LLMs can yield competitive results through few-shot prompting without prior fine-tuning. To our knowledge, the only work on lemmatization involving LLMs in an historical language is Riemenschneider (2025), which focuses on Akkadian. At the beginning of the lemmatization pipeline, three different systems are combined and asked to provide their own lemma for each form in the text. When they do not agree on the predicted lemma, the LLM is provided with the three predictions and the passage in which the word form occurs, and is asked to provide the correct lemma.

[1] Though it should be noted that among historical languages Latin is certainly the luckiest one in terms of resources for NLP.

Recent advances in the field have expanded the possibilities offered by deploying LLMs as standalone systems, by integrating them with other computational tools and architectures to form hybrid pipelines. For instance, recent work on Latin word sense prediction combines Large Language Models with Linguistic Knowledge Graphs (LKGs) in a Graph Retrieval-Augmented Generation framework, leveraging structured semantic and contextual information from multiple resources (including Latin WordNet and Wikidata) to support word sense disambiguation. Results show that while LLMs achieve competitive performance on their own, the integration of graph-based metadata can yield measurable improvements, particularly for larger models, highlighting the potential of structured knowledge to guide LLM reasoning in low-resource historical settings (Ghizzota et al., 2026). Similarly, hybrid architectures have been successfully applied to morphologically complex and low-resource languages, such as in Jungar Tuvan, where applying an LLM at the end of the pipeline significantly improves automatic glossing over the neural baseline (Liang et al., 2026). These approaches demonstrate that integrating structured or model-based linguistic processing with LLMs can enhance performance, especially in contexts characterized by morphological complexity and limited annotated data.

As described in this section, a lot of work has been done in Latin NLP, and a few initiatives have encouraged progress in this field. In parallel, the fast evolving landascape of NLP and AI offers new means of approaching old issues. More specifically to the focus of our paper, lemmatization of highly specialized texts e.g. legal documents in our case, remains under-explored, as well as experiments on the combination of state-of-the-art tools with the most recent advances in AI in specialized lemmatization tasks. In this paper, we explore this direction with the practical goal in mind of achieving high lemmatization performances to facilitate the development of our digital glossary of legal Latin.

3. Data and methodology

3.1. Latin data

The legal scholars working on the translation of the *Digest* use as a reference edition for the Latin text the *editio stereotypa duodecima* edited by Mommsen and Krüger (1911), widely adopted by the research community dealing with Justinian law and its translations (Schipani, 2005, 555–557). Therefore, the text in our corpus is largely based on this version of the Digest. However, as they proceed with the translation, they sometimes modify the text of the reference edition, for various reasons, explaining the implemented changes in a footnote in the published volume. For this reason, although the text of the *Digest* is publicly available online, we process the files as we receive them from the translators.

As the translation process is still ongoing, we worked with the books that had already been translated at the time of the experiments. Therefore, we processed 36 of the 50 books in the Digest. The corpus as of now thus contains around 635k tokens, with around 37,000 sentences.

Before automatic annotation, the texts were pre-processed to lowercase characters, separate punctuation, and remove non-standard characters e.g. diacritics. Numerical indications for each paragraph were also removed, and then re-assigned in the CoNLL-U file during post-processing. Lemmatizing the Digest implies various challenges, due to the following reasons. The first reason has already been addressed in the first two sections, namely an extremely specialised lexicon which might not have been seen by the models during training. Most of the annotated data publicly available for Latin are CL or Late Latin texts, poetry or prose, but not specialised or domain-bound texts such as legal or medical literature. The second issue is that although the Digest is mostly written in Latin, it also contains Ancient Greek (AG) script, which in some cases is very extensive, e.g. the 27th book is almost entirely written in AG. Finally, the text presents some *lacunae*, due to the fact that it was not transmitted to us in its entirety. Therefore, sometimes the text contains some gaps, interrupted sentences and similar textual and philological issues which might represent a challenge for automatic annotation.

Group	Thematic scope	Books
1	General principles	I–IV
2	Property and real rights	V–XI
3	Obligations and contracts	XII–XIX
4	Family law	XX–XXVII
5	Testate succession, possession, praetorian succession, and criminal law	XXVIII–L

Table 1: Thematic groups in the Digest corpus.

3.2. Gold Standard

As mentioned in section 3, the corpus comprises a total of around 37,000 sentences, from which we built a gold standard set consisting of 3,703 sentences (10% of the total). In order to ensure both

thematic and quantitative balance, the sentences were sampled proportionally across the five major thematic groups into which the Digest is traditionally divided (see Table 1). As we only have access to books 1–36 for now, we based the extraction on groups 1 to 5. The number of sentences extracted from each group was determined according to their relative size in the corpus. Formally, if S_n denotes the number of sentences in group n and S the total number of sentences in the corpus, then the number of sentences selected from group n is given by $0.1 \times S_n$. This guarantees that each thematic group contributes to the golden dataset in direct proportion to its representation in the corpus, while keeping the total gold size fixed at 3,703 sentences. The resulting distribution is shown in Table 2.

Group	N° of sentences (S_n)	Gold size (10%)
1	5,410	541
2	5,990	599
3	6,903	690
4	7,295	729
5	11,455	1,145

Table 2: Distribution of corpus and gold sentences by thematic group.

After the pre-processing phase, the sampled sentences were lemmatised and tagged using LatinPipe. This tool was selected among others due to its recent success in the latest EvaLatin campaign, described in Section 2. We manually corrected the output of LatinPipe, focusing on lemmas and PoS, to obtain our gold standard. Our gold standard dataset, as well as the code to run our experiments, is freely available on Github (see section 8).

3.3. List of legal terms in the *Digest*

In order to start building our digital glossary, we prepared a list of candidate lemmas to be validated by our team of Roman legal scholars. In order to build the list of candidate terms, we compared the *Digest* corpus with a general-interest corpus, developed by the LASLA laboratory in Li'ege (Laboratoire d'Analyse Statistique des Langues Anciennes)[2] and recently linked to the LiLa knowledge base (Fantoli et al., 2022, 2024). Specifically, for the initial word list we combined information about i) unique lemmas in the *Digest* ii) terms that appeared more frequently in the *Digest*, and iii) measured the semantic variation of the shared vocabulary between the *Digest* and the LASLA corpus. It should be noted that the comparing these two corpora does not come without risks: the LASLA corpus only contains texts from the Classical period, whereas the *Digest* dates back to the fourth century CE. However, we opted for this corpus instead of other resources for the excellent quality of its lemmatization, and ii) because the manual validation by the domain experts would anyway exclude terms emerging from the analysis because of the absence of temporal overlap. The final list, validated by the experts, contains 583 terms belonging to the legal domain, comprising 287 verbs, 264 nouns, and 32 adjectives.[3]

We first analysed the distribution of the technical terms in our gold dataset as a reference, to determine their frequency across the 3703 sentences. This step was necessary for performing in-depth evaluation of the performances of our hybrid pipeline on the technical terms specifically. In order to obtain this, we first counted both the number of occurrences for each term in the dataset, and the number of sentences in which each term occurs (noting that more than one term can occur in the same sentence). Then, we divided the list of terms into three ranges, depending on their distribution in our gold dataset: 'frequent', 'medium', 'rare'. The last category is 'absent', for technical terms that do not appear in our gold standard and therefore cannot be used for evaluation. To determine the three frequency ranges we adopted a data-driven approach: we calculated empirical quantiles based on the distribution of the technical lemmas in the corpus. Specifically, we computerd the 33rd and 67th percentiles over lemmas with at least one occurrence, and used these values as data-driven thresholds to build three balanced frequency ranges: rare lemmas have ≤ 10 occurrences, medium lemmas have 11–28 occurrences, and frequent lemmas have > 28 occurrences.

This decision was based on the fact that most of the technical legal terms in our list are rare in the gold dataset, and the very frequent ones are only a few. The highly skewed distribution of technical lemmas is shown in the barplot in Figure 1. Figure 2 shows the same distribution but cut up to 200 occurrences, for better readability of the first two ranges. Having long-tail distributions of low-frequency words with few high-frequency words is widely attested in natural language texts (Baroni, 2009, 810);(Zipf, 1949). In our specialized corpus, we observe a similar behaviour for our list of technical words.

[2] https://www.lasla.uliege.be/cms/c_8508894/fr/lasla

[3] It should be noted that the list only contains single lemmas. The list of candidate multi-word expressions is currently being evaluated by the domain experts.

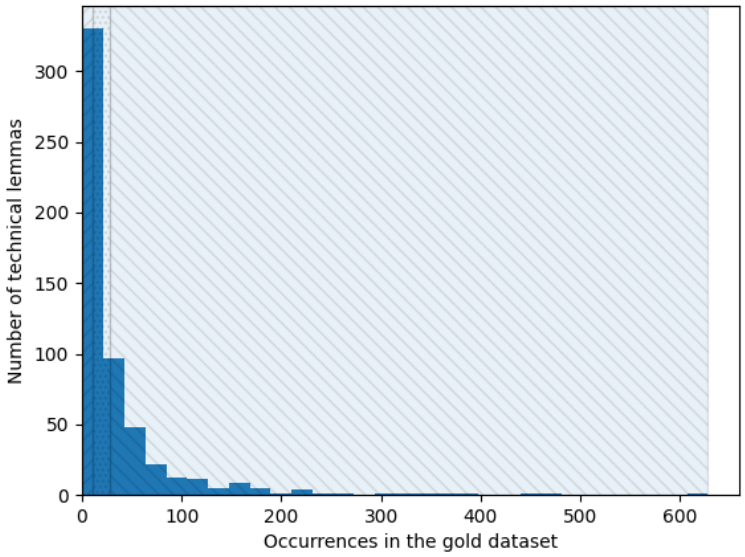


Figure 1: Frequency distribution of validated technical lemmas. Shaded regions represent quantile-based frequency bands: rare (≤10 occurrences), medium (11–28), and frequent (>28).

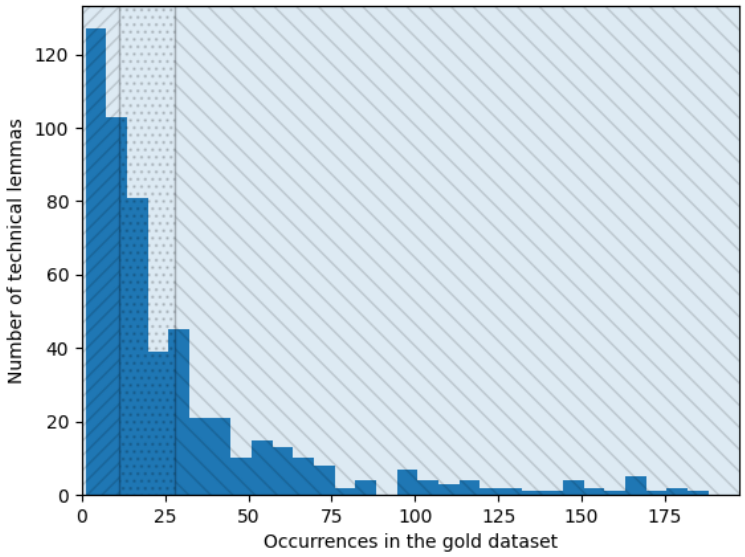


Figure 2: Frequency distribution of validated technical lemmas up to 200 occurrences.

4. Models and methodology

As mentioned in the previous sections, we obtained our baseline with LatinPipe.

The tool has a F1 score of 95.05 on the lemmatization of the Digest, which certainly leaves room for improvement. On a closer qualitative analysis of the results, we noticed that the model consistently missed lemmas like *pignus* 'pledge, security', often lemmatizing it as *pignor* and similar forms. This represents a strong limitation to the use of this tool to us, because technical words in the legal domain like *pignus* cannot be missed when pursuing the goal of extracting a specialised glossary from the corpus. Therefore, we combined the output of LatinPipe with GPT-4o-mini, to see if and how the baseline increases, and on which lemmas.

First, we compare our gold with the output of LatinPipe to spot wrongly predicted lemmas. We pass them over to GPT, with different prompt instruc-

tions (see following sections). Then, we replace the wrong lemmas in the CoNLL-U file produced with LatinPipe with the output of GPT, and perform the evaluation against our gold standard again.

4.1. Experimental settings

We tested GPT-4o-mini against the baseline obtained with LatinPipe in three different experimental settings.

Zero-shot 1. We provide the lemma as predicted by LatinPipe, and ask the model i) to evaluate if the predicted lemma is correct or not, and ii) if it is not correct, to provide the correct lemma. Below is the prompt used in this first experimental setting.

```
You are a Latin expert. I
will give you a Latin word
and the context sentence. The
lemma for this Latin word as
predicted by LatinPipe is:
'predicted lemma'
Instructions:

  • Check if the predicted
    lemma is correct
  • If it is correct give the
    same lemma in output
  • If it is wrong, correct it
  • Give me back ONLY THE
    CORRECT LEMMA, as a single
    Latin word
  • No additional comments or
    explanations

Word:
Context:
Output:
```

Zero-shot 2. We give the model the same prompt given in zero-shot 1, but we do not provide the lemma predicted by LatinPipe. This allows us to understand if providing the predicted lemma helps or hinders the accuracy of the prediction made by GPT.

Few-shot. We provide the model with a few examples of correct lemmatization. The shots consist of 30 randomly extracted sentences from the gold dataset. We give the model the word form, the associated gold lemma, and the sentence in which each word form appears. Then, we ask the model to repeat the task as formulated in zero-shot 2 i.e., without knowing the lemma predicted by LatinPipe on the remaining sentences.

Model	Setting	Precision	Recall	F1
LatinPipe (LP)		95.21	94.92	95.06
LP+GPT-4o-mini	Zero-shot 1	95.84	95.55	95.70
	Zero-shot 2	96.43	96.13	96.28
	Few-shot	96.71	96.41	**96.56**

Table 3: Results on the lemmatization task.

5. Results and discussion

Our baseline accuracy for the full sample of sentences in the dataset used for the evaluation is 95.05 (F1 score). Across all experimental settings, the combination of GPT-4o-mini with the LatinPipe output leads to an improvement of the lemmatization accuracy. Results are shown in Table 3.

When GPT is not given the lemma predicted by the system (zero-shot 2), its F1 score increases compared to when the predicted lemma is given (zero-shot 1). It seems that GPT is biased by the information given in the prompt in zero-shot 1, and performs better when it is left to freely assign the lemma to the word form.

The few shot experiment is provided with a new baseline value (F1 = 95.06), as we had to subtract from the annotation dataset the sentences selected as shots for the LLM. In this third experimental setting, GPT benefits from in-context few-shot examples and achieves its best performance (F1 = 96.49), improving the baseline by nearly 1.5 points.

On a closer qualitative analysis, we notice that the accuracy on problematic lemmas increases with the hybrid pipeline. For instance, *pignus* scores 37.3 with LatinPipe, and 88.1 with GPT (few-shot). As mentioned above, the Digest contains some AG text. LatinPipe recognizes AG script, and uses the label 'greek.expression' to fill the lemma field. GPT-4o-mini on the other hand, also recognizes AG script, but lemmatizes it as a Latin word. In particular, the conjunction καὶ 'and' is lemmatized by GPT-4o-mini as *et*, i.e. the corresponding conjunction in Latin.

5.1. In-depth analysis on technical legal terms

To better understand the actual impact of the hybrid pipeline on domain-relevant lexical material, we conducted a focused analysis on the subset of expert-validated technical lemmas. Rather than relying solely on global F1 improvements, we examined lemma-level behavior across frequency bands and part-of-speech categories, identifying where and how the hybrid configuration (LatinPipe + GPT) modifies the baseline system.

The large majority of technical lemmas (82.9%) remain stable, while 16.4% improve and only 0.7%

worsen. This confirms that the LLM layer acts as a conservative corrective component, selectively refining specific lemmas without introducing systematic regressions. The performance on lemmatization decreases for only four lemmas, suggesting that the LLM layer rarely overrides correct predictions made by LatinPipe on technical terms. In order to understand where the hybrid pipeline performs better, we conducted separate analyses on the technical terms based on their frequency range and their PoS.

Frequency range. The distribution of improvements across frequency ranges reveals a non-uniform pattern. The performance on frequent lemmas improves by 25.1%; 14.7% for medium-frequency lemmas; 6% for rare lemmas. Thus, frequent technical lemmas are approximately four times more likely to improve than rare ones. This suggests that the hybrid pipeline primarily intervenes where cumulative error impact is higher. Rather than rescuing marginal or extremely rare items, the LLM appears to stabilize lemmas that occur repeatedly in the corpus and therefore have greater influence on terminology extraction and distributional statistics. This distributional effect is particularly relevant for glossary construction, where high-frequency terms shape candidate ranking, collocational profiles, and bilingual alignment.

Part-of-speech. A clearer structural pattern emerges when considering part-of-speech categories. Lemmatization of verbs improves by 20.8%; on adjectives by 12.9%; on nouns by 9.7%.

Verbs are therefore more than twice as likely to improve compared to nouns. This indicates that the integration of GPT primarily contributes to resolving morphologically complex verbal forms rather than stabilizing already well-identified technical nouns.

Table 4 shows the results of our analysis by combining PoS and frequency range information. Frequent verbs show an improvement rate of 37.8%, compared to only 5.0% for rare verbs. In contrast, nouns remain comparatively stable across different frequency ranges, with improvement rates ranging between 6.7% and 13.2%.

This pattern suggests that LatinPipe already performs robustly on domain-specific nouns, while the LLM layer contributes primarily in contexts involving verbal inflectional ambiguity.

Examples from the list of technical terms. Table 5 presents representative examples of lemmas with the highest improvements. Among verbs, notable cases include *parco* ('to spare, have mercy upon'), improved from 0.50 to 1.00; *dirimo* ('to settle, dissolve'), improved from 0.50 to 0.83; *infero* ('to bring against, to charge'), improved from 0.56 to 0.88. All these verbs are central to the *Digest*, which describes various circumstances in which the defendant or a possible defendant may be charged, tried, convicted, or spared. These improvements likely reflect enhanced contextual disambiguation of inflected verbal forms.

The case of *creditor* ('creditor'), which improves from 0.55 to 0.91 across 102 occurrences, is particularly significant due to its high frequency in the corpus and central role in Roman legal discourse. Although morphologically ambiguous in some contexts, its improved consistency directly benefits glossary-oriented analyses. However, among nouns, the most emblematic example is *pignus*, improved from 0.37 to 0.88. The Digest deals extensively with how debts can be incurred, loans granted, debts discharged, and similar circumstances, and this substantial gain seems to suggest the hybrid system's ability to recover highly relevant technical terminology.

Adjectival improvements are more moderate, *praesens* ('present'), from 0.75 to 0.94, *ratus* ('ratified, valid'), from 0.77 to 0.92. Overall, adjectival lemmas appear more stable than verbal ones, and improvements tend to be smaller. It should also be noted that these are in fact adjectival forms derived from verbs (*praesum* and *reor*), and therefore considering them adjectives is entirely debatable. The assignment of this POS is due to the fact that the legal scholars specifically required the glossary to have a specific entry for these adjectival forms.

To address a reviewer's suggestion, we also conducted an additional experiment using the type-based morphological analyzer for Latin LEMLAT (Passarotti et al., 2017). In this setup, the lemma wrongly predicted by LatinPipe was compared with the set of candidate lemmas returned by LEMLAT for each word form belonging to the list of technical lemmas. The gold lemma was used to tackle potentially ambiguous word forms and include only those belonging to lemmas included in our list. The aim was to determine potentially problematic lemmas in the list of validated terms: if neither of the two systems was able to output the correct lemma for the word form, the lemma was flagged as 'problematic' for lemmatization.

The results show that, in the vast majority of cases, the correct lemma is present among the candidates generated by LEMLAT, although not explicitly disambiguated. Any other case of mismatch was due to inconsistencies in ortography (e.g. *submoueo* vs. *summoveo* 'to remove, to banish') or lemmatization rules (e.g. *melior* 'better', comparative form from the adjective *bonus* 'good' is lemmatized as *bonus* in our gold dataset and as *melior* by LEMLAT). Only one form in our dataset is not correctly captured by either system. This form is the participle *reliquatus*, from the verb *reliquor*, which is specifically used in legal contexts and means 'to be in arrears in respect of money

PoS	Range	I / S / W	Improvement (%)	Worsening (%)
ADJ	Frequent	1 / 7 / 0	12.5	0.0
	Medium	2 / 6 / 0	25.0	0.0
	Rare	1 / 14 / 0	6.7	0.0
NOUN	Frequent	10 / 65 / 1	13.2	1.3
	Medium	8 / 74 / 0	9.8	0.0
	Rare	6 / 83 / 0	6.7	0.0
VERB	Frequent	31 / 50 / 2	37.3	2.4
	Medium	14 / 58 / 1	19.2	1.4
	Rare	4 / 76 / 0	5.0	0.0

Table 4: Lemma-level changes on technical terms by PoS and frequency range. Raw counts are reported as Improved (I) / Stable (S) / Worsened (W). Percentages indicate improvement and worsening rates within each PoS × frequency range category.

PoS	Lemma	LP Acc.	LP+GPT Acc.	Δ	Occ.	Range
VERB	parco	0.50	1.00	+0.50	2	rare
	creditor	0.55	0.91	+0.36	102	frequent
	dirimo	0.50	0.83	+0.33	6	rare
	infero	0.56	0.88	+0.31	16	medium
NOUN	pignus	0.37	0.88	+0.51	59	frequent
	tribunal	0.25	0.75	+0.50	4	rare
	suspectus	0.83	1.00	+0.17	6	rare
	emptum	0.77	0.92	+0.15	13	medium
ADJ	praesens	0.75	0.94	+0.19	17	medium
	ratus	0.77	0.92	+0.15	13	medium
	absens	0.91	1.00	+0.09	11	rare
	bonus	0.87	0.88	+0.01	67	frequent

Table 5: Top four improved lemmas per PoS (LatinPipe → LatinPipe + GPT). Δ indicates the increase in lemma-level accuracy.

owed'.

These results highlight two main points. On the one hand, LEMLAT provides broad coverage of Latin legal technical vocabulary, at least with respect to the material attested in the Digest. Indeed, all forms except one are associated with the correct lemma. On the other hand, however, LEMLAT is not a lemmatizer in itself, but a morphological analyzer. Crucially, it does not perform contextual disambiguation: for each form, it outputs all possible morphological analyses and, consequently, all possible lemmas, which differ from one another. Therefore it does not constitute a direct alternative to the LLM in our specific experimental context. This observation further supports our choice of a hybrid pipeline in which ambiguity is resolved through context-sensitive models rather than through type-based analyses.

6. Conclusions and future work

Currently, LLMs do not represent valid replacements for state-of-the-art tools in lemmatization tasks. This limitation is due to several factors. Lemmatization is not a simple binary or multiple-choice classification task; rather, it involves a complex process that is computationally and economically costly, given the large number of input and output tokens and the need for high-performance GPUs. Moreover, the generation of CoNLL-U formatted data seems to be an excessively complex task, failing particularly when applied to datasets exceeding ten sentences.

Given this as a starting point, our evaluation was intentionally conceived as a goal-aligned assessment rather than as a generic improvement of lemmatization accuracy. Instead of optimizing for global performance, we restricted the analysis to a curated list of expert-validated technical lemmas that are directly relevant to glossary construction. This methodological choice reflects the downstream objective of the project: building a reliable glossary of Roman legal terminology. From this perspective, the hybrid pipeline shows desirable properties. It is largely conservative, introduces minimal regressions, and selec-

tively improves morphologically complex or distributionally central technical lemmas. The analysis demonstrates that improvements are not randomly distributed, but show patterns according to frequency range and PoS dimensions. In particular, the system disproportionately benefits frequent verbal lemmas, which are more prone to inflectional ambiguity.

However, this evaluation scenario should not be interpreted as a fully realistic setting, for the following reasons. Firstly, the analysis assumes access to a validated list of domain-specific lemmas. Such curated resources are rarely available in real-world scenarios, especially in low-resource historical domains, and for specialized languages like the Roman legal lexicon. The ability to focus evaluation on a predefined terminology set presupposes expert knowledge and manual curation, which may not scale to other corpora or domains. Moreover, the hybrid correction strategy implicitly assumes computational resources sufficient to post-process and evaluate lemma-level discrepancies. In large-scale or real-time applications, correcting or re-evaluating each lemma individually through an LLM layer may be computationally too heavy. Finally, the improvements observed here are selective rather than systemic. The hybrid pipeline does not substantially alter the overall lemmatization performances; most lemmas remain stable. While this stability is methodologically reassuring, it also means that the hybrid approach does not eliminate the need for human validation in large-scale terminology extraction workflows.

Our comparison with LEMLAT shows that, while it provides exhaustive coverage of morphologically plausible analyses of legal Latin terms, it does not perform contextual disambiguation. As a consequence, its integration into the lemmatization pipeline in place of the LLM would shift the task from lemma prediction to candidate selection, which would still require either manual intervention or an additional disambiguation component. In this perspective, LEMLAT can be effectively used as a consistency check, or as a candidate generator for future external validation via an LLM or a human annotator.

To conclude, the present results should not be interpreted as evidence that LLM-based post-processing provides a general solution to domain-specific lemmatization. Rather, they indicate that, under controlled conditions and with clearly defined terminology targets, hybrid pipelines can selectively improve structurally relevant portions of the lexicon. By shifting the focus from global accuracy to terminology-relevant behavior, we obtain a more nuanced understanding of where hybrid pipelines provide added value, and where they do not.

7. Author contributions

PM: Conceptualization; Data Curation; Formal Analysis; Investigation; Methodology; Validation; Visualization; Writing – original draft (all sections); Writing – review & editing.
ES: Supervision; Funding Acquisition; Writing – review and editing.
Contributions follow the Contributor Role Taxonomy: https://doi.org/10.5281/zenodo.18421448

8. Resources

We release all resources (gold standard and code) in the GitHub repository: https://github.com/paoma370/Justinian-Digest-project.git.

9. Acknowledgements

This work stems from the project PRIN 2022 PNRR P20224NJLK – SH2 – TESTO “Translating, Encoding, Sharing The Origins”. From the Littera Florentina to an open-access Italian translation of Justinian’s Digest”. Piano Nazionale Ripresa e Resilienza (PNRR) M4C2 – Investimento 1.1 “Progetti di Ricerca di Rilevante Interesse Nazionale (PRIN)- funded by the European Union – NextGeneration EU” – CUP B53D23032480001.
The authors also wish to thank Michele Pedone (PI of the TESTO project) and Alessandro Grillone (Pisa University) for lending their expertise to validate the list of legal Latin terms, which was foundational for carrying out this work.

10. Bibliographical References

Thomas M Banchich, John Marenbon, and Charles J Reid Jr. 2015. The revival of roman law and canon law. In *A Treatise of Legal Philosophy and General Jurisprudence: Volume 6: A History of the Philosophy of Law from the Ancient Greeks to the Scholastics*, pages 251–265. Springer.

Marco Baroni. 2009. Distributions in text. In Anke Lüdeling and Merja Kytö, editors, *Corpus linguistics: An international handbook. Volume 2*, pages 803–821. Mouton de Gruyter, Berlin.

Ruirui Chen, Chengwei Qin, Weifeng Jiang, and Dongkyu Choi. 2024. Is a large language model a good annotator for event extraction? In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 17772–17780.

Frederick W Dingledy. 2016. The corpus juris civilis: a guide to its history and use. *Legal Reference Services Quarterly*, 35(4):231–255.

Margherita Fantoli, Marco Passarotti, Dominique Longrée, and Concepción Cabrillana. 2024. Lemmas in dialogue: Linking the lasla corpus to the lila knowledge base. *Recent Trends and Findings in Latin Linguistics: Volume I: Syntax, Semantics and Pragmatics. Volume II: Semantics and Lexicography. Discourse and Dialogue*, page 297.

Margherita Fantoli, Marco Passarotti, Francesco Mambrini, Giovanni Moretti, and Paolo Ruffolo. 2022. Linking the LASLA corpus in the LiLa knowledge base of interoperable linguistic resources for Latin. In *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference*, pages 26–34, Marseille, France. European Language Resources Association.

Andrea Farina, Andrea Ballatore, and Barbara McGillivray. 2025. Mapping meaning in latin with large language models: A multi-task evaluation of preverbed motion verbs and spatial relation detection in llms. In *CLiC-it 2025 Italian Conference on Computational Linguistics. Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025) Cagliari, Italy, September 24-26, 2025.*

Eleonora Ghizzota, Paola Marongiu, Pierpaolo Basile, Stefano Ferilli, and Barbara McGillivray. 2026. Linguistic knowledge graphs for sense prediction: a case-study on latin. In *Proceedings of the 15th Language Resources and Evaluation Conference*. Anthology of Computers and the Humanities.

Saniya Irfan and Syed Juned Ali. 2025. Qallm: An llm-based ner dataset curation, annotation and evaluation in historical urdu elegies. In *Computational Humanities Research 2025*, pages 921–936. Anthology of Computers and the Humanities.

Siyu Liang, Talant Mawkanuli, and Gina-Anne Levow. 2026. Hybrid neural-LLM pipeline for morphological glossing in endangered language documentation: A case study of jungar tuvan. In *Proceedings of the Fifth Workshop on NLP Applications to Field Linguistics*, pages 16–30, Rabat, Morocco. Association for Computational Linguistics.

Theodor Mommsen and P. Krüger, editors. 1911. *Corpus iuris civilis: editio stereotypa duodecima. Institutiones*, volume 1. Weidmannos, Berlin.

Kseniia Petukhova and Ekaterina Kochmar. 2025. A fully automated pipeline for conversational discourse annotation: Tree scheme generation and labeling with large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15829–15852.

Marton Ribary and Barbara McGillivray. 2020. A corpus approach to roman law based on justinian's digest. In *Informatics*, volume 7, page 44. MDPI.

Frederick Riemenschneider. 2025. Beyond base predictors: Using LLMs to resolve ambiguities in Akkadian lemmatization. In *Proceedings of the Second Workshop on Ancient Language Processing*, pages 226–231, The Albuquerque Convention Center, Laguna. Association for Computational Linguistics.

Sandro Schipani. 2005. *Iustiniani Augusti Digesta seu Pandectae. Digesti o Pandette dell'Imperatore Giustiniano. Testo e traduzione*, volume 1. Giuffrè Editore.

Rachele Sprugnoli, Federica Iurescia, and Marco Passarotti. 2024. Overview of the EvaLatin 2024 evaluation campaign. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 190–197, Torino, Italia. ELRA and ICCL.

Rachele Sprugnoli, Marco Passarotti, Flavio Massimiliano Cecchini, Margherita Fantoli, and Giovanni Moretti. 2022. Overview of the EvaLatin 2022 evaluation campaign. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 183–188, Marseille, France. European Language Resources Association.

Rachele Sprugnoli, Marco Passarotti, Flavio Massimiliano Cecchini, and Matteo Pellegrini. 2020. Overview of the EvaLatin 2020 evaluation campaign. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 105–110, Marseille, France. European Language Resources Association (ELRA).

Olia Toporkov, Alan Akbik, and Rodrigo Agerri. 2025. Lemma dilemma: On lemma generation without domain-or language-specific training data. *arXiv preprint arXiv:2510.07434*.

Martin Volk, Dominic Philipp Fischer, Lukas Fischer, Patricia Scheurer, and Phillip Benjamin Ströbel. 2024. LLM-based machine translation and summarization for Latin. In *Proceedings of the Third Workshop on Language*

Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024, pages 122–128, Torino, Italia. ELRA and ICCL.

Leonie Weissweiler, Abdullatif Köksal, and Hinrich Schütze. 2025. Hybrid human-llm corpus construction and llm evaluation for the caused-motion construction. *Northern European Journal of Language Technology*, 11(1):27–57.

George Kingsley Zipf. 1949. *The Principle of Least Effort*. Addison-Wesley.

11. Language Resource References

David Bamman and Gregory Crane. 2011. The ancient greek and latin dependency treebanks. In *Language technology for cultural heritage: Selected papers from the LaTeCH Workshop Series*, pages 79–98. Springer.

Flavio Massimiliano Cecchini, Timo Korkiakangas, and Marco Passarotti. 2020a. A new latin treebank for universal dependencies: Charters between ancient latin and romance languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 933–942.

Flavio Massimiliano Cecchini, Rachele Sprugnoli, Giovanni Moretti, and Marco Passarotti. 2020b. Udante: First steps towards the universal dependencies treebank of dante's latin works. In *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020)*, pages 66–72.

Margherita Fantoli and Miryam de Lhoneux. 2022. Linguistic annotation of neo-latin mathematical texts: a pilot-study to improve the automatic parsing of the archimedes latinus. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 129–134.

Federica Gamba and Flavio Massimiliano Cecchini. 2024. De latinae linguae reparatione treebank. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Federica Gamba and Daniel Zeman. 2023. Latin morphology through the centuries: Ensuring consistency for better language processing. In *Proceedings of the Ancient Language Processing Workshop*, pages 59–67.

Dag TT Haug and Marius Jøhndal. 2008. Creating a parallel treebank of the old indo-european bible translations. In *Proceedings of the second workshop on language technology for cultural heritage data (LaTeCH 2008)*, pages 27–34. Prague.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Marco Passarotti. 2019. The project of the index thomisticus. *Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution*, 10:299.

Marco Passarotti, Marco Budassi, Eleonora Litta, and Paolo Ruffolo. 2017. The lemlat 3.0 package for morphological analysis of Latin. In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, pages 24–31, Gothenburg. Linköping University Electronic Press.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Milan Straka and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Milan Straka and Jana Straková. 2020. UDPipe at EvaLatin 2020: Contextualized embeddings and treebank embeddings. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 124–129, Marseille, France. European Language Resources Association (ELRA).

Milan Straka, Jana Straková, and Federica Gamba. 2024. ÚFAL LatinPipe at EvaLatin 2024: Morphosyntactic analysis of Latin. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 207–214, Torino, Italia. ELRA and ICCL.

UppsalaNLP at EvaLatin 2026: Multilingual parsing for Latin

Sara Stymne
Department of Linguistics and Philology
Uppsala University, Sweden
sara.stymne@lingfil.uu.se

Abstract

We describe the UppsalaNLP submission to the EvaLatin dependency parsing shared task. We explore using an out-of-the-box parser in combination with multi-treebank training on Latin and multilingual training on other ancient languages. Adding additional languages yields only small gains, but the results vary across treebanks and genres, with the largest positive effect for poetry. Our systems perform best in the shared task for prose but are less competitive for poetry, indicating the need for genre adaptation.

Keywords: Dependency parsing, Latin, Multilingual parsing

1. Introduction

We present the UppsalaNLP entry in the EvaLatin 2026 shared task on dependency parsing for Latin (Iurescia et al., 2026). Our main focus is to explore how an out-of-the-box system, MaChamp (van der Goot et al., 2021) in our case, can be improved through multilingual and multi-treebank training. Our main goal was to investigate how far it is possible to get with an existing strong parser, without any technical additions, only by combining existing treebanks during training. This is a feasible setup for many digital humanities projects.

Our base system is a multi-treebank system trained on eight Latin treebanks, compatible with the Universal Dependencies (UD, Nivre et al., 2020) annotation scheme, see Table 1 for an overview. We also follow Straka et al. (2024), who noticed that the Latin PROIEL treebank diverged from the other Latin treebanks, and used a harmonized version of it (Gamba and Zeman, 2023a,b).

For multilingual training, we also include other ancient languages, which have been shown to be useful for Latin parsing (Smith et al., 2018), including Ancient Greek, which has been shown to be useful in a Latin low-resource scenario (Karamolegkou and Stymne, 2021). In addition, we also include historical Romance languages, which are now available in UD. Many of the treebanks from the other languages stem from the PROIEL project, for which the Latin treebank has inconsistent annotations (Gamba and Zeman, 2023a,b). Thus, we also experiment with the impact of excluding them from the training. In addition, we explore the impact of multitask training on a mix of UD tasks.

We submit two systems to the EvaLatin shared task. Our primary system is trained on Latin in combination with other historical languages, but excluding PROIEL treebanks, except for Latin, where a harmonized version is used. Our secondary submission is further trained only on Latin. Our primary system ranks first on most metrics for the prose data in the EvaLatin evaluation, but it ranks fourth for the poetry data. Our secondary system is placed below the primary system for prose, second to third, but beats it for poetry, being placed third. This calls for further investigation into genre adaptation for parsing to better account for the peculiarities of poetry, an issue that has been previously investigated for other genres, domains, and languages (Stymne, 2020; Müller-Eberstein et al., 2021; Danilova and Stymne, 2023). Our multilingual approach yields only small gains compared to training only on Latin when evaluated on the test portions of the training treebank, but has a larger effect on the shared task test data, especially for poetry.

2. Related Work

EvaLatin 2026 (Iurescia et al., 2026) is a follow-up from EvaLatin 2024 (Sprugnoli et al., 2024). For both editions, the focus is on dependency parsing for Latin prose and poetry. In 2024, the winning team presented LatinPipe (Straka et al., 2024), an extension of UDPipe (Straka and Straková, 2017). LatinPipe contains both Latin-specific extensions and general architecture improvements. Specific to Latin, they train on multiple Latin treebanks, including a harmonized version of PROIEL (Gamba and Zeman, 2023a,b), use the gold UPOS (Universal part-of-speech tags de Marneffe et al., 2021) provided by the EvaLatin organizers, and combine multilingual and Latin-specific language models (LMs). They also add two BiLSTM layers on top of the LMs and perform ensembling. All of these techniques are shown to improve results on both UD and EvaLatin test sets. Other teams at EvaLatin 2024 explored using historical Latin sentence embeddings (Behr, 2024) and span-span prediction on top of a Latin LM (Mercelis, 2024), but with considerably worse performance than LatinPipe.

One approach for Latin dependency parsing is treebank combination, either for Latin alone or in

Proceedings of LT4HALA 2026 @ LREC 2026, pages 429–436.
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

a cross-lingual setting. An obstacle to the success of combining treebanks is the inconsistencies of the annotation, even within a standardized framework like UD. For Latin, Gamba and Zeman (2023b) show that there are considerable discrepancies across the UD Latin treebanks, especially for PROIEL. They design an automatic harmonization scheme that leads to improved results across treebanks. Kupari et al. (2024) also use these treebanks, further improving by using stronger parsers. Smith et al. (2018) explore treebank combination and find that small groups of treebanks for related languages tend to improve parsing results for most languages, especially for those with small treebanks. For Latin, they use a set of ancient-language treebanks, which improves performance on most Latin UD test sets, possibly due to shared annotation projects between these languages. In their study, they used UUParser (de Lhoneux et al., 2017), a BiLSTM-based parser. This study is revisited by van der Goot and de Lhoneux (2021), using the MaChamp parser (van der Goot et al., 2021), based on the transformer model mBERT (Devlin et al., 2019). Both overall and for Latin, they find that cross-treebank training is useful, but that it is often as effective to use all available treebanks as the designed language sets by Smith et al. (2018). In addition, they see a positive effect of using dataset embeddings. Karamolegkou and Stymne (2021) explore cross-lingual parsing in a low-resource scenario for Latin, and find that Ancient Greek is a better transfer language than modern romance languages like Italian. Kupari et al. (2024) explored the impact of using data from specific time periods on dependency parsing of Latin; however, the impact was minor.

Several works have investigated which LMs are most useful for parsing Latin. Nehrdich and Hellwig (2022) show that the Latin BERT model (Bamman and Burns, 2020) performs better than both static embeddings and mBERT. Straka et al. (2024) compare the use of two Latin-specific BERT models (Riemenschneider and Frank, 2023) to the multilingual XLM-RoBERTa (XLM-R) model (Conneau et al., 2020), and find that XLM-R is better than using the Latin models, but that it is possible to get additional gains by combining the Latin models with XLM-R. Behr (2024) investigated the usefulness of historical Latin sentence embeddings (Reimers and Gurevych, 2019) fine-tuned on top of Latin BERT (Bamman and Burns, 2020), which, however, did not improve the parsing scores.

Another line of work, for dependency parsing in general, focuses on improving parsing for specific genres. Stymne (2020) matches genres across languages, Müller-Eberstein et al. (2021) train parsers based on cross-lingual genre clusters, and Danilova and Stymne (2023) mine genre annotations from UD treebanks, and compare human genre annotations with genre clusters. None of these works specifically targets the goal of parsing poetry and prose in historical languages. This has been addressed for POS-tagging of historical Germanic languages, where it was shown that training on prose for tagging poetry, or vice versa, was not successful, and combining the genres had only a minor effect (Miani et al., 2026).

3. Data

Table 1 summarizes the Latin treebanks used. The majority of treebanks are from UD, release 2.17 (Zeman et al., 2025). Due to annotation discrepancies across UD treebanks, we follow LatinPipe (Straka et al., 2024) and explore the use of a set of harmonized treebanks (referred to as "H", Gamba and Zeman (2023a,b)) for all UD treebanks except CIRCSE, for which no harmonization is available. We try to use the harmonized treebanks for all treebanks for which they are available, or only for PROIEL. In addition, we use the same two small treebanks (and abbreviations) as LatinPipe, both annotated according to the UD scheme. For testing, we use the harmonized version of PROIEL, which we will refer to as H-PROIEL, and the UD2.17 version of all other UD treebanks. Arch has a small test set of 30 sentences, which we consider insufficient for reliable evaluation. The EvaLatin shared task provides two official testsets, covering prose, 203 sentences, and poetry, 308 sentences.

We also include UD treebanks from other ancient languages, grouped into three groups: Ancient Greek, Romance, covering Old French and Old Occitan, and Other, covering Gothic and Old Church Slavonic. Table 2 contains an overview, including the language groupings. In addition, we train a model using all these treebanks. In this set, several treebanks come from the PROIEL project (Haug and Jøhndal, 2008), which was shown to diverge from the UD framework in some respects for Latin. We thus also experiment with excluding those treebanks and train on only treebanks from other projects.

4. Dependency Parser

We use the MaChamp toolkit (van der Goot et al., 2021) for dependency parsing. MaChamp is a multipurpose toolkit targeted at a diverse set of NLP tasks, including dependency parsing, and is especially useful for multitask and multi-dataset training. It fine-tunes a language model for specific tasks. We use XLM-RoBERTa Large (Conneau et al., 2020), which performed better than any Latin-specific model in isolation and only slightly worse than in combination with two Latin-based models

Short	Treebank	Size in sentences Training	Test
Circse	UD_Circse (https://github.com/UniversalDependencies/UD_Latin-CIRCSE)	762	832
ITTB	UD_ITTB (Passarotti, 2019)	22,775	2,101
LLCT	UD_LLCT (Cecchini et al., 2020a)	7,289	894
Perseus	UD_Perseus (Bamman and Crane, 2011)	1,334	939
PROIEL	UD_PROIEL (Haug and Jøhndal, 2008)	3,387	1260
UDante	UD_UDante (Cecchini et al., 2020b)	926	421
Arch	Archimedes Latinus (Fantoli and de Lhoneux, 2022)	47	(30)
Sab	De Latinae Linguae Reparatione (Sabellicus) (Gamba and Cecchini, 2024)	246	–

Table 1: Latin treebanks used with sizes of training and test sets. Treebanks in the top of the table is part of UD, treebanks in the bottom are compatible with UD.

Language	Treebank	Training size	A. Greek	Romance	Other	No PROIEL	All
	UD_Perseus (Bamman and Crane, 2011)	11,476	X			X	X
Ancient Greek	UD_PROIEL (Haug and Jøhndal, 2008)	15,016	X				X
	UD_PTNK (Swanson et al., 2024)	727	X			X	X
Old French	UD_PROFTIEROLE (Stein and Prévost, 2013)	15,962		X		X	X
Old Occitan	UD_CorAG (Romanova et al., 2025)	912		X		X	X
Gothic	UD_PROIEL (Haug and Jøhndal, 2008)	3,387			X		X
Old Church Slavonic	UD_PROIEL (Haug and Jøhndal, 2008)	18,327			X		X

Table 2: Treebanks used for additional languages. The final columns show which groups of training languages each treebank is included in.

for LatinPipe (Straka and Straková, 2017). The MaChamp dependency parser is a non-projective graph-based parser based on the CLE algorithm (Chu and Liu, 1965; Edmonds, 1967), using deep biaffine attention (Dozat and Manning, 2018). In our multitasking experiments, we also predict UPOS, morphological tags, and lemmas, the latter by predicting conversions from word forms to lemmas. These tasks use a standard greedy sequence labelling approach. We use default parameters in MaChamp, except for the sample smoothing parameter a, used for balancing dataset size during training, which we set to 0.5 to prevent larger datasets from overwhelming smaller ones.

5. Experiments

Our main focus is on combining treebanks for Latin and other ancient languages. We also explore the impact of multitask training, comparing training MaChamp only for dependency parsing, training it also for UPOS, and training it also for UPOS, morphology prediction, and lemmatization.

For our treebank combination experiments, we first train models for each treebank, as a point of comparison. We then combine all available Latin treebanks in three settings: use only UD2.17 versions, use the harmonized versions of UD treebanks (where available), or use UD2.17 treebanks for all treebanks except PROIEL, for which the harmonized version was used.

We then combine the training on Latin with groups of other languages, as shown in Table 2, and with all treebanks, or all treebanks excluding PROIEL. For our final set of experiments, we take the multilingual model, trained on all data except non-Latin PROIEL treebanks, and further train it only on Latin, either on all Latin treebanks (with harmonized PROIEL) or only on CIRCSE, which we believe could be a better fit for poetry data.

Note that we do not use any gold UPOS-tags or morphology as input to our parser. While this information was available for the EvaLatin shared task and could have boosted the result, we prefer the more realistic setting of not using gold tags, which are not generally available for Latin texts.

For results on the testsets of the Latin treebanks used for training, we report labeled attachment score (LAS) without any subtypes. For the EvaLatin testdata, we report all official metrics of the EvaLatin shared task (Iurescia et al., 2026), which include LAS and CLAS, content-word LAS, which excludes relations to function words (Nivre and Fang, 2017), for both metrics with and without subtypes. All reported results are averages across three runs with different seeds, except for the official shared task results, which were calculated for the submitted systems.

6. Results

Table 3 shows LAS scores on the testsets from the training treebanks. Table 3.A shows results for multitasking experiments trained on all Latin treebanks using H-PROIEL. The effect of including other tasks than dependency parsing is minor and varies across treebanks. In the following, we will use systems trained on dependency parsing and UPOS that had the best average scores, by a small margin.

Experiment	H-PROIEL	Circse	ITTB	LLCT	Perseus	UDante	Average
A) Multitask training variants							
1 Dep only	86.83	68.43	**91.37**	**95.88**	78.52	**77.10**	83.02
2 Dep+UPOS+morph+lemma	86.54	**70.12**	91.06	95.62	**79.12**	76.70	83.19
3 Dep+UPOS	**86.93**	69.31	91.23	95.86	79.05	77.00	**83.23**
B) Latin treebank variants and combinations(Harmonized/UD)							
4 Single treebank	**88.84**	60.60	**92.54**	95.83	75.23	75.09	81.36
5 All H+ Arch+Sab	86.94	**70.30**	88.82	88.36	**79.09**	75.98	81.58
6 All UD + Arch+sab	81.98	69.49	91.19	95.83	78.93	**77.34**	82.46
3 H PROIEL+UD+Arch+Sab	86.93	69.31	91.23	**95.86**	79.05	77.00	**83.23**
C) Multilingual training, Latin (as in previous row) +							
7 Ancient Greek	87.31	70.54	91.31	95.95	79.01	77.22	83.56
8 Other	**87.36**	70.30	**91.43**	95.90	79.01	**77.40**	83.57
9 Romance	87.06	70.62	91.27	95.91	79.01	76.98	83.48
10 All	**87.36**	70.94	91.40	95.97	**79.33**	77.09	**83.68**
11 No PROIEL	86.66	**71.10**	91.32	**95.99**	**79.33**	77.22	83.60
D) Continued Latin only training (w.r.t in previous row)							
12 Circse only	84.00	70.43	75.20	77.65	63.66	62.79	72.29
13 H PROIEL+UD+Arch+Sab	**86.86**	**71.05**	**91.36**	**95.94**	**79.18**	**76.86**	**83.54**
E) Previous work							
MaChAmp (UD2.8)	–	–	92.45	95.41	74.67	74.01	–
LatinPipe (ensemble – UD2.13+H-PROIEL+Sab+Arch)	–	–	92.45	95.78	84.22	81.47	–

Table 3: LAS scores without subtypes on the testsets corresponding to the Latin training data. For PROIEL, we report scores in the harmonized version. Previous work are from Straka et al. (2024) (LatinPipe) and van der Goot et al. (2021) (MaChamp)

Table 3.B shows results with different combinations of Latin treebanks for training. For two of the larger treebanks, PROIEL and ITTB, we achieved the best results by training only on data from that treebank. For all other treebanks, training on the combination of all Latin treebanks was better, with large gains especially for the small CIRCSE treebank. Using original UD treebanks or harmonized treebanks had varying effects across treebanks, but only using the harmonized version of PROIEL led to the best average results and is used in all further experiments.

Table 3.C shows the effect of combining Latin data with other languages. Overall, this had little effect compared to training only on Latin, but led to small improvements for all treebanks, compared to the system with its corresponding Latin treebanks (system 3 in Table 3). The best language group varies across treebanks, but on average, models that used all languages performed slightly better than those that used a subset. While the model without non-Latin PROIEL treebanks performed slightly worse on average than the model trained on all treebanks, we still chose to use it in the following, since the model trained on all treebanks mainly improved H-PROIEL, indicating that the non-Latin treebanks may have the same discrepancies with UD annotation as Latin PROIEL.

The results in Table 3.D show the effect of further training the multilingual model only on Latin, either on all treebanks, or only on the small CIRCSE. Training on the small CIRCSE deterioriated the results for all treebanks, possibly due to overfitting, and even led to worse results on the CIRCSE test set. Using all Latin treebanks did not improve overall over the multilingual model, but kept the results relatively stable.

In Table 3.E, we compare our results with previous work, in the form of MaChamp and LatinPipe. Note, though, that those scores are for other UD versions than ours. For 3 out of 4 treebanks, our multilingual model is better than the original MaChamp model (van der Goot et al., 2021). For ITTB, our multilingual model falls behind, but the model trained only on ITTB is slightly ahead of MaChamp. We still fall behind the winner of the previous EvaLatin shared task, LatinPipe (Straka et al., 2024), except for the LLCT treebank. We believe that adding some of the LatinPipe features not used in this work, such as gold UPOS, Latin LMs, BiLSTM layers, and ensembling, could potentially bridge the gap between these systems.

Based on the results in Table 3, we decided to submit system 11, a multilingual system excluding non-Latin PROIEL and using harmonized Latin PROIEL as our primary system, UppsalaNLP 1, and system 13, further trained on Latin, as our secondary system, UppsalaNLP 2. Table 4 shows the official results for these systems, and also the results for a subset of contrastive systems, including the two single-source treebanks chosen for training the EvaLatin baselines, ITTB and CIRCSE (Iurescia et al., 2026). We also include ranks in the shared task for our submitted systems.

For prose, our primary system performed best on 3 out of 4 official metrics, whereas our secondary system performed better on poetry, where it ranked 3rd on all metrics. For prose, UppsalaNLP 1 performed on par with training only on ITTB, but better than all other systems. For poetry, UppsalaNLP 2 performed stronger than all our other systems. The difference between the two submitted systems was quite small on the test sets from the training treebanks, in Table 3, but considerably larger on the shared-task test sets. The multilingual training, in the submitted systems and in system 10 improves

		Prose				Poetry			
		LAS	LAS+s	CLAS	CLAS+s	LAS	LAS+s	CLAS	CLAS+s
11	UppsalaNLP 1 (Latin+no PROIEL)	**87.55**	**84.41**	**86.88**	**83.14**	69.59	66.63	68.20	67.34
	RANK	2	1	1	1	4	4	4	4
13	UppsalaNLP 2 (11 + further Latin training)	87.26	83.74	86.45	82.17	**70.55**	**67.56**	**68.47**	**68.47**
	RANK	3	2	3	2	3	3	3	3
3	Latin only	87.29	83.81	86.60	82.31	68.48	65.45	66.92	65.97
10	Latin + all	**87.37**	**84.03**	**86.65**	**82.79**	**69.88**	66.53	**68.41**	**67.23**
12	Fine-tuning on CIRCSE	73.36	70.36	78.01	76.34	69.23	**66.64**	67.74	67.18
4	CIRCSE	57.23	53.31	59.48	56.55	**59.69**	**57.35**	**58.69**	**59.14**
4	ITTB	**87.89**	**84.43**	**86.92**	**82.77**	45.43	42.73	43.33	42.02

Table 4: Official results for the submitted runs on all shared task metrics (top) and some contrastive results (bottom). Metrics marked with '+s' include the evaluation of subtypes. Numbers refer to systems presented in Table 3.

over Latin-only training, system 3, on all metrics for both genres, but especially for poetry, even though adding non-Latin languages only had a small effect on the devsets in Table 3.C.

For poetry, none of the single-treebank systems performed well, and ITTB, which was very competitive for prose, even fell behind the small CIRCSE treebank. Further training the multilingual system on only CIRCSE was competitive for poetry, but not for prose. For poetry, our multilingual system performed considerably better than the Latin-only systems. The multilingual system, including the non-Latin PROIEL treebanks, performed similarly to not using PROIEL.

In most cases, the four official metrics show the same tendencies. In the case of training on CIRCSE, system 12, though, we can see that this system performs comparatively better on the metrics with subtypes, indicating that CIRCSE is in better agreement with the test data for subtypes than the other treebanks.

We note that the accuracy for UPOS-tagging, which was performed during our multitask training, was just over 94% on poetry but over 97.5% on prose for our two submitted systems, indicating that not only is parsing poetry challenging, but also other tasks. The winners from 2024, (Straka et al., 2024) report a gain of over 1 LAS point when adding gold POS to their parser, indicating that improved tagging could also help parsing.

7. Conclusion

We have presented the UppsalaNLP contribution to the EvaLatin shared task, which placed first for parsing prose and 3–4 on parsing poetry. Our system uses multilingual training for Latin in combination with other ancient languages. The multilingual training mostly led only to minor improvements, except for poetry, for which it led to a substantial improvement. We believe that some of the most pressing future work is to improve parsing for poetry, possibly by including poetry training data for other languages.

8. Acknowledgement

Computations were enabled by resources provided by the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX).

9. Bibliographical References

David Bamman and Patrick J. Burns. 2020. Latin BERT: A contextual language model for classical philology. *CoRR*, abs/2009.10053.

David Bamman and Gregory Crane. 2011. The Ancient Greek and Latin dependency treebanks. In *Language Technology for Cultural Heritage*, pages 79–98. Springer Berlin Heidelberg, Berlin, Heidelberg.

Rufus Behr. 2024. Behr at EvaLatin 2024: Latin dependency parsing using historical sentence embeddings. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 198–202, Torino, Italia. ELRA and ICCL.

Flavio Massimiliano Cecchini, Timo Korkiakangas, and Marco Passarotti. 2020a. A new Latin treebank for Universal Dependencies: Charters between Ancient Latin and Romance languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 933–942, Marseille, France. European Language Resources Association.

Flavio Massimiliano Cecchini, Rachele Sprugnoli, Giovanni Moretti, and Marco Passarotti. 2020b. UDante: First steps towards the universal dependencies treebank of Dante's Latin works. In *Proceedings of the Seventh Italian Conference on Computational Linguistics*, pages 1–7. Italian Association for Computational Linguistics (AILC).

Yoeng-Jin Chu and Tseng-Hong Liu. 1965. On the shortest arborescence of a directed graph. *Scientia Sinica*, 14:1396–1400.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Vera Danilova and Sara Stymne. 2023. UD-MULTIGENRE – a UD-based dataset enriched with instance-level genre annotations. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 253–267, Singapore. Association for Computational Linguistics.

Miryam de Lhoneux, Sara Stymne, and Joakim Nivre. 2017. Arc-hybrid non-projective dependency parsing with a static-dynamic oracle. In *Proceedings of the 15th International Conference on Parsing Technologies*, pages 99–104, Pisa, Italy. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Timothy Dozat and Christopher D. Manning. 2018. Simpler but more accurate semantic dependency parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 484–490, Melbourne, Australia. Association for Computational Linguistics.

Jack Edmonds. 1967. Optimum branchings. *Journal of Research of the national Bureau of Standards B*, 71(4):233–240.

Margherita Fantoli and Miryam de Lhoneux. 2022. Linguistic annotation of neo-Latin mathematical texts: A pilot-study to improve the automatic parsing of the archimedes latinus. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 129–134, Marseille, France. European Language Resources Association.

Federica Gamba and Daniel Zeman. 2023a. Latin morphology through the centuries: Ensuring consistency for better language processing. In *Proceedings of the Ancient Language Processing Workshop*, pages 59–67, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Federica Gamba and Daniel Zeman. 2023b. Universalising Latin Universal Dependencies: a harmonisation of Latin treebanks in UD. In *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*, pages 7–16, Washington, D.C. Association for Computational Linguistics.

Dag Trygve Truslew Haug and Marius Jøhndal. 2008. Creating a parallel treebank of the Old Indo-European Bible translations. In *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, pages 27–34.

Federica Iurescia, Marco Passarotti, and Rachele Sprugnoli. 2026. Overview of the Dependency Parsing Task at EvaLatin 2026. In *Proceedings of the Fourth Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2026)*, Palma, Mallorca (Spain). ELRA.

Antonia Karamolegkou and Sara Stymne. 2021. Investigation of transfer languages for parsing Latin: Italic branch vs. Hellenic branch. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 315–320, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Hanna-Mari Kristiina Kupari, Erik Henriksson, Veronika Laippala, and Jenna Kanerva. 2024. Improving Latin dependency parsing by combining treebanks and predictions. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 216–228, Miami, USA. Association for Computational Linguistics.

Wouter Mercelis. 2024. KU Leuven / Brepols-CTLO at EvaLatin 2024: Span extraction approaches for Latin dependency parsing. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 203–206, Torino, Italia. ELRA and ICCL.

Irene Miani, Sara Stymne, and Gregory R. Darwin. 2026. Cross-lingual and cross-domain transfer learning for POS tagging in historical Germanic low-resource languages. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026)*, pages 542–558, Rabat, Morocco. Association for Computational Linguistics.

Max Müller-Eberstein, Rob van der Goot, and Barbara Plank. 2021. Genre as weak supervision for cross-lingual dependency parsing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4786–4802, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sebastian Nehrdich and Oliver Hellwig. 2022. Accurate dependency parsing and tagging of Latin. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 20–25, Marseille, France. European Language Resources Association.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Joakim Nivre and Chiao-Ting Fang. 2017. Universal Dependency evaluation. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 86–95, Gothenburg, Sweden. Association for Computational Linguistics.

Marco Passarotti. 2019. The project of the Index Thomisticus treebank. *Digital Classical Philology*, 10:299–320.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Frederick Riemenschneider and Anette Frank. 2023. Exploring large language models for classical philology. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15181–15199, Toronto, Canada. Association for Computational Linguistics.

Natasha Romanova, Rayan Ziane, and Barbara Francioni. 2025. Adaptation of models for parsing of Old Gascon. In *Proceedings of Journées scientifiques du réseau thématique LIFT2 linguistique informatique, formelle et de terrain*, Paris, France.

Aaron Smith, Bernd Bohnet, Miryam de Lhoneux, Joakim Nivre, Yan Shao, and Sara Stymne. 2018. 82 treebanks, 34 models: Universal Dependency parsing with multi-treebank models. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

Rachele Sprugnoli, Federica Iurescia, and Marco Passarotti. 2024. Overview of the EvaLatin 2024 evaluation campaign. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 190–197, Torino, Italia. ELRA and ICCL.

Achim Stein and Sophie Prévost. 2013. Syntactic annotation of medieval texts: The Syntactic Reference Corpus of Medieval French (SRCMF). In Paul Bennett, Martin Durrell, Silke Scheible, and Richard J. Whitt, editors, *New Methods in Historical Corpora*, Corpus Linguistics and International Perspectives on Language, pages 275–282. Gunter Narr Verlag.

Milan Straka and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Milan Straka, Jana Straková, and Federica Gamba. 2024. ÚFAL LatinPipe at EvaLatin 2024: Morphosyntactic analysis of Latin. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 207–214, Torino, Italia. ELRA and ICCL.

Sara Stymne. 2020. Cross-lingual domain adaptation for dependency parsing. In *Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories*, pages 62–69, Düsseldorf, Germany. Association for Computational Linguistics.

Daniel G. Swanson, Bryce D. Bussert, and Francis Tyers. 2024. Producing a parallel Universal Dependencies treebank of Ancient Hebrew and Ancient Greek via cross-lingual projection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13074–13078, Torino, Italia. ELRA and ICCL.

Rob van der Goot and Miryam de Lhoneux. 2021. Parsing with pretrained language models, multi-

ple datasets, and dataset embeddings. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 96–104, Sofia, Bulgaria. Association for Computational Linguistics.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

10. Language Resource References

Federica Gamba and Flavio Massimiliano Cecchini. 2024. De latinae linguae reparatione treebank. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, et al. 2025. Universal dependencies 2.17. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Contextual Probing for Low-Resource Named Entity Recognition in Latin

Maria Mihaela Trusca, Mark Depauw, Violet Soen,
Ine de Daele, Kevin Verbruggen, Tim Van de Cruys
Faculty of Arts, KU Leuven
Blijde Inkomststraat 21, Leuven, 3000, Belgium
mariamihaela.trusca@kuleuven.be, mark.depauw@kuleuven.be, violet.soen@kuleuven.be,
ine.dedaele@kuleuven.be, kevin.verbruggen@kuleuven.be, tim.vandecruys@kuleuven.be

Abstract

Named Entity Recognition (NER) for low-resource languages remains challenging due to limited annotated data and linguistic characteristics such as rich morphology and flexible word order. In this work, we propose a probing-based method that leverages the contextual knowledge encoded in pretrained language models to detect entities. Our approach uses a substitution strategy in which words in a sentence are replaced, one by one, with candidate entities of predefined entity types, referred to as probes. By measuring how well the probes of a certain entity type fit the surrounding context of the replaced word, we estimate the compatibility between the replaced word and the entity type. The resulting compatibility scores can be used either as a standalone zero-shot NER model or as an auxiliary feature during NER model decoding. We evaluate our method on the Latin dataset provided in the Fourth Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA). Our system ranked second in the coarse-grained NER task. For the fine-grained NER task, where no training data were available, we relied exclusively on the proposed scoring method without any model training and achieved third place. These results demonstrate that contextual probing can provide an effective signal for NER in low-resource settings.

Keywords: Contextual Probing, Latin Named Entity Recognition, Zero-Shot Named Entity Recognition

1. Introduction

Named Entity Recognition (NER) is a core task in Natural Language Processing (NLP), with applications in question answering, information retrieval, and digital humanities research. While NER models have achieved strong performance for high-resource languages such as English, their application to low-resource languages, such as Latin, remains limited. This is partly due to linguistic factors such as rich morphology and flexible word order, which are difficult to model effectively when annotated training data are limited.

The current NER standard approach relies on pre-trained language models for the target language, which are subsequently fine-tuned for NER. Although this strategy is generally robust, the final performance of the NER system largely depends on the size and quality of the available training data.

In this paper, we propose a method designed to reduce the negative impact of limited NER data by more effectively leveraging the linguistic knowledge acquired by the language models during their pretraining. This knowledge is learned through a self-supervised language modeling task, in which the model is trained to predict missing or contextually appropriate tokens, encouraging it to capture rich contextual relationships between words. In our implementation, we exploit these contextual relationships to evaluate how well substitutions of well-known entities fit in the surrounding context, allowing us to infer whether a given position in the text likely corresponds to an entity.

The core idea of our method is simple yet effective. Given a predefined list of entities (or probes) of a specific type, we use a substitution-based probing approach to identify entities. For each word in the text, we replace it individually with each probe from a predefined list of known entities of a given type. When the language model predicts that the substitution fits naturally in context, this indicates that the original word is functionally similar to the probe, and we therefore infer that it is an entity of that type.

The proposed method does not require additional training and is suitable for zero-shot NER. However, when NER training data are available, the compatibility between probes and the context can be used as a complementary signal to better guide NER models during decoding in the fine-tuning process. We demonstrate the utility of our approach by evaluating it on Latin, a low-resource language, using the test dataset provided by the Fourth Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA).

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 presents our method. Section 4 describes the experimental setup and Section 5 reports the results. Section 6 concludes the paper.

Proceedings of LT4HALA 2026 @ LREC 2026, pages 437–442.
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

2. Related Works

2.1. Probing for Named Entity Recognition

Recent work has explored probing and few-shot techniques to improve NER performance in low-resource settings. Jiang et al. (2024) propose point in-context learning (P-ICL), where lists of representative entities are provided as few-shot prompts to a language model. Cocchieri et al. (2025) instead leverage natural language definitions of entity types, assigning NER labels based on the similarity between work or span embeddings and entity type embeddings. Shachar et al. (2025) use a similar embedding-based similarity but reformulate NER as a retrieval task, ranking candidate entities rather than producing NER labels. Zhang et al. (2023) and Liang et al. (2024) combine the strengths of both approaches by constructing prototypes that merge embeddings of example entities or probes with embeddings of type definitions, and assign NER labels to tokens based on their similarity to these prototypes.

While the above models use probing either with prompting or focusing on similarity between candidate entities and type definition, in our work, probing is used to leverage the sentence context, which serves as the primary signal for determining NER labels.

2.2. Latin Named Entity Recognition

Research on Latin NER has explored a variety of approaches, ranging from traditional sequence-labelling techniques such as CRF and rule-based methods (Erdmann et al., 2016; Chastang et al., 2021; Burns, 2023; Novotny et al., 2023) to multilingual or Latin-pretrained transformer-based models (Torres Aguilar, 2022; Beersmans et al., 2023). Building on this trend, in this work, we define NER models based on Latin-pretrained transformers, since pretraining on the target language allows the model to better capture the morphology and syntax of that language than multilingual models, especially when sufficient monolingual data is available (Rust et al., 2020).

3. Methodology

3.1. Preliminaries

Pseudo-log-likelihood for Word Scoring To compute the compatibility between a word and its context, we use the Pseudo-Log-Likelihood (PLL) metric defined by Salazar et al. (2020). Given a sentence s and a word at position i, y_i, with the surrounding context $y_{\setminus i}$, the PLL is defined as

$$\text{PLL}(y_i \mid y_{\setminus i}) = -\log P(y_i \mid y_{\setminus i}). \quad (1)$$

Probes as Candidate Entities Given that our NER task involves K entity types (e.g., “place” or “person”), we define, for each type k ($k \in K$), a set of m entities, referred to as probes. To evaluate the compatibility of these probes with the sentences, it is important to select well-known or representative entities.

3.2. PLL-Based Entity Compatibility Scoring

Our hypothesis is that language models naturally encode NER-like knowledge due to their pretraining on a language modeling objective, which facilitates the learning of rich contextual representations between words. Consequently, even without any NER fine-tuning, a pretrained language model can already provide an indication of whether a word is likely to be an entity based on its surrounding context.

Problem Definition. Let s be a sentence composed of n words $(y_1, \ldots, y_n)$, and let $\mathcal{E}_k$ denote a set of m probe entities associated with entity type k. Our goal is to assign a compatibility score to each word position i for each entity type k, indicating whether the word y_i behaves like an entity of that type.

Contextual Compatibility Scoring For each word position i, we compute the compatibility score $comp_i(k)$ with respect to entity type k using PLL scores between the surrounding context of position i, denoted $y_{\setminus i}$, and the probe entities in $\mathcal{E}_k$:

$$comp_i(k) = \frac{1}{m} \sum_{e \in \mathcal{E}_k} \text{PLL}(e \mid y_{\setminus i}) \quad (2)$$

Here, $\text{PLL}(e \mid y_{\setminus i})$ measures how well the substitution of y_i with the probe entity e fits the sentence according to the pretrained language model, exploiting the contextual representations induced during pretraining.

After computing the scores $comp_i(k)$ for all K entity types, each word position i is associated with K compatibility scores, indicating its affinity to each entity type. Finally, we normalize these scores to prevent differences in scale across entity types from biasing the NER decision.

3.3. Integration with NER Model

The contextual compatibility scoring presented above has the advantage of being able to operate as an independent NER system without any fine-tuning. An example of this scoring method is shown in Table 1 for the entity types “person" and “place". While the method correctly identifies entities such as “Cicero" as a person and “Pompeiumm" as a

location, it can also make errors, as in the case of the word “habuit" Therefore, if annotated NER training data are available, we recommend integrating this scoring as an auxiliary signal during the fine-tuning of language models for NER. Specifically, given the logits $logits_i(k)$ computed by the language model for word y_i and entity type k during NER training, we adjust these logits based on the contextual compatibility scores as follows:

$$logits_i(k) = \begin{cases} logits_i(k) + \alpha \cdot comp_i(k), \\ \quad \text{if } softmax(logits_i(k)) < \tau, \\ logits_i(k), \text{otherwise.} \end{cases} \tag{3}$$

Word	Person	Place
Cicero	0.89	0.44
in	0	0
Roma	0.28	0.81
orationem	0.87	0.58
habuit	0.84	0.66
et	0	0
Pompeium	0.70	0.73
salutavit	0.1	0.04

Table 1: Contextual Compatibility Scoring Method used as a NER model for the sentence: “Cicero in Roma orationem habuit et Pompeium salutavit".

Here, α controls the strength of the contextual compatibility scoring, and τ is a threshold that ensures the scoring is applied only in uncertain or difficult cases, i.e., when the NER model is not confident about the entity type of a word. Since the NER knowledge learned from a specialized training dataset is typically more reliable than relying solely on contextual similarity learned during the pretraining of the language models, the threshold τ prevents the compatibility scores from overriding the model’s confident predictions.

4. Experimental Setup

Label	[1]	[2]	[3]	[4]	[5]
O	90.45	92.05	88.59	98.24	92.42
“person"	5.74	5.19	6.42	1.69	3.05
“place"	3.63	1.79	1.50	0.07	2.93
“collective"	0.00	0.97	3.49	0.00	1.25
“work"	0.00	0.00	0.00	0.00	0.26
“time"	0.00	0.00	0.00	0.00	0.09
“object"	0.00	0.00	0.00	0.00	0.01

Table 2: Percentage distribution of NER labels across the five datasets used for the training of the NER model.

Datasets To train the NER models, we rely on five datasets whose distribution of NER labels is presented in Table 2. The first dataset [1] was introduced by Torres Aguilar (2022) and contains Latin NER annotations for manuscripts written between the 10th and 15th centuries. The second dataset [2], presented by Beersmans et al. (2023), contains annotations for Latin texts written between the 2nd century BCE and the 2nd century CE. The third dataset [3], described by Erdmann et al. (2016), consists of Latin texts written between the 1st century BCE and the early 2nd century CE. The fourth dataset [4] is an internal resource that will soon be released as part of our work at STUDIUM.AI. It contains NER annotations for Latin student notes written between the 16th and 18th centuries.

Since the first four datasets described above contain annotations for only three of the ten NER categories defined in the LT4HALA evaluation (“person",“place", and “collective"), we extend the set of available NER labels by creating a fifth dataset [5]. This dataset is generated by translating existing HIPE datasets (Ehrmann et al., 2022) in French, English, and German into Latin using the Google Translate API. To propagate the NER labels from the source texts to the Latin translations, we first extract the entities in the source language—hereafter referred to as source entities—and then apply the following two-step alignment procedure:

- First, we translate the source entities separately from the texts into Latin. The Latin NER labels of the first step are obtained by matching the translated entities to words in the translated Latin texts.
- In the second step, for each unaligned source entity from the first step and each word in the translated Latin text, we compute the alignment score below:

$$align = \beta \cdot \cos_{sim}(emb_{src}, emb_{Latin}) + (1 - \beta) \cdot Lev_{dist}(w_{src}, w_{Latin}) \tag{4}$$

  where:

  - emb_{src} is the language-agnostic BERT (LaBERT) embedding (Feng et al., 2022) of the remaining unaligned source entity,
  - emb_{Latin} is the LaBERT embedding of a candidate word in the translated Latin text,
  - $\cos_{sim}$ is the cosine similarity between embeddings,
  - Lev_{dist} is the Levenshtein distance between the remaining unaligned source entity w_{src} and the sentence word w_{Latin}.
  - β is a weighting factor balancing semantic similarity and orthographic similarity.

Here, $\cos_{\sim}$ measures the semantic similarity, while Lev_{dist} measures the orthographic similarity. Each

remaining unaligned source entity from the first step is assigned to the word in the Latin translation with the highest $score$ exceeding the threshold ϕ. These assignments constitute the Latin NER labels produced in the second step of our procedure. The final NER labels for the translated Latin texts are obtained by merging the labels from both steps of the alignment process.

Implementation details All our experiments are performed on an NVIDIA GeForce RTX 3090 GPU. The language models are trained for 600 epochs, with a learning rate of 5×10^{-5} and a batch size of 16. The hyperparameters of our method are set as follows: $\alpha = 1$, $\tau = 0.7$, $\beta = 0.9$, and $\phi = 0.97$.

Models To evaluate our PLL-based entity compatibility scoring, we use LaBERTa (Riemenschneider and Frank, 2023), a RoBERTa-based (Liu et al., 2019) encoder pretrained on Latin corpora. In Section 5 dedicated to the experiments, our LaBERTa-based NER model with the proposed PLL method is referred to as "KULeuven," which is the name of our team for the LT4HALA competition. For comparison, we also report results using LatinBERT (Bamman and Burns, 2020). Both LaBERTa and LatinBERT are encoder-only transformers suitable for token-level tasks such as NER. Preliminary experiments showed that LaBERTa outperforms LatinBERT (79.45% vs. 74.24% F1 on a 90/10 train/validation split, both integrated with PLL scoring), so only the LaBERTa-based results were submitted to the LT4HALA competition.

Metrics The NER results are evaluated using precision, recall, and F1 scores under two settings: *strict* and *fuzzy*. The *strict* setting requires an exact match of both entity type and boundaries, while the *fuzzy* setting requires the correct entity type with at least one work overlap.

5. Experiments

5.1. LT4HALA Tasks

We evaluate our model on the two shared tasks of the LT4HALA competition.

Task 1 – NERC Coarse-grained: Recognize top-level entity categories (i.e., entities like "person" and "place"). Since annotated NER training data is available, labels were generated using LaBERTa fine-tuned on the data and combined with our PLL-based entity compatibility scoring. Our model, referred to as "KULeuven", the name of our team for the LT4HALA competition, ranked second among competitors (Table 3).

Task 2 – NERC Fine-grained: Recognize top- and sub-level entity categories (e.g., "person.entity", "person.ancesty", or "collective.organization"). For this task, no training data is available. Instead, we define lists of probes for all pairs of top- and sub-level entities and use them to assign NER labels via the PLL-based scoring method, without integrating a language model. While our method ranks third, it is a zero-shot approach that requires no fine-tuning and relies solely on the knowledge encoded in the pretrained LaBERTa model.

Evaluation	System	P	R	F1
strict	uOttawa	0.899	0.917	0.908
	KULeuven	0.736	0.694	0.714
	argo-navis	0.575	0.498	0.534
fuzzy	uOttawa	0.932	0.950	0.941
	KULeuven	0.794	0.749	0.771
	argo-navis	0.635	0.550	0.590

Table 3: Task 1 – NERC Coarse-grained. Strict and fuzzy comparison between the results of our team (KULeuven) and the results of the other two teams (uOttawa and argo-navis). For each team, we present the best obtained results. The results of our team are obtained by integrating the proposed PLL-based method into LaBERTa.

Evaluation	System	P	R	F1
strict	uOttawa	0.841	0.890	0.865
	argo-navis	0.466	0.327	0.384
	KULeuven	0.131	0.128	0.129
fuzzy	uOttawa	0.862	0.913	0.887
	argo-navis	0.523	0.367	0.432
	KULeuven	0.141	0.138	0.140

Table 4: Task 1 – NERC Fine-grained. Strict and fuzzy comparison between the results of our team (KULeuven) and the results of the other two teams (uOttawa and argo-navis). For each team, we present the best obtained results. The NER labels generated by our model are extracted using only our PLL-based scoring method with no fine-tuning.

5.2. Effect of the PLL-based Method on NER Performance

While the results submitted to LT4HALA rely on LaBERTa due to its strong NER performance on Latin texts, this subsection examines in more detail the general impact of the proposed PLL-based method on the NER capabilities of Latin models, including not only LaBERTa but also LatinBERT.

In Table 5, we compare LaBERTa and LatinBERT integrated with our PLL-based scoring method for Task 1 of the LT4HALA competition. We also report results for the zero-shot setting, where NER labels are obtained solely using our PLL-based entity compatibility scoring method computed from the pretrained LaBERTa or LatinBERT models; these

results are marked with * in the table. Additionally, we include results for LaBERTa and LatinBERT trained on the NER data without applying the PLL-based method.

As expected, adding our PLL-based method during the decoding process of NER fine-tuning improves performance for both Latin language models. Among the evaluated setups, LaBERTa combined with PLL (referred to as “KULeuven” in Table 3) achieves the best overall performance. Notably, the zero-shot approach remains competitive, with an F1 score exceeding 0.304 under strict evaluation and 0.338 under fuzzy evaluation, despite the absence of NER fine-tuning.

Evaluation	System	P	R	F1
strict	PLL+LaBERTa	0.736	0.694	0.714
	PLL+LatinBert	0.679	0.652	0.665
	PLL+LaBERTa*	0.336	0.387	0.359
	PLL+LatinBERT*	0.285	0.325	0.304
	LaBERTa	0.658	0.631	0.644
	LatinBERT	0.612	0.624	0.618
fuzzy	PLL+LaBERTa	0.794	0.749	0.771
	PLL+LatinBert	0.713	0.684	0.697
	PLL+LaBERTa*	0.384	0.424	0.403
	PLL+LatinBERT*	0.341	0.335	0.338
	LaBERTa	0.724	0.704	0.714
	LatinBERT	0.677	0.635	0.656

Table 5: Task 1 – NERC Coarse-grained. Strict and fuzzy NER performance of LaBERTa and LatinBERT with PLL. PLL+LaBERTa is refered as “KULeuven" in Table 3. Results marked with * correspond to the zero-shot setting with no fine-tuning. Additionally, we report results for the fine-tuned LaBERTa and LatinBERT on the NER data without the PLL-based method.

In Table 6, we compare our PLL-based scoring method applied in a zero-shot framework on top of LaBERTa and LatinBERT for Task 2 of the LT4HALA competition. The PLL-based method is applied as a standalone NER model due to the lack of NER training data for this task. Despite the large number of probes used to differentiate all pairs of top- and sub-level entity categories and the absence of training data, the PLL-based method produces competitive results on both models, with LaBERTa performing slightly better.

Evaluation	System	P	R	F1
strict	PLL+LaBERTa*	0.131	0.128	0.129
	PLL+LatinBERT*	0.109	0.102	0.105
fuzzy	PLL+LaBERTa*	0.141	0.138	0.140
	PLL+LatinBERT*	0.135	0.118	0.126

Table 6: Task 1 – NERC Fine-grained. Strict and fuzzy NER performance of LaBERTa and LatinBERT with PLL in the zero-shot setting (*). PLL+LaBERTa is refered as “KULeuven" in Table 4.

6. Conclusion

We introduce a novel approach to better exploit the implicit NER knowledge acquired by language models during pretraining and demonstrate its effectiveness for Latin. Our method can be integrated into NER models to enhance fine-tuning, or used as a standalone zero-shot approach when no annotated data is available. It achieved second place in coarse-grained NER and third in fine-grained NER at the LT4HALA workshop, highlighting that pretrained models encode rich entity knowledge that can be leveraged effectively in low-resource NER scenarios.

7. Acknowledgements

This work is part of the Studium.AI project funded by FW0 I004022N.

8. Bibliographical References

David Bamman and Patrick J. Burns. 2020. Latin bert: A contextual language model for classical philology. *ArXiv*, abs/2009.10053.

Marijke Beersmans, Evelien de Graaf, Tim Van de Cruys, and Margherita Fantoli. 2023. Training and evaluation of named entity recognition models for classical Latin. In *Proceedings of the Ancient Language Processing Workshop*, pages 1–12. INCOMA Ltd., Shoumen, Bulgaria.

Patrick J. Burns. 2023. Latincy: Synthetic trained pipelines for latin nlp. *ArXiv*, abs/2305.04365.

Pierre Chastang, Sergio Torres Aguilar, and Xavier Tannier. 2021. A named entity recognition model for medieval latin charters. *Digit. Humanit. Q.*, 15.

Alessio Cocchieri, Marcos Martínez Galindo, Giacomo Frisoni, Gianluca Moro, Claudio Sartori, and Giuseppe Tagliavini. 2025. Zeroner: Fueling zero-shot named entity recognition via entity type descriptions. In *Annual Meeting of the Association for Computational Linguistics*.

Maud Ehrmann, Matteo Romanello, Antoine Doucet, and Simon Clematide. 2022. Introducing the hipe 2022 shared task: Named entity recognition and linking in multilingual historical documents. In *Advances in Information Retrieval*, pages 347–354. Springer International Publishing.

Alexander Erdmann, Christopher Brown, Brian Joseph, Mark Janse, Petra Ajaka, Micha Elsner,

and Marie-Catherine de Marneffe. 2016. Challenges and solutions for Latin named entity recognition. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 85–93. The COLING 2016 Organizing Committee.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891. Association for Computational Linguistics.

Guochao Jiang, Zepeng Ding, Yuchen Shi, and Deqing Yang. 2024. P-icl: Point in-context learning for named entity recognition with large language models. *ArXiv*, abs/2405.04960.

Yueqing Liang, Liangwei Yang, Chen Wang, Xiongxiao Xu, Philip S. Yu, and Kai Shu. 2024. Taxonomy-guided zero-shot recommendations with llms. *ArXiv*, abs/2406.14043.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Vit Novotny, Kristina Luger, Michal Štefánik, Tereza Vrabcova, and Ales Horak. 2023. People and places of historical Europe: Bootstrapping annotation pipeline and a new corpus of named entities in late medieval texts. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14104–14113. Association for Computational Linguistics.

Frederick Riemenschneider and Anette Frank. 2023. Exploring large language models for classical philology. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL'23)*. Association for Computational Linguistics, Toronto, Canada.

Phillip Rust, Jonas Pfeiffer, Ivan Vulic, Sebastian Ruder, and Iryna Gurevych. 2020. How good is your tokenizer? on the monolingual performance of multilingual language models. *ArXiv*, abs/2012.15613.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2699–2712. Association for Computational Linguistics.

Or Shachar, Uri Katz, Yoav Goldberg, and Oren Glickman. 2025. Ner retriever: Zero-shot named entity retrieval with type-aware embeddings. In *Conference on Empirical Methods in Natural Language Processing*.

Sergio Torres Aguilar. 2022. Multilingual named entity recognition for medieval charters using stacked embeddings and bert-based models. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 119–128. European Language Resources Association.

Mozhi Zhang, Hang Yan, Yaqian Zhou, and Xipeng Qiu. 2023. Promptner: A prompting method for few-shot named entity recognition via k nearest neighbor search. *ArXiv*, abs/2305.12217.

Classificatio Sine Iactu – That Is, Zero-Shot NERC in Latin

Luisa Ripoll-Alberola[1,3], **Fernando Nicolás-Flores**[2], **Francisco Javier Muñoz Acebes**[3]
[1]Computational Humanities, Leipzig University, [2]Universidad de Alicante,
[3] Filología Digital, Universidad de Valladolid
Augustusplatz 10, 04109 Leipzig, Germany
ripoll_alberola@informatik.uni-leipzig.de, fernando.nicolas@ua.es, fjavier.munoz@uva.es

Abstract

This paper presents a zero-shot approach to Named-Entity Recognition and Classification (NERC) in Latin, applied to the EvaLatin shared task. Given the novelty and granularity of the annotation guidelines, which preclude the use of existing annotated resources, we employ the zero-shot model GLiNER2, a general information extraction system capable of CPU-efficient inference, within a cross-lingual pipeline. Latin texts are first translated into English via the Google Translate API, processed by the model, and the resulting annotations are aligned back to the original Latin using word-alignment techniques. Rule-based post-processing addresses labelling inconsistencies and low-confidence predictions. We evaluate two model variants, a large monolingual and a multilingual model, under both strict and fuzzy evaluation. The large model delivers the best results for the coarse-grained task (F1: 0.590 fuzzy), while the multilingual model outperforms it on the fine-grained task (F1: 0.432 fuzzy). Results indicate that multilingual embeddings confer an advantage for fine-grained semantic distinctions, that English embeddings introduce systematic bias in cross-lingual transfer, and that zero-shot NER represents a viable, reproducible baseline for low-resource historical languages. Fine-tuning on guideline-compliant annotated data remains a priority for future work.

Keywords: Named-Entity Recognition, Zero-Shot Learning, Latin, GLiNER.

1. Introduction

When we first read the annotation guidelines proposed by the EvaLatin coordinating team, we were impressed by their precision and ambition. The proposed schema does not merely identify broad categories of named entities; it articulates a remarkably fine-grained ontology capable of capturing subtle distinctions within historical texts. Such granularity raises a methodological challenge: how can a Named-Entity Recognition (NER) model efficiently apply this schema in a language for which annotated resources remain scarce?

Rather than normalising existing annotations to previous guidelines and fine-tuning a model, the present study explores zero-shot learning. Zero-shot models generalise to new entity types without training and achieve F1 scores of approximately 0.6 on a range of tasks in English (Alhoshan et al., 2023, Del Moral-González et al., 2025, Marchitan et al., 2025), a performance level sufficient to bootstrap silver-standard corpora. This property is of particular relevance for Latin, where philologically informed manual annotation is costly. The proposed approach therefore integrates a zero-shot model into a reusable pipeline[1].

Prior work has applied LLMs in zero-shot or few-shot settings to historical documents (Hiltmann et al., 2025, Zhang and Colavizza, 2025), including studies on classical languages (Akavarapu et al., 2025), and the reception of antiquity in modern languages (Poibeau, 2024, Ripoll-Alberola and Burghardt, 2025). Performance on CPU-executable systems remains unexplored, despite its importance for reproducibility in resource-constrained projects in the humanities. For this reason, we adopt GLiNER2 (Zaratiana et al., 2025), which extends the original GLiNER model (Zaratiana et al., 2023) to become a general information extraction system. GLiNER2 achieves performance comparable to GPT-4o in NER, operates 2.6 times faster on CPU, and is openly available[2].

The GLiNER family of architectures concatenates task specifications with the input text, computes contextualised representations jointly, and refines them through feedforward networks to produce task-specific and span embeddings. A matching score (dot-product + sigmoid) is then computed for each span–task pair, producing a probability that a given span corresponds to a given category. GLiNER2 extends this framework with a richer special token vocabulary, enabling NER, hierarchical structured extraction, and text classification within a single forward pass, whereas the original GLiNER was NER-only. For NER specifically, however, the underlying mechanism remains essentially identical. GLiNER2 also improved NER performance over the original model[3].

[1]All code is documented and available in: https://github.com/luisarip/classificatio_sine_iactu/

[2]https://huggingface.co/collections/fastino/gliner2-family

[3]We may refer to the model by the general name of

Proceedings of LT4HALA 2026 @ LREC 2026, pages 443–447.
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

Whether these advantages – computational accessibility, label flexibility, and the elimination of task-specific pretraining – translate into sufficient performance for low-resource historical languages is the question this study addresses.

GLiNER relies on pretrained bidirectional transformer encoders (DeBERTa or mDeBERTa) as its backbone. Since these embeddings cannot be replaced with Latin-specific representations without full retraining, a cross-linguistic approach is adopted, following Soffiantini (2024): the model is applied to an English translation, and identified entities are aligned back to the original Latin text.

2. Description of the system

These are the main steps of our system:

a. Translation LAT>ENG.

b. GLiNER call.

c. Rule-based post processing.

d. Alignment ENG>LAT.

e. Projection of the results back to the original TSV.

In step (a), sentence boundaries are identified using the "EndOfSentence" marker in the MISC column, and sentences are passed individually to the translation engine. Preliminary checks demonstrated that longer, entity-dense passages caused the model to miss named entities. Sentence-level inputs show consistent performance over varying sentence lengths and were therefore adopted.

For translation, the deep-translator Python package[4] was used to query the Google Translator API, with an inter-request delay to avoid rate limiting. Translation outputs were cached to avoid redundant API calls. This was considered reliable given the deterministic behaviour of the Google Translate API.

Two model variants were compared: the multilingual[5] and the monolingual large[6]. Coarse and fine-grained labels were defined in a separate JSON file, selected at runtime via the `taxonomy_name` argument. All tag names were lowercased. The GLiNER architecture accepts an optional natural-language description per tag; accordingly, the descriptions provided in the annotation guidelines were adapted for zero-shot use by removing examples and converting singular forms to plural where the category was "collective". NER is used rather than hierarchical structured extraction, as it does not perform adequately for this task.

Development on the sample data revealed two recurring issues: labelling inconsistencies among occurrences of the same entity (e.g., *Iuno* [the goddess Juno] was classified as both person and creature), and a substantial proportion of low-confidence false positives. A rule-based post-processing step was introduced: labels are normalised by majority vote, and predictions falling below a confidence threshold of 0.65 are discarded.

These results, obtained on the English text, are aligned back to the original Latin text. The alignment is performed using simalign (Jalili Sabet et al., 2020), a word alignment method which uses static and contextualised embeddings, meaning that parallel training is not required. We run simalign using the embeddings of UGARIT/grc-alignment (Yousef et al., 2022b), an XLM-RoBERTa-based alignment model trained with Ancient Greek-Latin and Ancient Greek-English pairs. These embeddings generalise well to Latin-English alignment despite the absence of training for this language pair (Yousef et al., 2022a). English-Latin alignments are computed and projected onto the original TSV file, converted to BIO format, and subjected to a final rule-based check: tokens in lowercase, assumed to be common nouns rather than named entities, are filtered out and relabelled as O. This mechanism was not always enforced. In the sample data, for instance, *populi Romani* [the Roman people] was labelled as B-COLLECTIVE/.ETHNIC, I-COLLECTIVE/.ETHNIC. The current function requires at least one uppercased token, but does not exclude cases in which additional non-entity tokens are assigned entity labels, as in *Albanique patres* [Alban fathers], were only *Albanique* bears the COLLECTIVE.ETHNIC label.

3. Results

Table 1 presents aggregate results. Per-entity results for the best-performing model in each task are also given: the large model for the coarse-grained task (Table 2) and the multilingual model for the fine-grained task (Table 3). The large model achieves higher precision for all configurations, while the multilingual model delivers higher recall.

Notably, in the NERC coarse task, the categories benefiting most from fuzzy evaluation are LANGUAGE (which doubles in performance with the multilingual model, 0.667/1.000/0.800, and improves in 50.1% with the large model, 1.000/0.750/0.857) and WORK (which reaches 0.500/1.000/0.667 with the multilingual model under fuzzy evaluation). In the NERC fine task, LANGUAGE similarly improves under fuzzy evaluation

"GLiNER". However, we are always applying GLiNER2, the most recent version.

[4] https://github.com/nidhaloff/deep-translator

[5] https://huggingface.co/fastino/gliner2-multi-v1

[6] https://huggingface.co/fastino/gliner2-large-v1

Task	Evaluation	Model	Precision	Recall	F1 Score
Task 1: NERC Coarse	Strict	Multilingual	0.552	**0.500**	0.525
		Large	**0.575**	0.498	**0.534**
	Fuzzy	Multilingual	0.619	**0.561**	0.589
		Large	**0.635**	0.550	**0.590**
Task 2: NERC Fine	Strict	Multilingual	0.466	**0.327**	**0.384**
		Large	**0.491**	0.218	0.302
	Fuzzy	Multilingual	0.523	**0.367**	**0.432**
		Large	**0.534**	0.237	0.328

Table 1: Performance over all entities in the test sample for Task 1 (NERC Coarse) and Task 2 (NERC Fine). Results are reported using strict and fuzzy evaluation metrics. Highlighted, the best values per evaluation.

(0.800/1.000/0.889). WORK is not correctly identified by any model, unlike in the coarse task.

Performance drops between the sample (validation) data and the test set, suggesting sensitivity to textual complexity, entity density, and entity type. The *Metamorphoses*, for instance, contains a high density of mythological figures lacking a dedicated label, which may introduce ambiguity for the model. On the sample data, the large model achieved average F1 of 0.66 (COLLECTIVE 0.68, EVENT 0.67, PERSON 0.67, PLACE 0.72).

4. Discussion

The principal finding is that multilingual embeddings improve performance on the fine-grained NERC task for Latin texts, even if the input is in English. In the coarse-grained task, the large model achieves marginally higher precision at the cost of a steeper drop in recall, making the two models comparable. The multilingual model's stronger recall, more pronounced in the fine task, likely reflects residual sensitivity to Latin morphology and vocabulary, given that mDeBERTa was trained on 390 million Latin tokens (Conneau et al., 2020).

Persons and locations are the best-identified categories, which is expected, given (i) their prevalence in most NER benchmarks, and (ii) overlap between Latin toponyms and proper names and their English or Romance-language equivalents. Nevertheless, some mistakes persist: for instance, as "Troy" is a given name in English, the city of Troy is many times labelled as PERSON, even when the Latin form *Troiae* is provided as input. It can thus be concluded that English exerts a disproportionate influence on the embedding space, even in multilingual models.

LANGUAGE benefits most from fuzzy evaluation: language mentions typically appear as multi-token expressions (e.g., *Graecis litteris* [Greek letters], *lingua Britannicae* [in British language]), but the model retrieves only the capitalised adjective (*Graecis*, *Britannicae*).

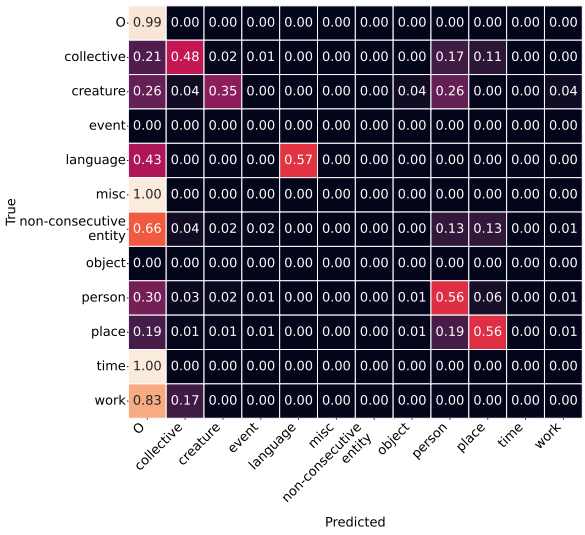


Figure 1: Normalised confusion matrix for NERC coarse task with large model.

Evaluation using seqeval (Nakayama, 2018) produces the normalised confusion matrix in Figure 1. Multitoken entities are predominantly categorised as one-token entities, thereby explaining the high prevalence of O among entities. Two problems recur in the classification of person entities. First, many mythological figures such as Iuno are sometimes misclassified as CREATURE, bypassing the anthropomorphic reading. This points to a potential gap in the tagset: a PERSON.MYTH subclass would guide the model in such cases, and serve annotation tasks focused on mythological rather than historical persons. Second, COLLECTIVE.ETHNIC, PLACE.DERIVATIVE and PERSON.ETHNIC are highly specific categories, all three of which are derived adjectives with toponymical roots. While they offer annotators the opportunity to capture nuances, they also present under-

Entity Type	P	R	F1	TP	FP	FN	F1 Fuzzy
ALL	0.575	0.498	0.534	1327	982	1338	0.590
COLLECTIVE	0.555*	0.467*	0.507*	197	158	225	0.533
CREATURE	0.320*	0.348*	0.333*	8	17	15	0.333
LANGUAGE	0.667	0.500	0.571	2	1	2	0.857
PERSON	0.633	0.515	0.568	262	152	247	0.628
PLACE	0.604	0.504	0.549	858	563	846	0.614

Table 2: Performance (strict evaluation) for the NERC Coarse task using the best-performing model *gliner2-large-v1*. Entity types EVENT, MISC, OBJECT, TIME and WORK were omitted because precision, recall, and F1 were zero. Entity types marked with an asterisk (*) were better identified by the multilingual model: COLLECTIVE (0.729/0.645/0.684) and CREATURE (0.389/0.609/0.475). The last column reports the F1 score obtained using fuzzy evaluation.

Entity Type	P	R	F1	TP	FP	FN	F1 Fuzzy
ALL	0.466	0.327	0.384	872	999	1793	0.432
COLLECTIVE.ETHNIC	0.167	0.105	0.129	41	204	350	0.142
CREATURE	0.500	0.294	0.370	5	5	12	0.370
LANGUAGE	0.400*	0.500	0.444*	2	3	2	0.889
PERSON	0.652*	0.365*	0.468*	133	71	231	0.518
PERSON.AUTHOR	0.147*	0.109*	0.125*	5	29	41	0.200
PLACE	0.605	0.417	0.494	686	447	961	0.554

Table 3: Performance (strict evaluation) for the NERC Fine task using the best-performing model *gliner2-multi-v1*. The categories COLLECTIVE, COLLECTIVE.ANCESTRY, COLLECTIVE.DERIVATIVE, CREATURE.ANIMAL, CREATURE.ASTRONOMY, EVENT, MISC, OBJECT, PERSON.ANCESTRY, PERSON.DERIVATIVE, PERSON.EPITHET, PERSON.ETHNIC, PLACE.ASTRONOMY, PLACE.DERIVATIVE, TIME and WORK were omitted because precision, recall and F1 were zero. Entity types marked with an asterisk (*) were better identified by the large model: LANGUAGE (0.667/0.500/0.571), PERSON (0.662/0.393/0.493) and PERSON.AUTHOR (0.400/0.304/0.346). The last column reports the F1 score obtained using fuzzy evaluation.

standing challenges for LLMs. The translation step is crucial in the differentiation of these three fields, see an example below:

> *Progeniem sed enim Troiano a sanguine duci audierat* (Verg. A. 1.19)
> Troiano → place.derivative
>
> *but anxiously she heard that of the Trojan blood there was a breed then rising* (trans. Williams 1910)
>
> *But he had heard of the race of the Trojans, led by blood* (Google Translate)
> Trojans → collective

When analysing the fine-grained task, a broad pattern is observed: the model strongly favours first-level tags, which are more frequent in the data, and rarely commits to second-level distinctions. As a result, the model has an acceptable performance but completely misses its purpose. PERSON.AUTHOR, PERSON.DERIVATIVE, PLACE.DERIVATIVE, CREATURE.ANIMAL, and CREATURE.ASTRONOMY are largely collapsed into their parent categories or labelled as O. Among subcategories, only COLLECTIVE.ETHNIC and PERSON.AUTHOR generate true positives, and PERSON is the only category with true positives in more than two tags, given that all collectives are labelled as COLLECTIVE.ETHNIC. Therefore, the model retains coarse-grained knowledge but does not generalise to finer distinctions.

5. Conclusion

Although performance ranged approximately between 0.4 and 0.6 over entities and models, zero-shot NER with GLiNER may suffice for simpler Latin texts. Multilingual models are a better option, even in a cross-linguistic setting. The primary limitation is the impossibility of applying Latin-specific embeddings; conversely, the ability to process Latin texts without any annotated data represents a direction worth further investigation. All code has been

made available on GitHub to ensure reproducibility in Latin texts of choice. Future work should address three directions: replicating these experiments with higher-quality Latin–English translation; adapting the annotation guidelines for better understanding of the transformer model; and fine-tuning the GLiNER model on additional annotated data following the guidelines.

6. Acknowledgements

Luisa Ripoll-Alberola acknowledges funding from the European Union's Horizon Europe Research and Innovation Programme through the Marie Sklodowska-Curie Actions Doctoral Network MECANO, Grant Agreement No. 101120349.

7. Bibliographical References

V.S.D.S.Mahesh Akavarapu, Hrishikesh Terdalkar, Pramit Bhattacharyya, Shubhangi Agarwal, Dr. Vishakha Deulgaonkar, Chaitali Dangarikar, Pralay Manna, and Arnab Bhattacharya. 2025. A case study of cross-lingual zero-shot generalization for classical languages in llms. In *Findings of the Association for Computational Linguistics: ACL 2025*, page 2745–2761, Vienna, Austria. Association for Computational Linguistics.

Waad Alhoshan, Alessio Ferrari, and Liping Zhao. 2023. Zero-shot learning for requirements classification: An exploratory study.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Rodrigo Del Moral-González, Helena Gómez-Adorno, and Orlando Ramos-Flores. 2025. Comparative analysis of generative llms for labeling entities in clinical notes. *Genomics Informatics*, 23(1):3.

Torsten Hiltmann, Martin Dröge, Nicole Dresselhaus, Till Grallert, Melanie Althage, Paul Bayer, Sophie Eckenstaler, Koray Mendi, Jascha Marijn Schmitz, Philipp Schneider, Wiebke Sczeponik, and Anica Skibba. 2025. Ner4all or context is all you need: Using llms for low-effort, high-performance ner on historical texts. a humanities informed approach.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1627–1643, Online. Association for Computational Linguistics.

Teodor-George Marchitan, Claudiu Creanga, and Liviu P. Dinu. 2025. Few-shot text-based emotion detection.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Thierry Poibeau. 2024. Annotating references to mythological entities in french literature. (arXiv:2412.18270). ArXiv:2412.18270 [cs].

Luisa Ripoll-Alberola and Manuel Burghardt. 2025. Llm-based approaches to canonical reference extraction in academic texts: Initial results. *AII-EEKE 2025 Joint Workshop of the 5th AI + Informetrics (AII) and the 6th Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE), in 20th International Conference on Scientometrics Informetrics (ISSI)*.

Laura Soffiantini. 2024. Cross-Linguistic Annotation Transfer in Geoparsing Experiments with Classical Texts. *DH Benelux Journal*, (6):155–168.

Tariq Yousef, Chiara Palladino, Farnoosh Shamsian, Anise d'Orange Ferreira, and Michel Ferreira dos Reis. 2022a. An automatic model and gold standard for translation alignment of Ancient Greek. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5894–5905, Marseille, France. European Language Resources Association.

Tariq Yousef, Chiara Palladino, David J. Wright, and Monica Berti. 2022b. Automatic translation alignment for ancient greek and latin. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 101–107, Marseille, France. European Language Resources Association.

Urchade Zaratiana, Gil Pasternak, Oliver Boyd, George Hurn-Maloney, and Ash Lewis. 2025. Gliner2: An efficient multi-task information extraction system with schema-driven interface.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2023. Gliner: Generalist model for named entity recognition using bidirectional transformer.

Shibingfeng Zhang and Giovanni Colavizza. 2025. Named entity recognition of historical text via large language model.

Extending `omnes flores` for the EvaLatin 2026 Dependency Parsing Tasks

Hiroshi Matsuda, Masayuki Asahara
Megagon Labs, Tokyo, Recruit Co., Ltd., National Institute for Japanese Language and Linguistics
Marunouchi 1-9-2, Chiyoda, Tokyo, Japan, Midorichou 10-2, Tachikawa, Tokyo, Japan
hiroshi_matsuda@megagon.ai, masayu-a@ninjal.ac.jp

Abstract

`omnes flores` is an NLP framework based on Universal Dependencies (UD) that utilizes multilingual Large Language Models (LLMs), and its default model is trained on data from 40 UD languages comprising 40 treebanks. For the EvaLatin 2026 Dependency Parsing Tasks, we extended the training data of `omnes flores` by incorporating six public Latin treebanks from UD and trained a dependency parsing model using the extended training data. The dependency parser of `omnes flores` normally takes a list of word FORM values as input. However, since the EvaLatin 2026 test data includes an UPOS column, we investigated whether incorporating both FORM and UPOS during both training and inference could improve parsing accuracy. Our experiments show that training using both FORM and UPOS improves performance by 0.5-1.0 LAS points on Prose compared with training using only FORM, but decreases performance by 5 points on Poetry.

Keywords: Universal Dependencies, dependency parsing, LoRA SFT, EvaLatin

1. Introduction

In recent years, Large Language Models (LLMs) have improved to the point that they can be applied to tasks that were traditionally handled by Natural Language Processing (NLP) techniques. Tuning LLMs for specific tasks or domains has become easier in recent years. This is largely due to few-shot prompting and the spread of Parameter-Efficient Fine-Tuning (PEFT) techniques used for supervised fine-tuning (SFT) on instruction—response pairs. As a result, end users can increasingly treat LLMs as black boxes and apply them to a wide range of downstream tasks.

Quantitative analysis and NLP technology. Because LLMs are large models, inference is computationally expensive, and LLMs by themselves perform poorly on quantitative analysis of text (they struggle even with simple tasks such as word-frequency statistics). As a result, the use of LLMs for mining tasks over large-scale text collections has not progressed very far, and conventional NLP techniques continue to be used for this kind of task.

Dependency parsing is often used not only in computational linguistics research but also in industrial applications such as text mining, and Universal Dependencies (UD) (Nivre et al., 2016, 2020) is widely used for training and evaluating dependency parsing models. UD began to be developed around 2015, and the latest version at the time of writing (r2.17) supports 186 languages.

Sustainability of NLP technology. Most open-source NLP frameworks in active use today were designed before the widespread adoption of LLMs. They mainly implement functional components such as part-of-speech tagging and dependency parsing with encoder-based models, and assemble those components into a processing pipeline. Developing and maintaining such conventional NLP pipelines requires advanced specialized knowledge and engineering skills. However, most market demand has already shifted from conventional NLP to LLMs, and it may become difficult to maintain the organizational capacity needed to improve and support conventional NLP pipelines.

Syntactic parsing with LLMs. With an eye toward ensuring the sustainability of NLP development in the technology market, we propose a practical approach in which conventional NLP pipelines are replaced with LLMs while improving accuracy. Based on the method of the prior work (Matsuda et al., 2025), we perform SFT with Low-Rank Adapters (LoRA) (Hu et al., 2022) on a multilingual LLM whose weights are publicly available by using a further improved version of the dependency parsing prompt from Matsuda et al. (2025).

2. Related Work

2.1. NLP Frameworks

Below we briefly survey NLP frameworks whose parsing functionality continues to be improved at the time of writing, in order of public release. `spaCy` performs deterministic or beam-search dependency parsing with the Non-Monotonic Arc-Eager Transition System (Honnibal and Johnson, 2015). Multitask learning is possible between the transformer and the parsing components. `Stanza`

Proceedings of LT4HALA 2026 @ LREC 2026, pages 448–452.
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

can learn the entire pipeline end to end, from language identification to dependency parsing (Qi et al., 2020). Its dependency parser uses an extended biaffine model with Bi-LSTMs. In recent years, constituency parsing has also been addressed (Bauer and Manning, 2025). `Spark NLP` is an NLP framework that claims support for more than 200 languages and a wide range of development environments (Python/Scala/Java/R) (Kocaman and Talby, 2021). Details such as the dependency parsing algorithm are not publicly available. `HanLP` performs multitask learning between a shared encoder and multiple parsing components. A biaffine-derived model is used for dependency parsing (He and Choi, 2021).

2.2. Dependency Parsing

Many of the NLP frameworks mentioned above perform dependency parsing with variants of the biaffine model (Dozat and Manning, 2017). Recently, methods such as Hexatagger (Amini et al., 2023), which represent local tree structure with a small number of tags and learn it with an encoder model, have been reported to achieve high parsing accuracy. For dependency parsing with autoregressive decoder models such as LLMs, methods that express dependency structure through word indices in a tabular format such as `CoNLL-U` are considered more accurate than methods that output trees in bracketing notation (Matsuda et al., 2025). In this paper, we extend the method of the prior work (Matsuda et al., 2025) and propose a method in which processing steps other than dependency parsing are also carried out by LLMs.

3. Proposed Method

We modified `omnes flores`[1], a unified NLP framework which employs LLMs for tasks such as language identification, sentence delimitation, word segmentation, part-of-speech tagging, and dependency parsing, in terms of both training data and input features for EvaLatin Dependency Parsing Tasks (Iurescia et al., 2026). For details of the methods of SFT and inference of the underlying `omnes flores`, refer to the official website.

3.1. Adding public UD Latin Treebanks

The default model of `omnes flores` uses the following 40 treebanks each of which has at least 40,000 words in its training split and a license that permits commercial use.

UD_Armenian-ArmTDP, UD_Belarusian-HSE, UD_Bororo-BDT, UD_Chinese-GSD, UD_Chinese-GSDSimp, UD_Croatian-SET, UD_Czech-CAC, UD_Danish-DDT, UD_Dutch-Alpino, UD_English-EWT, UD_Estonian-EWT, UD_Finnish-TDT, UD_French-GSD, UD_German-GSD, UD_Haitian_Creole-Adolphe, UD_Hebrew-IAHLTwiki, UD_Icelandic-GC, UD_Indonesian-GSD, UD_Irish-IDT, UD_Japanese-GSDLUW, UD_Korean-Kaist, UD_Latvian-LVTB, UD_Lithuanian-ALKSNIS, UD_Naija-NSC, UD_Norwegian-Nynorsk, UD_Persian-PerDT, UD_Portuguese-Porttinari, UD_Romanian-RRT, UD_Russian-GSD, UD_Scottish_Gaelic-ARCOSG, UD_Serbian-SET, UD_Sindhi-Isra, UD_Slovak-SNK, UD_Slovenian-SSJ, UD_Spanish-GSD, UD_Swedish-Talbanken, UD_Thai-TUD, UD_Turkish-BOUN, UD_Ukrainian-ParlaMint, UD_Western_Armenian-ArmTDP.

We added the following six Latin treebanks to the 40 treebanks listed above and finally performed LoRA SFT on the merged training set.

UD_Latin-ITTB, UD_Latin-LLCT, UD_Latin-UDante, UD_Latin-CIRCSE, UD_Latin-Perseus, UD_Latin-PROIEL.

3.2. Dialog Template

The dialog template used for dependency parsing is shown in Figure 1. We input `language`, `sentence`, and `indexed word list` together with a prompt instructing UD part-of-speech tagging, head prediction, and dependency relation labeling, and train the correspondence between the input and the response, which is a TSV-formatted dependency structure output.

Prior work (Matsuda et al., 2025) reported that the best accuracy is obtained when prompts are used to perform UD part-of-speech tagging, head prediction, and dependency relation labeling in three steps. However, increasing the number of steps lengthens the output context, which strongly affects inference efficiency. In addition, because of LLM context-length limits, truncation can occur at the end of the output for long inputs. In order to shorten the output context, `omnes flores` performs all dependency parsing subtasks in a single step and trains the model with the word-form column omitted from the output TSV.

In preliminary experiments, we confirmed that the context length of the proposed method (instruction plus response) is less than half that of Matsuda et al. (2025), while achieving accuracy comparable to Matsuda et al. (2025) on three metrics: UD part-of-speech tagging accuracy (UPOS), unlabeled attachment score (UAS), and labeled attachment score (LAS).

[1] https://megagonlabs.github.io/omnes-flores

{role: system} `You are a linguist and specialize in dependency parsing based on Universal Dependencies.`

{role: user} `We will now perform dependency parsing on` **`Latin`** `sentence.`
`After splitting the input sentence into words with indexes as shown below, create a TSV with four fields:`
`word index from 1 to` **`7`** `+ part of speech + the dependent word index + the Universal Dependencies relation.`

`input sentence:`
species intelligibilis similitudo est alicuius intellecti.

`words:`
1 species
2 intelligibilis
3 similitudo
4 est
5 alicuius
6 intellecti
7 .

{role: assistant}

```
1   NOUN    3   nsubj
2   ADJ     1   amod
3   NOUN    0   root
4   AUX     3   cop
5   PRON    3   nmod
6   VERB    5   acl
7   PUNCT   3   punct
<eos>
```

Figure 1: An example of the **1-step-no-form** dialog instance. The parts that change from instance to instance are shown in **`bold`**. The `shaded region` in the assistant-role corresponds to the range over which the loss gradient is computed during training, and to the decoded text during inference. At inference time, the span from the system-role up to the assistant-role header is provided as input, and decoding of the subsequent segment continues until `<eos>` is generated.

Train	Test - LAS						
Dataset	circse	ittb	llct	perseus	proiel	udante	Macro AVG
circse	51.3	25.4	23.5	26.7	53.6	21.2	33.6
ittb	24.3	90.5	56.8	71.0	30.5	70.1	57.2
llct	1.0	63.2	94.1	60.8	1.4	55.4	46.0
perseus	13.9	57.2	42.5	65.0	10.2	45.8	39.1
proiel	59.4	51.5	47.6	53.6	84.6	43.4	56.7
udante	5.3	69.3	52.5	65.9	9.5	63.3	44.3
latin6	61.6	85.8	90.6	74.2	78.7	71.0	77.0
40lang+latin6	61.9	86.5	90.8	75.0	80.5	72.8	77.9

Table 1: LAS scores under various settings when using FORM as input. `latin6` is obtained by merging the training sets of `circse`, `ittb`, `llct`, `perseus`, `proiel`, and `udante`. `40lang+latin6` is the model trained on data obtained by merging the 40-language training data of `omnes flores` and `latin6`.

4. Experiments

4.1. Settings and Accuracy Measures

Hardware. Google Cloud G4 instance (8 x RTX Pro 6000 GPUs, 384-core AMD Turin CPUs, 768 GB RAM), 250 GB Hyperdisk, 10 TB Filestore.
Software. Ubuntu 24.04, CUDA 12.8, Python 3.12.11, PyTorch 2.8.0, Transformers 4.57.0, Tokenizers 0.22.1, TRL 0.19.1, PEFT 0.17.1.
Base model. We use gemma-2-9b, which showed good accuracy in Matsuda et al. (2025).
Hyperparameters. We follow the settings of Matsuda et al. (2025), but reduce the number of epochs from 3 to 2 in order to shorten training time. The remaining settings are as follows: max_seq_length: 8,192, lr: 3e-4, lr_scheduler: cosine_with_min_lr, min_lr: 0.1, lora_r: 8, lora_dropout: 0.05, target_modules: all-linear.
Accuracy measures. We use UD part-of-speech tagging accuracy (UPOS), unlabeled attachment score (UAS), and labeled attachment score (LAS).

4.2. Evaluation on public UD Latin Treebanks

We performed LoRA SFT on `gemma-2-9b` with the dialog instances generated in the way explained in Section 3.2, and then carried out inference and evaluation with various settings. The results are shown in Tables 1 and 2.

Table 1 shows the following types of results when using FORM on input:
- Trained on each of the UD Latin treebanks
- Trained on a merged set of the six UD Latin treebanks (`latin6`)
- Trained on the default `omnes flores` model with the addition of `latin6` to the training data for 40 languages (`40lang+latin6`)

Table 2 shows the same types of results as Table 1 when using both FORM and UPOS.

The `40lang+latin6` model shows the best LAS across all public UD Latin treebanks. In addition, when UPOS is used together with FORM as input, an improvement in LAS of 2.1 points is achieved.

Regarding the addition of `40lang`, the following

Train	Test - LAS							Effect of Adding UPOS
Dataset	circse	ittb	llct	perseus	proiel	udante	Macro AVG	
circse	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-33.6
ittb	0.1	80.0	55.5	67.6	0.0	63.8	44.5	-12.7
llct	9.3	57.8	81.2	57.6	9.5	51.3	44.4	-1.5
perseus	0.1	3.7	0.2	3.7	0.0	2.1	1.6	-37.5
proiel	54.2	58.6	50.5	61.2	72.5	49.3	57.7	1.0
udante	6.2	7.5	5.2	3.4	9.5	8.0	6.6	-37.7
latin6	61.5	82.3	88.2	73.9	77.8	70.0	75.6	-1.4
40lang+latin6	65.3	87.2	91.3	77.9	81.8	76.6	80.0	2.1

Table 2: The LAS scores when using both FORM and UPOS on input.

System	Setting	Poetry		Prose	
		CLAS (P/R/F1/Rank)	LAS (P/R/F1/Rank)	CLAS (P/R/F1/Rank)	LAS (P/R/F1/Rank)
Omnes Flores_1	with subtypes	57.65 / 64.29 / 60.79 / 5	60.83 / 60.83 / 60.83 / 5	81.32 / 79.90 / 80.60 / 5	83.26 / 83.26 / 83.26 / 4
Omnes Flores_1	no subtypes	63.22 / 63.11 / 63.17 / 5	65.64 / 65.64 / 65.64 / 5	84.08 / 84.26 / 84.17 / 5	86.10 / 86.10 / 86.10 / 5
Omnes Flores_2	with subtypes	52.36 / 58.74 / 55.37 / 6	54.74 / 54.74 / 54.74 / 6	82.69 / 80.05 / 81.35 / 3	**83.74 / 83.74 / 83.74 / 2**
Omnes Flores_2	no subtypes	58.16 / 58.57 / 58.37 / 6	60.20 / 60.20 / 60.20 / 6	85.14 / 85.42 / 85.28 / 4	**87.09 / 87.09 / 87.09 / 4**

Table 3: Summary of EvaLatin 2026 Syntactic Parsing Results submitted by the Omnes Flores team. Model names are as follows: models that reference only FORM in the input are named Omnes Flores_1, and models that reference both FORM and UPOS are named Omnes Flores_2. Each cell reports Precision / Recall / F1 / Rank among all 8 models submitted to the EvaLatin 2026 Dependency Parsing Tasks. The evaluation results for all models submitted by all teams are publicly available in the official EvaLatin repository.

three reasons can be considered for the improved accuracy in Latin, even though almost all of the treebanks comprising 40lang are modern languages:

- Increased training volume
- Generalization from other Romance languages
- Generalization from languages with many Latin-derived words, such as English

Based on these results, we submitted the outputs of the 40lang+latin6 model to the EvaLatin 2026 Dependency Parsing Tasks (Iurescia et al., 2026). Model names are as follows: models that reference only FORM in the input are named Omnes Flores_1, and models that reference both FORM and UPOS are named Omnes Flores_2.

4.3. EvaLatin 2026 Results

The official evaluation results for Omnes Flores team submissions in EvaLatin 2026 Dependency Parsing Tasks are shown in Table 3. The Omnes Flores_1 refers to the outputs from the model trained with reference to FORM only, and Omnes Flores_2 refers to the outputs from the model trained with reference to both FORM and UPOS.

Across the two submissions, performance is clearly stronger on Prose than on Poetry. Removing subtypes improves every official F1 score, with gains ranging from 2.4 to 5.5 points. Omnes Flores_1 is consistently stronger on Poetry, whereas Omnes Flores_2 is consistently stronger on Prose.

The best Poetry result among the two systems is 65.6 LAS F1 from Omnes Flores_1 without subtypes, while the best Prose result is 87.1 LAS F1 from Omnes Flores_2 without subtypes. In the final ranking tables, Omnes Flores_2 is relatively competitive on Prose—it places 3rd in CLAS with subtypes and ties for 2nd in LAS with subtypes—while both models remain 5th or 6th in the Poetry rankings.

5. Conclusion

We propose a method that extends omnes flores by adding both training data and input features and show that it improves dependency parsing accuracy on Latin treebanks.

We submitted the outputs of the following two models to EvaLatin 2026 Dependency Parsing Tasks: (1) a model trained by adding six Latin treebanks to the training data of the default model of omnes flores, and (2) a model trained by adding UPOS to the input features of the model in (1).

In the final ranking table, one of our systems tied for 2nd in Prose LAS with subtypes, whereas both systems ranked 5th or 6th on Poetry.

Based on these results, we released an expanded version of the omnes flores model[2], trained on 84 languages comprising 99 treebanks, by adding 44 languages (including Latin), represented by 59 treebanks with independent training sets in the UD v2.17 treebank suite, in order to support languages not included among the 40 languages used to train the standard omnes flores model[3].

Future work will pursue two directions. First, we will investigate whether incorporating morphological features during both training and inference can further improve parsing accuracy. Second, we will analyze the factors underlying the performance degradation observed on the Poetry data when UPOS information is used in addition to FORM.

6. Acknowledgment

This work was conducted as part of a collaborative research project between Recruit Co., Ltd. and the National Institute for Japanese Language and Linguistics.

[2] https://huggingface.co/megagonlabs/omnes-flores-84-lang-99-treebank-non-commercial-v0

[3] https://huggingface.co/megagonlabs/omnes-flores-40-lang-41-treebank-v0

7. Bibliographical References

Afra Amini, Tianyu Liu, and Ryan Cotterell. 2023. Hexatagging: Projective dependency parsing as tagging. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1453–1464, Toronto, Canada.

John Bauer and Christopher D. Manning. 2025. High-accuracy transition-based constituency parsing. In *Proceedings of the 18th International Conference on Parsing Technologies (IWPT, SyntaxFest 2025)*, pages 26–39, Ljubljana, Slovenia.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *International Conference on Learning Representations*.

Han He and Jinho D. Choi. 2021. The stem cell hypothesis: Dilemma behind multi-task learning with transformer encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5555–5577, Punta Cana, Dominican Republic.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Federica Iurescia, Marco Passarotti, and Rachele Sprugnoli. 2026. Overview of the Dependency Parsing Task at EvaLatin 2026. In *Proceedings of the Fourth Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2026)*, Palma, Mallorca (Spain). ELRA.

Veysel Kocaman and David Talby. 2021. Spark NLP: Natural language understanding at scale. *Software Impacts*, page 100058.

Hiroshi Matsuda, Chunpeng Ma, and Masayuki Asahara. 2025. Step-by-step instructions and a simple tabular output format improve the dependency parsing accuracy of LLMs. In *Proceedings of the 18th International Conference on Parsing Technologies (IWPT, SyntaxFest 2025)*, pages 11–19, Ljubljana, Slovenia.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043. European Language Resources Association.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108.

Pre-Editorial Normalization for Automatically Transcribed Medieval Manuscripts in Old French and Latin

Thibault Clérice,[1] **Rachel Bawden,**[1] **Anthony Glaise,**[2]
Ariane Pinche,[3] **David A. Smith**[4]
[1]Inria, Paris, France [2]CNRS (UMR 7323), Tours, France,
[3]CNRS (UMR 5648), Lyon, France, [4]Northeastern University
{firstname.lastname}@{inria,cnrs}.fr, dasmith@ccs.neu.edu

Abstract

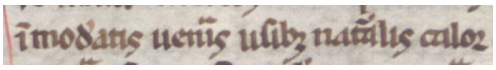
Recent advances in Automatic Text Recognition (ATR) have improved access to historical archives, yet a methodological divide persists between palaeographic transcriptions and normalized digital editions. While ATR models trained on more palaeographically-oriented datasets such as CATMuS have shown greater generalizability, their raw outputs remain poorly compatible with most readers and downstream NLP tools, thus creating a usability gap. On the other hand, ATR models trained to produce normalized outputs have been shown to struggle to adapt to new domains and tend to over-normalize and hallucinate. We introduce the task of Pre-Editorial Normalization (PEN), which consists in normalizing graphemic ATR output according to editorial conventions, which has the advantage of keeping an intermediate step with palaeographic fidelity while providing a normalized version for practical usability. We present a new dataset derived from the CoMMA corpus and aligned with digitized Old French and Latin editions using *passim*. We also produce a manually corrected gold-standard evaluation set. We benchmark this resource using ByT5-based sequence-to-sequence models on normalization and pre-annotation tasks. Our contributions include the formal definition of PEN, a 4.66M-sample silver training corpus, a 1.8k-sample gold evaluation set, and a normalization model achieving a 6.7% CER, substantially outperforming previous models for this task.

Keywords: Latin, Old French, Normalization, Text Recognition, Medieval manuscripts

1. Introduction

The progress of Automatic Text Recognition (ATR) has facilitated access to historical archives (Pinche and Stokes, 2024). However, there exists a methodological divide in how to transcribe manuscripts on a spectrum ranging between two extremes: the graphetic approach and a normalized approach (Figure 1). The graphetic approach, defined by Robinson and Solopova (1993), treats transcription as a graphical representation of the manuscript's text, which captures each character variant, and maintains original punctuation and abbreviation without expansion. Normalized transcription conflates transcribing with textual editing (Pierazzo, 2011), thus adhering more to users expectations by "[...] [making] the manuscript text accessible to the modern reader by expanding abbreviations, separating words, inserting diacritical marks, and adding modern punctuation" (Duval, 2012, p.330). However, this approach, combining two tasks at the same time, a graphical interpretation and a linguistic interpretation, has shown to poorly resolve abbreviations and to hallucinate content (Aguilar, 2024; Torres Aguilar and Jolivet, 2023), without providing an ability for the reader to distinguish normalization errors from graphical interpretation errors.

It can be useful to follow an approach between the two extremes, whereby character variants are merged but where abbreviations and other typical manuscripts phenomena are preserved: this is known as the graphemic approach. This enables the transcriptions to be treated as their own epistemic objects, which is an advantage in itself. It also results in transcriptions that result in models, such as those trained on the CATMuS dataset (Clérice et al., 2024), that better generalize to new domains and languages (Coll Ardanuy et al., 2025).



Graphetic	īmoꝺatis ueniꝭ uſibꝫ naᷓtᷓlıs calo2
Graphemic	īmodatis ueniīs usib; naᷓtᷓalis calor
Normalized	inmoderatis veneris usibus naturalis calor

Figure 1: The outputs of different transcription approaches for the same line of text. Manuscript: BSB CLM 13027. The task defined in this paper consists in producing normalized transcriptions from graphemic outputs.

However, graphemic transcriptions create a "usability gap" where the most accurate transcriptions of the material reality of the text are the least compatible with most available tools for computational analysis and expectations from the humanities (Kiessling, 2025).[1] It can therefore be use-

[1]Tools such as lemmatizers, part-of-speech (PoS) taggers, and Named Entity Recognition (NER) systems

Proceedings of LT4HALA 2026 @ LREC 2026, pages 453–468.
11 May 2026. *©ELRA Language Resources Association (ELRA), 2026*

ful to automatically produce normalized versions of these transcriptions. It holds an advantage over all-in-one normalized ATR approaches, as it provides an intermediate layer that makes it possible to distinguish ATR error (such as ‹rn›/‹m›) from normalization errors (‹ẽ&› for ‹est et› or ‹esset›).

In this paper, we introduce the task of automatically producing normalized transcriptions from graphemic ATR transcriptions, which we name pre-editorial normalization (PEN). The task definition and rationale is followed by the introduction of a novel dataset, based on the ATR silver corpus CoMMA (Clérice et al., 2025) and digitized and born-digital editions of Old French and Latin texts. Using *passim* (Smith et al., 2015), we automatically aligned both corpora, of which we then manually corrected some of the alignments to produce a gold-standard test set. We benchmark the silver dataset using models trained on a sequence-to-sequence transformer, ByT5 (Xue et al., 2022), across two tasks: (i) a normalization task, for which we consider our gold test set as the goal, and (ii) a pre-annotation task for lemmatization and PoS-tagging.

2. Related Work

2.1. Automatic Text Recognition with Normalized Output

ATR research has generally paid limited attention to the distinction between imitative transcription of historical sources and their subsequent normalization. Historical text recognition models are often evaluated on resources such as the IAMHistDB medieval datasets, the Perzival (Fischer et al., 2009), and the Saint Gall Database (Fischer et al., 2011), despite the fact that these corpora rely on different transcription paradigms. Camps et al. (2022) benchmarked both approaches on genre-constrained documents but without providing detailed qualitative analyses of the results.

By contrast, several studies within the Digital Humanities community have explicitly addressed the risks associated with over-normalization. In particular, two recent works have examined this issue as a central research question. Bottaioli et al. (2024) show that models trained to jointly transcribe and normalize middle names in digitized birth certificates may hallucinate names for individuals who did not possess one. Similarly, in the context of medieval manuscripts and, more broadly, in languages with highly variable orthography, Torterolo-Orta et al. (2025) demonstrate that increased accessibility through normalization entails a latent risk of systematically erasing historically and linguistically meaningful variation.

Conversely, qualitative error analyses conducted in several studies by S. T. Aguilar highlight persistent limitations of models trained on normalized outputs, particularly with respect to abbreviation resolution. Torres Aguilar and Jolivet (2023) show that morphosyntactic errors constitute one of the most frequent sources of abbreviation-related errors. Moreover, Aguilar (2024) report several critical failures, including difficulties in handling line-end hyphenation, the misinterpretation of simple abbreviations (e.g., ‹ĩfusũ› for ‹infusum› transcribed as ‹in futurum›) even in relatively straightforward contexts, and errors in local semantic disambiguation, such as interpreting ‹dia^{9}› as ‹diaconus› ‘deacon’ instead of ‹diabolus› ‘devil’.

2.2. Historical Text Normalization

Datasets for historical text normalization remain scarce, particularly for Latin. To our knowledge, the only comprehensive and readily available non-synthetic dataset for Latin historical normalization is the work of Schonhardt (2025a,b), which is limited to a single text attested in six manuscripts. This restricted scope limits its suitability for studying abbreviation expansion in a broader diachronic and bilingual context. A small number of additional datasets have been produced or identified in the work of Camps et al. (2022), some of which include Old French. However, their coverage remains limited in terms of genre and textual diversity.

Related work by Büttner (2025) addresses the normalization of Latin spelling by aligning medieval forms with more classical orthographic standards using a ByT5-based approach. While this study provides a relevant example of normalization in the Latin domain, it focuses on orthographic regularization rather than pre-editorial normalization.

By contrast, normalization for vernacular languages has received greater attention, beginning with late Middle French and extending into the early modern period. Solfrini et al. (2025) present a study on the normalization of printed late Middle French texts from the 1530s. More broadly, the automatic standardization of late Middle French has been explored in several works (Rubino et al., 2024b,a). These approaches are similar in methodology but differ in purpose from ours, as their primary aim is to reduce endemic linguistic variation rather than to preserve philological and historical features. On this topic, Bawden et al. (2022) compare normalization models for early modern French (17th c. and later), including datasets and comparative assessments of statistical and neural machine translation approaches.

are typically trained on normalized, modern-style editions. When faced with the raw, non-resolved output of a graphemic ATR model, the performance of these tools drops significantly (Clérice and Pinche, 2025).

Finally, digital editions may encode such information using a variety of representation frameworks, as illustrated in doctoral work (Franzini, 2019) and other projects (Nury, 2024). However, beyond the technical challenges involved in interpreting and exporting data from these platforms, identifying suitable editions remains a difficult task. To date, the only available catalogue providing filters for downloadable TEI resources (Andorfer and Franzini, 2025) indicates that, among more than 29 digital editions with accessible XML-TEI files for Latin and French texts, none preserves abbreviated forms with explicit markup.[2]

2.3. Normalization of User-Generated Content (UGC)

Another type of data associated with the lexical normalization task is user-generated content (UGC), of which social media texts are the most prevalent (Baldwin et al., 2013). UGC, like the historical data dealt with in this article, is characterised by non-standard phenomena such as acronyms, abbreviations, truncations and variations in spelling. Such texts can present challenges for natural language processing (NLP) tools, particularly those trained on standard, edited texts, which has led to a body of work in normalisation as a preprocessing step (Melero et al., 2012; Baldwin et al., 2015; van der Goot et al., 2021; Nishimwe et al., 2024a).

Initial approaches relied on rules, lexicons and language models, and it was common to decompose the task into the detection of words to be modified and their normalisation (Baranes, 2015; Ikeda et al., 2016). Later approaches opt for a translation-like approach, where the task is treated as the translation of a non-standard to standard language sequence (Aw et al., 2006; De Clercq, Orphée and Desmet, Bart and Schulz, Sarah and Lefever, Els and Hoste, Veronique, 2013). This approach has the advantage of increased use of context with the sequence and flexibility, helpful notably for segmentation and reordering differences.

The choice of representation has been a subject of much research in the handling of UGC texts (both for normalisation and other NLP tasks). Phonetic representations have been explored (Rosales Núñez et al., 2019; Mandal and Nanmaran, 2018), particularly for spelling errors. Character representations have also shown promise in terms of increasing robustness against phenomena such as spelling errors, devowelling, acronyms and abbreviations (Belinkov and Bisk, 2018; Watson et al., 2018; Domingo and Casacuberta, 2021; Scherrer and Ljubešić, 2021). Character-level representations have the advantage of better capturing similarities between similar but non-identical lexical forms. It is an approach that has previously been shown to work well for morphologically rich languages (which are also characterised by a higher level of variation) and low-resource scenarios (in order to generalize better across forms). For this reason, another successful approach is to either collect and annotate more data or to generate synthetic data (Belinkov and Bisk, 2018; Vaibhav et al., 2019; Nishimwe et al., 2024b).

Recent work has shown that models trained on large quantities of data, and in particular large language models (LLMs) are already "more robust" in that they perform better than previously on non-standard UGC data, without being specifically trained to do so and without having to apply a specific normalization pre-processing step (Bawden and Sagot, 2023; Peters and Martins, 2025; Bawden and Sagot, 2025).[3] However, this is unlikely to be the case across all languages, and in particular low-resource ones.

2.4. Automatic Alignment

Aligning multiple data representations has been an important method for dataset creation since the application of dynamic time warping in speech recognition (Sakoe and Chiba, 1978) or dynamic programming for aligning ATR outputs or multiple translations of the same source (Pial and Skiena, 2023; Thai et al., 2022). This paper uses methods developed for text-reuse detection, which is a broad task with many possible approaches and evaluations (Moritz et al., 2016). As described below (§4.2), the main inference task in creating our dataset is to match and align the ATR output of manuscript pages with passages in normalized editions. Unlike in other ATR evaluation tasks, these normalized editions do not have page breaks that match any particular manuscripts and the manuscripts are often miscellanies with several works. The size of the corpus makes it impractical to compare all pairs of ATR pages and passages from digital editions (Paju et al., 2022). We therefore employ passim 2.0(Smith et al., 2023), which uses hashed character n-gram fingerprinting to filter the set of candidate passages for alignment. Once a pair of passages is selected, passim uses a character-level hidden Markov alignment model (Vogel et al., 1996). Unlike text-reuse packages that use the Smith-Waterman alignment algorithm (Smith et al., 2014; Funk and Mullen, 2018), the HMM approach is more robust to the rearrangement of passages in the ATR output due to uncertainty in page layout analysis.

[2] Sahle et al. (2020) does not provide such metadata.

[3] Given that they are trained on large amounts of web data, it can be debated whether the models are actually more robust or whether the non-standard forms are simply sufficiently common in the training data.

3. Pre-editorial Normalization

3.1. Task Definition

The output of graphemic ATR[4] is inherently raw and reflects both manuscript-specific phenomena and transcription guidelines. As a result, it exhibits several sources of variability: spacing is inconsistent (including but not limited to unmarked line-break hyphenation), capitalization—especially in earlier centuries—may serve a decorative rather than linguistic function, numerical expressions vary across systems (ranging from Roman numerals to fractional notation in later periods), ramist letter pairs (‹u/v› and ‹i/j›) are not systematically distinguished, punctuation practices frequently differ from those adopted in modern editorial standards, and abbreviations are preserved. Finally, especially in the context of raw ATR output, recognition errors are superimposed on top of the multiple sources of variation described above.

These characteristics considerably limit the direct usability of ATR outputs in standard NLP and text-processing pipelines. Moreover, they diverge substantially from the normalized representations typically expected by scholars in the humanities, thereby motivating the need for dedicated normalization and post-processing methods.

The PEN task consists in:

- normalizing letters that were considered allographs in the middle ages and that diverged over time, such as ramist letters (‹u›/‹v›) in consonantal and vocalic usage. Historical spelling variants are not normalized; for instance, ‹f› replacing an etymological ‹ph› is preserved.
- resolving abbreviations in a context-sensitive manner: expansions should reflect not only the semantic content of words, but also their morphosyntactic features and a coherent orthographic system within the text.
- addressing recognition errors in the context of raw ATR output, such as ‹rn›/‹m› confusions.
- normalizing capitalization around punctuation and proper names according to modern editorial standards.
- preserving original punctuation.

[4] Robinson and Solopova (1993) defines graphemic as a transcription keeping original practices such as abbreviation but merging variants of same characters, such as ‹s› and ‹ſ›. They are implemented and more broadly defined in Clérice et al. (2024).

3.2. Limits of the Task

Certain abbreviations, especially those referring to personal names or place names, should not be expanded when the available knowledge in context is insufficient to support a reliable interpretation. For example, the abbreviation ‹G. dixit› ‘G. said’ in a medical text can be reliably expanded to ‹Galenus dixit› based on contextual evidence. On the other hand, in a charter or accounts, there might not be any contextual elements to resolve the name of S. in ‹S. Parisiensis› ‘S. the Parisian’, which could be many things (Stephanus, Severinus, etc.) and the initial should therefore be kept as it is.

Severe or unprocessable ATR errors resulting in extended spans of noise should not be corrected by the model. Attempting to normalize such segments risks introducing over-normalization and, more critically, hallucinated content. Given the text-based nature of the task, only recoverable textual information should be considered.

Numerical expressions written in Roman or Arabic numerals are likewise left unchanged. This form of variation remains an open issue within the community and requires further standardization before being fully integrated into the task definition. Moreover, the distinction between ordinal and cardinal numbers is often expressed through contextual or morphological markers and may require complex morphosyntactic analysis to be resolved reliably on top of the inherent variation of Roman numeral spelling.

Sufficient contextual information is required at all levels (semantic, morphosyntactic, and spelling) to support reliable normalization decisions. However, contextual evidence may itself be conflicting. For instance, if both ‹molt› and ‹moult› occur within the same Old French text, it remains unclear how the abbreviated form ‹młt› should be expanded.

In this paper, we frame the task as a single input–single output problem. In future work, the PEN task should also involve a processing layer capable of representing a distribution over possible outputs, to account for spelling variations within a single document, which can be seen in Old French (‹segnor› and ‹signor› on the same page).

4. Dataset Construction

4.1. Source Data

Our data for historical transcriptions is drawn from the CoMMA corpus, a graphemic-transcription corpus of 33,000 manuscripts (3.3B whitespace-delimited tokens). In our approach, we use the main text of each page to build full-text manuscript pages, ignoring running headers, marginal zones, and interlinear comments, thanks to the dataset’s

fine-grained details concerning layout classification. Library manuscript records carry language metadata; however, as this metadata operates at the document level, it is not used for filtering — our alignment process may identify passages in Latin or Old French within manuscripts bearing no such annotation.

For the normalized texts to align with CoMMA, we use publicly available data for Latin, and a mix of closed and open data for Old French, to ensure a large enough pool of data to align with the manuscripts. Both the Latin pool and the Old French one are drawn from the same data cited by CoMMA for their ModernBert model (Clérice et al., 2025, Table 6). Some of the corpus contains highly technical data (mathematics, medical, etc.), which we expect to be the hardest to normalize in Latin, as it contain higher levels of special vocabulary and specific abbreviations.

4.2. Alignment

We align the ATR transcripts of the CoMMA corpus with the plain text of digital editions of Latin and Old French texts using passim 2.0 (Smith et al., 2023). This version of the package implements a character-level HMM alignment algorithm (Vogel et al., 1996), which permits rearrangement of passages between two versions of a text. This is helpful in working with ATR output, when the sequence of page regions in the transcript does not match the sequence as transcribed in a scholarly edition. In contrast to other ATR ground-truth datasets produced by forced alignment, we do not need to specify in the input which manuscripts are versions of which texts; rather, we specify to passim that each page from CoMMA should be aligned to one or more pages from the corpus of digital editions. We performed unicode decomposition normalization on both the CoMMA ATR output and the digitial editions so that the common use of diacritics for, e.g., nasalization (‹cõsul› for ‹consul›) or truncation (‹ć› for ‹cum›) would align more closely to their standard forms. To improve recall, we index the collection with character 10-grams (ignoring whitespace and punctuation) that occur in fewer than 100 pages. It then aligns pages having at least five 10-grams in common and drops alignments shorter than 50 characters. We use beam search ($b = 200$) when aligning pages.

From the automatic alignment, we filter the data with at least five lines of continuous aligned content, a sufficiently large match rate (60%), and only where the alignment comes from the same source text. This resulted in 3.2M alignments. We then reprocessed them into smaller alignments, with a maximum and minimum size of 1000 and 300 bytes respectively, to get down to a total of 4,665,509 pairs.

Corpus	Work	# Mss.
PL	Auctor Incertus - Missale mixtum	2090
PL	Gratianus - Concordia discordantium canonum	2026
encyMe	Speculum doctrinale	2002
CSEL	Augustine - Liber qui appellatur Speculum	1795
PL	Martinus Legionensis - Sermones	1735
PL	Auctor incertus - Breviarium	1625
PL	Ivo Carnotensis - Decretum	1613

Table 1: Top works by manuscript presence.

4.3. Limits

As our approach relies on the automatic alignment of raw ATR output with digitized editions, several sources of noise may affect the resulting annotations regarding the task definition. First, since the distance between an abbreviated token and its expanded form influences alignment quality, there is a non-negligible probability that a manuscript form is aligned with a different lexical variant in the critical edition. This situation is common in textual transmission and may result in semantically or syntactically divergent alignments. Additionally, over-normalization may also occur, in a manner similar to variant-swapping phenomena, such as ‹f›/‹ph› alternations in Latin or ‹moult›/‹molt› variation in Old French. These cases may lead to normalized forms that contradict the intended task definition.

4.4. Description

Unsurprisingly, classical, late-antique, and Christian "literary" Latin texts are over-represented in the alignments between our source corpora and CoMMA, while later medieval administrative texts are under-represented. This reflects the fact that CoMMA is largely derived from library collections rather than archives. Patrologia Latina, our largest source (97M tokens, Roelli and Ctibor (2024)), accounts for 64.5% of the alignment pairs. By contrast, CEMA, an archival corpus and the second-largest source (57M tokens, Perreaux (2021)), represents only 1.5% of the alignments, whereas Perseus (6.1M tokens, Crane et al. (2021b)) accounts for 6.5%.

As expected from studies on manuscript transmission, some works are much more represented than others (see Table 1), and all most common authors are either tied to religious practice (missals, sermons and homilies), theory (commentaries), or canon law (*decretum*). One reason for this over-representation of some texts might be the presence of numerous bible citations.

Over-normalization, in the form of variant insertion, spelling standardization (in Latin), or lexical replacement (in Old French), constitutes one of the main identified risks. At this stage, however, we are unable to reliably distinguish over-

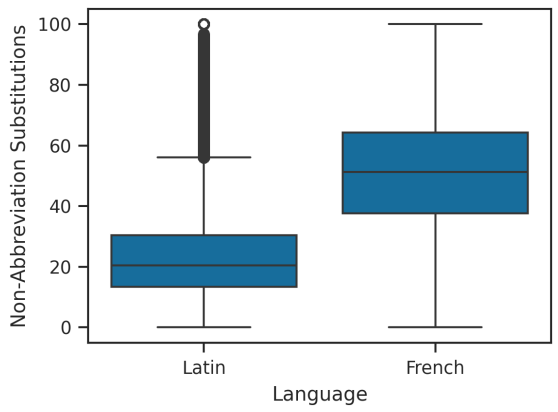


Figure 2: Distribution of the percentage of tokens per alignment in the train set that do not correspond to abbreviation resolution pattern.

normalization from two related phenomena: first, cases in which abbreviation markers are not recognized by the ATR system (e.g. ‹ifusa› for ‹ĩfusã› with missed diacritics), and second, post-ATR editorial corrections, which are likely more prevalent.

Nevertheless, we estimate the extent of over-normalization by automatically aligning tokens in the training set and measuring the proportion of tokens that differ between the source and aligned texts when no explicit abbreviation markers are present. This analysis reveals a marked contrast between the two languages (Figure 2). In Latin, 20.5% of tokens exhibit substitutions consistent with non-standard abbreviation resolution. This proportion broadly aligns with expectations based on the 9.7% character error rate (CER) reported for CoMMA, including additional effects from the limited possibilities of over-normalization in Latin and variant substitutions.

By contrast, Old French displays a substantially higher level of non-abbreviation related substitutions, with a median substitution rate of 51% (and a mean of 50%). This discrepancy suggests that, compared to Latin, Old French editions impose a greater degree of spelling normalization on manuscript sources, thereby introducing more extensive spelling variation between manuscripts and their corresponding editorial representations.

4.5. Manual Correction and Gold Standard

To create a gold set, we sampled some data to be representative of main corpora from the automatic alignment. The resulting pairs are all manually corrected by a philologist to resolve alignment errors, normalization inconsistencies, and segmentation issues. This results in a gold-standard dataset

		Word Length			
		Samples	**Average**	**Std**	**Median**
Raw ATR	French	189	95	34	89
	Latin	352	93	39	88
GT ATR	Latin	45	288	256	259
	Mixed	2	276	101	276

Table 2: Distribution of languages, sample counts, and token counts across the two alignment sources: raw ATR output and ground truth (GT).

that serves both for evaluation and for qualitative error analysis. To that test data, we add ground truth data from CATMuS data repositories (both archives and literary documents) that was manually normalised by the same expert.

In the automatically aligned ATR output-part of the gold dataset, two texts are substantially over-represented: *Oresme* (75 alignments) and the *Histoire Ancienne* (43). This imbalance reflects the limited diversity of available Old French datasets, which made it challenging to construct a fully balanced sample. All other corpora are represented by between 11 and 21 alignments. Among these, some technically demanding texts are included, such as the *Mathematica* corpus. It is represented by 20 alignments and contains atypical elements such as geometrical point labels (e.g. ‘A’), which resemble abbreviations but must not be treated as such.

The ground truth ATR focused specifically on two series of texts: the first are drawn from archives, small letters and charters, mostly from the Cartularies of Notre-Dame de Paris (Claustre and Smith, 2022); the second are drawn from a single manuscript bearing the text *Liber de Coitu* (Constantine the African, xı c.). The first series contains administrative text, with various anonymous people and occasional switching between Latin and French, while the second contains very abbreviated content, such as “eliq̊t̊́ hũor q̣̈ ẽ.ĩ cėbº ȝ eliq̊tº” (corresponding to *eliquatur humor qui est in cerebro, et eliquatus* ‘the humor that is in the brain liquefies, and once it is liquefied’).

The final gold dataset comprises 588 samples (64,230 tokens) from 316 different manuscripts. 47 alignments were drawn from ground truth ATR, with no or very few recognition errors, and which are therefore longer than other samples (287 tokens on average vs. 94, see Table 2).

5. Benchmarking Normalization

Our first experiment evaluates the impact of the proposed dataset in the context of ATR output normalization. Specifically, we train a model for abbreviation resolution and compare its performance

against off-the-shelf approaches, including zero-shot and domain-limited pretrained models.

5.1. Setup

We formulate text normalization as a sequence-to-sequence task. We select ByT5 as the base architecture, as it can handle a wide range of characters (including Unicode-decomposed forms), since it operates at the byte level. This property has been associated with improved performance on languages exhibiting high levels of spelling variation (Wolters and Van Cranenburgh, 2024). ByT5 has also been successfully applied to related tasks such as ATR correction (Pavlopoulos et al., 2024), medieval Latin normalization (Büttner, 2025), and abbreviation expansion (Schonhardt, 2025a).

Baselines We consider two baseline systems. The first is a ByT5-based model trained on the dataset of Schonhardt (2025a). This model relies on different Unicode-level guidelines for abbreviations and is strongly domain-constrained. Despite the shared scribal conventions for abbreviation between Old French and Latin, we expect this model to perform poorly on Old French, which is out of domain. In addition, the dataset's source text were manually corrected, which may hinder robustness to ATR noise and to inconsistencies in spacing.

Our second baseline consists of zero-shot prompting Llama 3 8B Instruct (Grattafiori et al., 2024). The model is prompted to perform PEN, with an auxiliary instruction to predict the input language, and common abbreviation patterns and glyphs (Appendix D).

Training data The training data is bilingual, combining Latin and Old French. Owing to a strong imbalance between the two languages and the relative scarcity of Old French data, we upsample the French portion of the training set by a factor of ten, reducing the imbalance from approximately 100:1 to 10:1 (Latin to Old French), and thereby ensuring sufficient exposure to Old French during training.

Experimental design We evaluate on our manually curated test set, as well as on the first 2,000 entries of the dataset introduced by Schonhardt (2025a), for which the Schonhardt baseline model should be at an advantage. All experiments are evaluated using CER and word error rate (WER), using the Levenshtein edit distance normalized over the length of the ground truth.

5.2. Results

Evaluated on our gold dataset, our proposed model performs well for both languages, improving over the zero-shot model and the only other available model for the task by significant margins (Table 3). On the Schonhardt dataset, our model shows a stronger performance than the original model itself, the original model showing a tendency to undergenerate.

Model	CER			WER		
	All	*Latin*	*Old French*	All	*Latin*	*Old French*
On our gold data						
Llama 3	139.9	160.4	79.9	163.7	183.1	107.3
ByT5 (Schonhardt)	64.8	75.6	33.3	92.3	104.8	55.9
ByT5 (Ours)	**6.7**	**6.3**	**7.6**	**20.7**	**19.6**	**23.8**
On Schonhardt data (Latin only)						
Llama 3		26.4			54.5	
ByT5 (Schonhardt)		130.0			109.4	
ByT5 (Ours)		**11.1**			**38.5**	

Table 3: Comparison of PEN results.

Qualitative analysis reveals a recurrent *classical bias* in the normalization of Latin, whereby the model tends to favor forms aligned with standardized spelling over historically attested medieval variants, except for final ‹-e› instead of ‹ae›. This tendency is illustrated by systematic preferences for ‹ph› over ‹f› and for ‹-tio-› over ‹-cio-›, reflecting a reversion to classical spellings. It also overpredicts names without context knowledge, such as ‹Johannis› for ‹Io.› and often struggles to adapt abbreviation resolution to the context, not unlike some of the models of Torres' work.

In Old French, the absence of stabilized orthographic conventions similarly leads the model to privilege dominant variants in its training set, normalizing forms such as ‹leurs› to ‹lors› and ‹pueple› to ‹pople›. This results in the systematic erasure of localized *scripta* traditions, as the model prioritizes alignment targets over historically meaningful scribal variation. It however shows a strong capacity to retokenize or correct wrong ATR predictions.

Finally, we observe cases of variant insertion by the model, or at least semantic failure such as ‹iustum› for ‹iterum›. While this phenomenon constitutes a limitation of the baseline model introduced in this study, it nevertheless highlights the importance of decoupling PEN from the ATR process itself. From a data-centric perspective, this separation enables us to trace normalization errors back to pre-editorial stages without requiring direct recourse to the original manuscript sources.

6. Impact on a Downstream Task

Given these results and the observed overnormalization effects, a remaining question is whether our normalization model—trained on the alignment dataset—can nevertheless outperform a previously established pipeline for downstream text analysis. To address this question, we compare our approach with the pipeline proposed by Clérice and Pinche (2025) in the context of stylometric au-

Pipeline Name	Lemma Pre.	Lemma Rec.	Lemma F1	POS Pre.	POS Rec.	POS F1	POS 3-grams Prec.	POS 3-grams Rec.	POS 3-grams F1
All tokens									
Lambert	88.2	**89.4**	88.8	95.8	**97.2**	96.5	85.6	**86.8**	86.2
Ours	**90.9**	89.0	**89.9**	**98.0**	95.9	**96.9**	**87.7**	85.8	**86.7**
100 most frequent words (MFW)									
Lambert	92.0	**92.6**	**92.3**						
Ours	**94.3**	89.5	91.8						

Table 4: Impact of the normalizer on downstream tasks (PoS tagging and lemmatization).

thorship analysis of *Vie de Saint Lambert*, focusing on the extraction of lemmas and PoS information, including PoS 3-grams.

6.1. Setup

Stylometric analysis of Old French texts is strongly hindered by spelling variation, particularly in inter-manuscript settings. Their pipeline addresses this issue by combining a convolutional neural network for space error correction with a model for both PoS-tagging and lemmatization (Clérice et al., 2019) fine-tuned on artificially abbreviated data (hereafter, the *Lambert* pipeline). On the other hand, our pipeline uses our normalization model and the same PoS-tagger/lemmatizer from (Clérice et al., 2019).

Data We conduct our evaluation on their transcription of *Vie de Saint Lambert* in the manuscript BnF Fr. 412. We compare their lemmatization and PoS-tagging outputs and the annotations produced by our own pipeline against the ground truth annotation of the normalized version. The *Vie* displays a very low amount of contextual abbreviation, and is mostly hindered by spacing issues and ATR errors. We compare the ground-truth based on the full results and on the 100 most frequent words (MFW), in a setting similar to stylometry practices (ignoring any punctuation).

Evaluation To assess the impact of these different preprocessing pipelines, we adopt a Bag-of-Words (BoW) evaluation framework and compute precision and recall in order to measure their suitability for frequency-based analyses. Given G and P the sequences of gold and predicted labels, and $c_G(x)$ and $c_P(x)$ the number of times a label x occurs in G and P, we define the number of true positives as $\text{TP} = \sum_x \min\big(c_G(x), c_P(x)\big)$. We then have $\text{Precision}_{\text{BoW}} = \frac{\text{TP}}{|P|}$, and $\text{Recall}_{\text{BoW}} = \frac{\text{TP}}{|G|}$.

6.2. Results

Overall, both models produce a comparable number of tokens. The Lambert pipeline generates approximately 2% more tokens, whereas the ByT5-based approach results in 3.8% fewer tokens than the edited text. For both systems, our error analysis reveals that annotation errors are not directly correlated with the correctness of the normalization step applied to individual tokens. For instance, the word “uie” (life), which is a non-ambiguous input for manual lemmatization, is poorly lemmatized 66% of the time by the Lambert pipeline and 30% by ours, with two out of three errors due to a non-normalization to “vie” and a lemmatization to UIE instead of VIE1.

The results (Table 4) exhibit a consistent general trend across all evaluated subsets. Precision and F1 are systematically higher for our approach, within a few percentage points, while recall is consistently higher for the Lambert pipeline by a comparable margin (except for MFW lemmatization). Our model introduces less noise than Lambert’s pipeline but undergenerates and produces inadequate content for the tagger: for example, it does not produce apostrophes (“qu il” instead of “qu’il”), whereas they seem to be an overused marker by the tagger.

In the context of stylometric analysis, both approaches therefore present complementary advantages, depending on their respective impact on the distribution of the 100 most frequent words. The advantage of our approach lies in a potential simplification of the pipeline, contrary to the multiplication of steps and ad-hoc models from Lambert’s.

7. Discussion and Conclusion

We introduced pre-editorial normalization (PEN) as an intermediate task between graphemic ATR output and fully edited (normalized) text, and presented a large-scale aligned dataset for medieval Latin and Old French. Our experiments show that models trained on this resource substantially outperform existing baselines on normalization, and that PEN shows promising results as a first step for downstream linguistic analysis while simplifying processing pipelines.

At the same time, our results highlight persistent risks of over-normalization and editorial bias, particularly in the tendency to favor more classical spellings for Latin and alternative ones for Old French. These findings underline the importance of treating normalization as an intermediate layer rather than an implicit part of transcription.

Future work will focus on expanding the dataset, modeling uncertainty in normalization, and more importantly page-wide and document-level normalization settings, in line with recent practices in machine translation to use broader contextual windows. We hope that this resource and task definition will support more transparent and reusable approaches to research on historical texts.

8. Bibliographical References

Sergio Torres Aguilar. 2024. Handwritten text recognition for historical documents using visual language models and gans.

Peter Andorfer and Greta Franzini. 2025. Catalogue digital editions. https://dig-ed-cat.acdh.oeaw.ac.at/.

AiTi Aw, Min Zhang, Juan Xiao, and Jian Su. 2006. A phrase-based statistical model for SMS text normalization. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 33–40, Sydney, Australia. Association for Computational Linguistics.

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How Noisy Social Media Text, How Diffrnt Social Media Sources? In *Proceedings of the 6th International Joint Conference on Natural Language Processing*, pages 356–364, Nagoya, Japan.

Timothy Baldwin, Marie Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. 2015. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 126–135, Beijing, China. Association for Computational Linguistics.

Marion Baranes. 2015. *Normalisation orthographique de corpus bruités*. Theses, Université Paris-Diderot - Paris VII.

Rachel Bawden, Jonathan Poinhos, Eleni Kogkitsidou, Philippe Gambette, Benoît Sagot, and Simon Gabay. 2022. Automatic normalisation of early Modern French. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3354–3366, Marseille, France. European Language Resources Association.

Rachel Bawden and Benoît Sagot. 2023. RoCS-MT: Robustness challenge set for machine translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 198–216, Singapore. Association for Computational Linguistics.

Rachel Bawden and Benoît Sagot. 2025. RoCS-MT v2 at WMT 2025: Robust challenge set for machine translation. In *Proceedings of the Tenth Conference on Machine Translation*, pages 834–849, Suzhou, China. Association for Computational Linguistics.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

Natalia Bottaioli, Solène Tarride, Jérémy Anger, Seginus Mowlavi, Marina Gardella, Antoine Tadros, Gabriele Facciolo, Rafael Grompone von Gioi, Christopher Kermorvant, Jean-Michel Morel, et al. 2024. Normalized vs diplomatic annotation: A case study of automatic information extraction from handwritten uruguayan birth certificates. In *International Conference on Document Analysis and Recognition*, pages 40–54. Springer.

Pascale Bourgain. 1992. Sur l'édition des textes littéraires latins médiévaux. *Bibliothèque de l'École des chartes*, 150(1):5–49.

Andreas Büttner. 2025. byt5-base-latin-normalize: A latin orthography normalization model. https://huggingface.co/andbue/byt5-base-latin-normalize.

Jean-Baptiste Camps, Chahan Vidal-Gorène, Dominique Stutzmann, Marguerite Vernet, and Ariane Pinche. 2022. Data Diversity in handwritten text recognition. Challenge or opportunity? In *Digital Humanities 2022. Responding to Asian diversity. Conference Abstracts (The University of Tokyo, Japan, 25-29 July 2022)*, Digital Humanities 2022. Responding to Asian diversity. Conference Abstracts (The University of Tokyo, Japan, 25-29 July 2022), pages 160–165, Tokyo, Japan. DH2022 Local Organizing Committee.

Julie Claustre and Darwin Smith. 2022. e-NdP notre-dame de paris et son cloitre (1326-1504). *Revue Mabillon*, 33:218–226.

Thibault Clérice, Simon Gabay, Malamatenia Vlachou-Efstathiou, Ariane Pinche, and Benoît Sagot. 2025. CoMMA, a Large-scale Corpus of Multilingual Medieval Archives.

Thibault Clérice and Ariane Pinche. 2025. Wauchier, Is That You? A multi-manuscript authorship analysis of Saint Lambert's life. In *CHR 2025 - 6th Conference on Computational Humanities Research*, Luxembourg, Luxembourg.

Thibault Clérice, Ariane Pinche, Malamatenia Vlachou-Efstathiou, Alix Chagué, Jean-Baptiste Camps, Matthias Gille Levenson, Olivier Brisville-Fertin, Federico Boschetti, Franz Fischer, Michael Gervers, Agnès Boutreux, Avery Manton, Simon Gabay, Patricia O'connor, Wouter Haverals, Mike Kestemont, Caroline

Vandyck, and Benjamin Kiessling. 2024. CATMuS Medieval: A multilingual large-scale cross-century dataset in Latin script for handwritten text recognition and beyond. In *Lecture Notes in Computer Science*, volume 14806 of *Lecture Notes in Computer Science*, pages 174–194, Athens, Greece. Springer Nature Switzerland.

Thibault Clérice, Jean-Baptiste Camps, and Ariane Pinche. 2019. Deucalion, modèle ancien francais (0.2.0).

Mariona Coll Ardanuy, Iban Berganzo-Besga, Ramon Sarobe, and Coral Cuadrada. 2025. Evaluating handwritten text recognition in medieval notarial manuscripts: A new dataset and comprehensive analysis. In *International Conference on Document Analysis and Recognition*, pages 340–357. Springer.

De Clercq, Orphée and Desmet, Bart and Schulz, Sarah and Lefever, Els and Hoste, Veronique. 2013. Normalization of Dutch user-generated content. In *Proceedings of Recent Advances in Natural Language Processing*, pages 179–188. INCOMA.

Miguel Domingo and Francisco Casacuberta. 2021. A comparison of character-based neural machine translations techniques applied to spelling normalization. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10-15, 2021, Proceedings, Part VII*, page 326–338, Berlin, Heidelberg. Springer-Verlag.

Frédéric Duval. 2012. Transcrire le français médiéval: De l'« Instruction » de Paul Meyer à la description linguistique contemporaine. *Bibliothèque de l'École des chartes*, 170(2):321–342.

Andreas Fischer, Volkmar Frinken, Alicia Fornés, and Horst Bunke. 2011. Transcription alignment of latin manuscripts using hidden markov models. In *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing*, pages 29–36.

Andreas Fischer, Markus Wuthrich, Marcus Liwicki, Volkmar Frinken, Horst Bunke, Gabriel Viehhauser, and Michael Stolz. 2009. Automatic transcription of handwritten medieval documents. In *2009 15th International Conference on Virtual Systems and Multimedia*, pages 137–142. IEEE.

Greta Hayley Franzini. 2019. *Towards a first digital edition of the oldest surviving manuscript of St Augustine's De civitate Dei*. Ph.D. thesis, UCL (University College London).

Kellen Funk and Lincoln A. Mullen. 2018. The Spine of American Law: Digital Text Analysis and U.S. Legal Practice. *The American Historical Review*, 123(1):132–164.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan,

Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang,

Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The Llama 3 Herd of Models.

Bo Han and Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 368–378, Portland, Oregon, USA. Association for Computational Linguistics.

Taishi Ikeda, Hiroyuki Shindo, and Yuji Matsumoto. 2016. Japanese text normalization with encoder-decoder model. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 129–137, Osaka, Japan. The COLING 2016 Organizing Committee.

Benjamin Kiessling. 2022. The Kraken OCR system.

Benjamin Kiessling. 2025. Transcription Guidelines for Generalized Automatic Text Recognition.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2023. Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 13094–13102.

Soumil Mandal and Karthick Nanmaran. 2018. Normalization of transliterated words in code-mixed data using Seq2Seq model & Levenshtein distance. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 49–53, Brussels, Belgium. Association for Computational Linguistics.

Héctor Martínez Alonso, Djamé Seddah, and Benoît Sagot. 2016. From noisy questions to Minecraft texts: Annotation challenges in extreme syntax scenario. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 13–23, Osaka, Japan. The COLING 2016 Organizing Committee.

Claudia Matos Veliz, Orphee De Clercq, and Veronique Hoste. 2019. Comparing MT approaches for text normalization. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 740–749, Varna, Bulgaria. INCOMA Ltd.

Maite Melero, Marta R. Costa-Jussà, Judith Domingo, Montse Marquina, and Martí Quixal. 2012. Holaaa!! writin like u talk is kewl but kinda hard 4 NLP. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3794–3800, Istanbul, Turkey. European Language Resources Association (ELRA).

Paul Michel and Graham Neubig. 2018. MTNT: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 543–553, Brussels, Belgium. Association for Computational Linguistics.

Maria Moritz, Andreas Wiederhold, Barbara Pavlek, Yuri Bizzoni, and Marco Büchler. 2016. Non-literal text reuse in historical texts: An approach to identify reuse transformations and its application to Bible reuse. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1849–1859, Austin, Texas. Association for Computational Linguistics.

Lydia Nishimwe, Benoît Sagot, and Rachel Bawden. 2024a. Étude sur la normalisation lexicale de contenus produits par les utilisateurs [a study on the lexical normalisation of user-generated content]. *Traitement Automatique des Langues*, 64(2):15–41.

Lydia Nishimwe, Benoît Sagot, and Rachel Bawden. 2024b. Making sentence embeddings robust to user-generated content. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10984–10998, Torino, Italia. ELRA and ICCL.

Elisa Nury. 2024. Codex bobbiensis (vl 1) dataset 8 (f.36r to f.39v).

Petri Paju, Heli Rantala, and Hannu Salmi. 2022. Towards an ontology and epistemology of text reuse. In Estelle Bunout, Maud Ehrmann, and Frédéric Clavert, editors, *Digitised Newspapers – A New Eldorado for Historians?: Reflections on Tools, Methods and Epistemology*. De Gruyter.

Jungsoo Park, Mujeen Sung, Jinhyuk Lee, and Jaewoo Kang. 2020. Adversarial subword regularization for robust neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1945–1953, Online. Association for Computational Linguistics.

John Pavlopoulos, Vasiliki Kougia, Esteban Garces Arias, Paraskevi Platanou, Stepan Shabalin, Konstantina Liagkou, Emmanouil Papadatos, Holger Essler, Jean-Baptiste Camps, and Franz Fischer. 2024. Challenging Error Correction in Recognised Byzantine Greek. In *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)*, pages 1–12, Hybrid in Bangkok, Thailand and online. Association for Computational Linguistics.

Ben Peters and Andre Martins. 2025. Did Translation Models Get More Robust Without Anyone Even Noticing? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2445–2458, Vienna, Austria. Association for Computational Linguistics.

Tanzir Pial and Steven Skiena. 2023. GNAT: A General Narrative Alignment Tool. ArXiv:2311.03627 [cs].

Elena Pierazzo. 2011. A rationale of digital documentary editions. *Literary and Linguistic Computing*, 26(4):463–477.

Ariane Pinche. 2021. *Edition nativement numérique du recueil hagiographique "Li Seint Confessor" de Wauchier de Denain d'après le manuscrit fr. 412 de la Bibliothèque nationale de France*. Ph.D. thesis.

Ariane Pinche and Peter Stokes. 2024. Historical documents and automatic text recognition: Introduction. *Journal of Data Mining & Digital Humanities*, Historical Documents and automatic text recognition.

Peter Robinson and Elizabeth Solopova. 1993. Guidelines for transcription of the manuscripts of the wife of bath's prologue. *The Canterbury Tales Project Occasional Papers*, 1:19–52.

José Carlos Rosales Núñez, Djamé Seddah, and Guillaume Wisniewski. 2019. Phonetic normalization for machine translation of user generated content. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 407–416, Hong Kong, China. Association for Computational Linguistics.

Raphael Rubino, Sandra Coram-Mekkey, Johanna Gerlach, Jonathan David Mutal, and Pierrette Bouillon. 2024a. Automatic normalisation of Middle French and its impact on productivity. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 176–189, Torino, Italia. ELRA and ICCL.

Raphael Rubino, Johanna Gerlach, Jonathan Mutal, and Pierrette Bouillon. 2024b. Normalizing without modernizing: Keeping historical wordforms of Middle French while reducing spelling variants. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3394–3402, Mexico City, Mexico. Association for Computational Linguistics.

Patrick Sahle, Georg Vogeler, Jana Klinger, Stephan Makowski, and Nadine Sutor. 2020. Digital scholarly editions. Online resource.

Hiroaki Sakoe and Seibi Chiba. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26:159–165.

Yves Scherrer and Nikola Ljubešić. 2021. Sesame street to mount sinai: BERT-constrained character-level Moses models for multilingual lexical normalization. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 465–472, Online. Association for Computational Linguistics.

Michael Schonhardt. 2025a. Bdd abbreviations augmented: A parallel corpus for abbreviation expansion in medieval latin. https://huggingface.co/datasets/mschonhardt/bdd-abbreviations-augmented.

Michael Schonhardt. 2025b. Die digitale edition als schnittstelle: Editorische workflows im ki-zeitalter. *Das Mittelalter. Perspektiven mediävistischer Forschung*, 30(1):54–69.

Djamé Seddah, Benoit Sagot, Marie Candito, Virginie Mouilleron, and Vanessa Combet. 2012. The French Social Media Bank: a treebank of noisy user generated content. In *Proceedings of COLING 2012*, pages 2441–2458, Mumbai, India. The COLING 2012 Organizing Committee.

David A. Smith, Ryan Cordell, Elizabeth Maddock Dillon, Nick Stramp, and John Wilkerson. 2014. Detecting and modeling local text reuse. In *ACM-IEEE Joint Conference on Digital Libraries (JCDL)*.

David A. Smith, Ryan Cordell, and Abby Mullen. 2015. Computational Methods for Uncovering Reprinted Texts in Antebellum Newspapers. *American Literary History*, 27(3):E1–E15.

David A. Smith, Jacob Murel, Jonathan Parkes Allen, and Matthew Thomas Miller. 2023. Automatic collation for diversifying corpora: Commonly copied texts as distant supervision for handwritten text recognition. In *Computational Humanities Research*.

Sonia Solfrini, Mylène Dejouy, Aurélia Marques Oliveira, and Pierre-Olivier Beaulnes. 2025. Normaliser le moyen français : du graphématique au semi-diplomatique. In *Actes des 18e Rencontres Jeunes Chercheurs en RI (RJCRI) et 27ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL)*, pages 239–252, Marseille, France. ATALA \\& ARIA.

Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. 2022. Exploring Document-Level Literary Machine Translation with Parallel Paragraphs from World Literature. ArXiv:2210.14250 [cs].

Sergio Torres Aguilar and Vincent Jolivet. 2023. Handwritten Text Recognition for Documentary Medieval Manuscripts. *Journal of Data Mining and Digital Humanities*, Historical Documents and automatic text recognition.

Yanco Amor Torterolo-Orta, Jaione Macicior-Mitxelena, Marina Miguez-Lamanuzzi, and Ana García-Serrano. 2025. Transcribing Spanish Texts from the Past: Experiments with Transkribus, Tesseract and Granite.

Vaibhav, Sumeet Singh, Craig Stewart, and Graham Neubig. 2019. Improving robustness of machine translation with synthetic noise. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1916–1920, Minneapolis, Minnesota. Association for Computational Linguistics.

Rob van der Goot, Alan Ramponi, Arkaitz Zubiaga, Barbara Plank, Benjamin Muller, Iñaki San Vicente Roncal, Nikola Ljubešić, Özlem Çetinoğlu, Rahmad Mahendra, Talha Çolakoğlu, Timothy Baldwin, Tommaso Caselli, and Wladimir Sidorenko. 2021. MultiLexNorm: A shared task on multilingual lexical normalization. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 493–509, Online. Association for Computational Linguistics.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *Proceedings of the 16th Conference on Computational Linguistics (COLING)*, volume 2, page 836.

Daniel Watson, Nasser Zalmout, and Nizar Habash. 2018. Utilizing character and word embeddings for text normalization with sequence-to-sequence models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 837–843, Brussels, Belgium. Association for Computational Linguistics.

Andre Wolters and Andreas Van Cranenburgh. 2024. Historical dutch spelling normalization with pretrained language models. *Computational Linguistics in the Netherlands Journal*, 13:147–171.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

9. Language Resource References

Assmann, B. and Sahle, Patrick. 1826–. *Monumenta Germaniae Historica*.

Bauduin, Pierre and Buard, Pierre-Yves and Goloubkoff, Anne and Mancel, Émeline and Fournier-Fujimoto, Tamiko and Buquet, Thierry and Brossard, Stéphane and Trohel, Benoît and Combalbert, Grégory and Bisson, Marie and Berthelot, Clémentine and Dubois, Adrien and Bloche, Michaël. 2023. *Scripta*.

Jean-Baptiste Camps and Alice Cochet and Lucence Ing and Elena Albarran. 2019. *Geste: un corpus de chansons de geste, 2016-…*. Zenodo.

Clérice, Thibault. 2021. *Corpora of rare texts, Lasciva Roma*.

Crane, Gregory Ralph and Babeu, Alison and Cerrato, Lisa and Munson, Matthew and Dee, Stella and Gessner, Annette and Robertson, Bruce and Franzini, Greta and Stoyanova, Simona and Selle, Tabea and Springmann, Uwe and Clérice, Thibault. 2021a. *Corpus Scriptorum Ecclesiasticorum Latinorum digitized by Open Greek And Latin*.

Crane, Gregory Ralph and Mylonas, Elli and Smith, Neel and Almas, Bridget and Babeu, Alison and Berra, Aurélien and Bonhomme, Marie-Laurence and Brooks, David and Cerrato, Lisa and Clérice, Thibault and Dee, Stella and Ehwald, R. and Frering, Léa and Gessner, Annette and Harrington, J. Matthew and Himes, Zach and Jovanovic, Neven and Konieczny, Michael and Mahoney, Anne and Merrill, William and Mimno, David and Mueller, Lucian and Munson, Matthew and Singhal, Rashmi and Smith, David and Taounza-Jeminet, Alyx and Weaver, Gabriel A. and Wulfman, C. 2021b. *Perseus Canonical Latin Literature*.

Anthonij Dees and Achim Stein and Pierre Kunstmann and Martin Gleßgen and Giles Souvay. 2006. *Nouveau Corpus d'Amsterdam*.

Draelants, Isabelle and Kuhry, Emmanuelle. 2020. *Sources des Encyclopédies Médiévale*.

Glessgen, Martin-Dietrich and Carles, Hélène and Duval, Frédéric and Videsott, Paul. 2016. *Documents linguistiques galloromans*.

Goetz, Georg and Gitner, Adam, et al. *CGLO*. 23-12.

Guillot, Céline and Heiden, Serge and Lavrentiev, Alexei. 2022. *Base de français médiéval*.

Hasse, Dag Nikolaus. 2023. *Ptolemaeus Arabus et Latinus*.

Neven Jovanović. 2024. *Croatiae Auctores Latini Textus*. Zenodo.

Hannah Morcos and Simon Gaunt and Maria Teresa Rachetta and Henry Ravenhall and Simone Ventura and Luca Barbieri and Paul Caton and Geoffroy Noël and Ginestra Ferraro and Natasha Romanova. 2020. *The Histoire ancienne jusqu'à César: A Digital Edition (The Values of French)*.

Perreaux, Nicolas. 2021. *Cartae Europae Medii Aevi (CEMA)*.

Corinne Pierreville and Jules Nuguet and Ariane Pinche and Alexei Lavrentiev. 2025. *Projet Fabliaux*.

Pinelli, Luca and Ctibor, Jan and Roelli, Philipp and Degl'Innocenti, Antonella and Gamberini, Roberto and Carlamaria Crespi, Serena and Fizzarotti, Luisa and Montepaone, Olivia and Vangone, Laura and Santarelli, Riccardo. 2023. *Mirabile Digital Archive*.

Roelli, Philipp and Ctibor, Jan. 2024. *Corpus Corporum*.

Russo, Emilio. 2023. *Biblioteca Italiana*.

Schibel, Wolfgang and Kredel, Heinz and Niehl, Rüdiger. 1999–2008. *CAMENA: Latin Texts of Early Modern Europe*. DFG-funded project; online access.

Clive R. Sneddon. 1982. *Old French corpus*. Literary and Linguistic Data Service.

Francesco Stella and Edoardo D'Angelo and Giorgio Di Maria and Marina Buzzoni and Fulvio Delle Donne and Roberto Rosselli Del Turco. 2025. *ALIM: Archivio della Latinità Italiana del Medioevo*. Accessed: 2025-09-27.

Stutzmann, Dominique and Torres Aguilar, Sergio and Chaffenet, Paul. 2021. *HOME-Alcar: Aligned and Annotated Cartularies*. Zenodo.

Tabacco, Raffaella and Lana, Maurizio and Balossino, Michele and Bessi, Giancarlo and Bognini, Filippo and Buffa, Martina and Caso, Daniela and Cattaneo, Gianmario and Ciotti, Fabio and Ciusani, Mauro and Colombo, Piero and Cuzzotti, Claudia and del Core, Vincenzo and Della Calce, Elisa and Denicola, Luciano and Digirolamo, Letizia and Ferrandi, Etienne and Ferroni, Manuela and Fontana, Davide and Guiglia, Valeria and Loberti , Martina and Lucciano, Melanie and Maconi, Ludovica and Malaspina, Ermanno and Manuela, Ferroni and Marini, Alessia and Maronet, Léa and Massano, Federico and Mazzucco, Clementina and Mellano, Anastasia and Miglietta, Chiara and Mollea, Simone and Mosca, Laura and Musso, Simona and Naso, Manuela and Paniagua, David and Poncina, Fabio and Ramires, Giuseppe and Rinaldi, Valentina and Rosso, Nadia and Rota, Simona and Rozzi, Stefano and Rugnone, Elisa and Senore, Corinna and Stok, Fabio and Strona, Beatrice and Verny, Romain and Vittoria Martino, Maria. 2021. *Digital Library of Late-Antique Latin Texts*.

Théry, Julien. 2022. *APOSCRIPTA database. Unified Corpus of Papal Letters*. Zenodo.

A. Dataset and models

The ByT5 model can be found on HuggingFace at https://huggingface.co/comma-project/normalization-byt5-small.

The dataset is available on HuggingFace at https://huggingface.co/datasets/comma-project/pen-alignment-pairs.

B. Corpus Used for the Alignment

Corpus (Latin then French)	#kTokens
Corpus Corporum (div.) (Roelli and Ctibor, 2024)	96,995
CEMA (Perreaux, 2021)	57,601
CAMENA (Schibel et al., 1999–2008)	7,916
(Assmann and Sahle, 1826–)	6,930
DigilibLT (Tabacco et al., 2021)	6,674
Perseus (Crane et al., 2021b)	6,148
CSEL (Crane et al., 2021a)	6,046
CroaLa (Jovanović, 2024)	5,233
Mirabile (Pinelli et al., 2023)	4,655
Bib. Italiana (Russo, 2023)	4,520
ALIM (Stella et al., 2025)	4,284
Add. Texts (Clérice, 2021)	3,844
Scripta (Bauduin et al., 2023)	2,226
Aposcripta (Théry, 2022)	1,736
ALCAR (Stutzmann et al., 2021)	973
Dig. Ptolemaeus (Hasse, 2023)	880
CGLO (Goetz and Gitner)	855
SourceEncyMe (Draelants and Kuhry, 2020)	777
NCA Dees et al. (2006)	3,185
BFM (Guillot et al., 2022)	6,293
Fabliaux (Pierreville et al., 2025)	120
Geste (Camps et al., 2019)	324
DocLing (Glessgen et al., 2016)	1,278
TVOF (Morcos et al., 2020)	361
OFC (Sneddon, 1982)	14
Total	229,870

Table B.5: Normalized corpus used by CoMMA for the ModernBert model.

C. Ranking of Corpora according to the Number of Alignments

Corpus	#alignments
Patrologia Latina	2,838,542
Lasciva Roma	413,580
CSEL	377,601
PerseusDL	306,297
DigilibLT	179,899
Auctores Scientiarum Varii	110,535
CEMA-latin	78,570

Table C.6: The most frequent aligned corpora in the dataset.

D. Llama3 8B Prompt

system: You are a specialist in Medieval Paleography and HTR correction.

Your goal is to transform ”raw” HTR transcriptions into normalized Medieval Old French or Latin.

LINGUISTIC & PALAEOGRAPHIC RULES:

- HTR ERROR CORRECTION: Fix character confusions typical of HTR, such as ‘m’ being misread for ‘in’ (e.g., ‘mexper’ -> ‘inexper’) or ‘f’ for ‘s’.
- U/V & I/J NORMALIZATION: Distinguish between vocalic and consonantal uses. Use ‘v’ and ‘j’ for consonants and ‘u’ and ‘i’ for vowels.
- EXPANSION: Fully expand all scribal abbreviations and suspensions into their full character forms.
- GRAPHICAL VARIATION: Fully respect graphical variation, such as filosofia should be kept instead of normalizing to philosophia.
- APOSTROPHE: Add apostrophe where necessary (e.g., ‘dun’ -> ‘d’un’).
- SPACES: Remove extra spaces (e.g., ‘co eur’ -> ‘coeur’) and add necessary ones (e.g., ‘letriste’ -> ‘le triste’).

ABBREVIATION KEY:

- �underscore or ꝑ = com / con (e.g., ꝑme -> comme)
- ꝰ = us (e.g., voꝰ -> vous)
- q̇ = qui
- ł = ol (e.g., młt -> molt)
- q̃ (tilde/macron) = nasalization (n/m) or missing letters (e.g., q̃ -> que)
- vertical tilde for suspension (missing letters) or R based phonemes (e.g., -re/er/-ra/ar-)

Return ONLY JSON.

user: Normalize this HTR output: text

assistant:

OldBERTur: Named Entity Recognition for Medieval Icelandic

Pontus Henningsson, Eva Pettersson, Erik Lenas
Swedish National Archives, Uppsala University, Swedish National Archives
pontus.henningsson@riksarkivet.se, eva.pettersson@lingfil.uu.se, erik.lenas@riksarkivet.se

Abstract

We present OldBERTur, a Named Entity Recognition (NER) model for Old Icelandic available in two variations, one for normalised texts, and one for diplomatic texts. Using a BERT-based model architecture, we fine-tune an existing BERT language model, and due to training data scarcity, we employ multiple training configurations, including pre-training domain adaptation, sentence-level data resampling, and modern Icelandic data augmentation; achieving a 93 F1 score for normalised texts, and 79 for diplomatic texts. We find that additional training configurations, such as resampling entity-annotated Old Icelandic texts, significantly improve performance in low-resource settings, while the effectiveness of added training configurations diminishes as the available training data increases. Our models can be used to automatically identify and classify person and location names in texts sourced from the rich Icelandic medieval literary tradition. Our models, along with their data and code, are made publicly available to allow for reuse and future research into medieval Scandinavian NLP and beyond.

Keywords: Named Entity Recognition, NER, Old Icelandic, Natural Language Processing, NLP, Digital Humanities

1. Introduction

Old Icelandic, the written language of Iceland during the Middle Ages,[1] contains some of the most important literary works written in a vernacular language in Scandinavia (Clunies Ross, 2000). The medieval Icelandic literary texts, including the Poetic and Prose Eddas and numerous sagas, written mainly between the 12th and 15th century, have been studied extensively for many decades; playing a key role in the understanding of the medieval Icelandic and wider Scandinavian world along with its languages and poetry, as well as providing a tantalising glimpse into pre-Christian Scandinavia and its mythology (Acker and Larrington, 2002; Clover and Lindow, 2019; Tómasson, 2006).

Considerable efforts have been made to digitise many of these Old Icelandic texts, with, for instance, the Medieval Nordic Text Archive (Menota, 2025) offering free access to medieval Nordic manuscripts. Despite increased digitisation, Old Icelandic still remains relatively unexplored by modern Natural Language Processing (NLP) techniques. While Old Icelandic and modern Icelandic are relatively similar due to the conservative nature of Icelandic and its focus on language preservation (Hilmarsson-Dunn and Kristinsson, 2010; Árnason and Leonard, 2011), there are still differences in older texts that may be hard for NLP tools trained on modern language to analyse. This is particularly true for Named Entity Recognition (NER), the task of automatically extracting named entities, such as persons and places, since naming conventions tend to shift over time. The lack of dedicated NER tools for Old Icelandic thus limits the possibilities for researchers to extract and analyse people, places, and their relationships, at scale from the literary Old Icelandic tradition. Furthermore, seeing as Icelandic, and indeed Old Icelandic, is a language with rich and complex morphology (Loftsson and Rögnvaldsson, 2007), off-the-shelf methods from related languages, such as NLP tools created for other Scandinavian languages, cannot necessarily be directly applied to the Old Icelandic material.

In this paper, we present and discuss OldBERTur, an openly available BERT-based (Devlin et al., 2019) NER model for Old Icelandic available in two variations, one for normalised transcriptions of Old Icelandic and one for diplomatic transcriptions. We provide the code, data, and models as freely available resources.[2]

2. Related Work

Cultural heritage institutions, such as archives, libraries, and museums, have increasingly made their historical material available online through digitisation. These institutions now provide instant digital access to a multitude of historical content, for

[1] Often also referred to as Old Norse, we use the term Old Icelandic to signify the main language used for Icelandic literary production in medieval Iceland. For a deeper contextual discussion of Old Norse-Icelandic, the reader is directed to Clover and Lindow (2019); Clunies Ross (2000).

[2] Normalised NER Model: huggingface.co/Riksarkivet/oldbertur-normalised-old-icelandic-ner, diplomatic NER model: huggingface.co/Riksarkivet/oldbertur-diplomatic-old-icelandic-ner, code and data: github.com/phenningsson/Medieval-Icelandic-NER.

Proceedings of LT4HALA 2026 @ LREC 2026, pages 469–481.
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

instance scanned images of historical documents and digital editions featuring transcriptions (Birnbaum et al., 2017; Kasperowski et al., 2024). While this digitisation serves a way to make the material more accessible and findable, as well as a way to preserve and reduce the need for its physical use; digitisation has also created an abundance of machine-readable historical material (Ehrmann et al., 2023). This "Big Data of the Past" as coined by Kaplan and di Lenardo (2017), combined with modern computational methods hold great potential for enabling increased digitisation, accessibility, and analysis of the past at scale (Ehrmann et al., 2023; Nockels et al., 2024).

One area where historical texts and computational methods increasingly converge is through the development of NER models for historical texts. The HIPE shared tasks (Ehrmann et al., 2020, 2022) established benchmarks for NER models for historical newspapers in English, French, German, Swedish, and Finnish; with the results demonstrating that neural NER models, in particular those relying on domain-adapted BERT-models, show solid performance for historical NER. Notably, the HIPE tasks exhibit how fine-tuning and adapting a NER-model for both language and domain can improve its performance, especially when supplied with sufficient training data (Ehrmann et al., 2020, 2022).

2.1. NLP for Icelandic

Icelandic is a North Germanic language belonging to the West Scandinavian branch together with Norwegian and Faroese, while Danish and Swedish make up the East Scandinavian languages (Hovdhaugen et al., 2000). While relatively less-resourced compared to major European languages, Icelandic has some key NLP resources. The IceNLP toolkit (Loftsson, 2019) serves as a comprehensive rule-based toolbox for analysis of contemporary Icelandic texts, including a tokeniser, a part-of-speech tagger, NER, and more. IceBERT (Snæbjarnarson et al., 2022), a RoBERTa-based (Liu et al., 2019) model trained on 16GB of modern Icelandic text, achieves impressive performance on many downstream NLP tasks, including NER, with an F1 score of 91.43 on the MIM-GOLD-NER dataset (Ingólfsdóttir et al., 2020), and subsequently improved to 92.73 (Guðjónsson et al., 2021). However, computational tools for older versions of the Icelandic language remain scarce.

2.2. Medieval NER

Multiple studies have showcased that neural NER models can perform well on medieval texts. Torres Aguilar (2022) created an annotated corpus and NER models for multilingual medieval charters written in Spanish, French, and Latin from the 10th to 15th century, consisting of ∼2,3 million tokens and ∼177k annotated entities. Leveraging pre-trained models and neural networks through both Bi-LSTM-CRF and BERT-based approaches, the study achieves F1 scores over 94 for all developed models. Torres Aguilar and Stutzmann (2021) created an annotated corpus of around ∼500k tokens stemming from ∼1.2k charters written in medieval French from the 13th–14th century, using similar architectures as Torres Aguilar (2022) for the NER models. The results show competitive performance, with F1 scores averaging well above 90. Similarly, Díez Platas et al. (2021) developed a NER model for 12th–15th century Spanish on a corpus consisting of medieval literature and administrative texts containing more than 3k entities, achieving F1 scores in the 74–87 range, using a rule-based approach relying on linguistic features and gazetteers.

However, there are some challenges inherent to applying computational methods to medieval texts. Many medieval texts, especially texts written in vernacular languages, showcase frequent variations in orthography, regular use of context-dependent nicknames or titles, dialectal nuances, and use of abbreviations which often require domain knowledge to resolve (Aguilar, 2025; Clérice et al., 2024; Schoen and Saretto, 2022). Despite difficulties and diachronic changes, previous studies show that it is possible to create competent NER models for medieval texts when given enough data and computational support, although previous work has mainly focused on medieval Romance languages.

For Old Scandinavian languages in particular, Besnier and Mattingly (2021) created a small entity-annotated dataset consisting of the Old Icelandic poem Völuspá; something which can be significantly expanded upon, leaving a research gap that this study sets out to address.

3. Data

Despite the mass digitisation of historical material, not all languages are served the same. Similar to contemporary low-resource languages, many historical languages are faced with a lack of texts and data available for computational methods and analysis (Tekgürler, 2025). Historical corpora annotated with named entities are generally concentrated to the 19th–20th century; large corpora from other time periods are scarce (Ehrmann et al., 2023). Indeed, while there are digital editions of Old Icelandic texts available online, annotated corpora for NER is limited. We therefore use data augmentation to complement our Old Icelandic training data, further discussed in the sections below. We use

the traditional Beginning-Inside-Outside (BIO) tagging, and follow the CoNLL 2002 format (Tjong Kim Sang, 2002) for the annotated entities.

3.1. Medieval Nordic Text Archive

The Medieval Nordic Text Archive (Menota, 2025) provides free access to digital editions of medieval Nordic texts, often with multiple levels of transcription: *facsimile*, which copies the manuscript as faithfully as possible; *diplomatic*, which resolves abbreviations and reduces allographic variation while retaining many features of the original text; *normalised*, which resolves abbreviations and standardises orthography, spelling, and punctuation to a modern editorial standard (Haugen, 2004). We use only diplomatic and normalised transcriptions as training data; facsimile transcriptions are too difficult for our limited training data.

Texts available through Menota are encoded in a customised TEI (TEI Consortium, 2025) XML format, called Menotic XML (Haugen, 2004). There are 57 available Old Icelandic manuscripts (∼860k words total). Out of these 57 texts, nine are entity-annotated on a diplomatic level, and out of these nine texts annotated on a diplomatic level, seven are also annotated on a normalised level. Thus, we have nine texts for our diplomatic Menota corpus, and seven texts for our normalised Menota corpus, with the annotation carried out by experts and available through Menota (Haugen, 2019). All the entity-annotated texts contain information on person and place names as entities, and we only take person and place name entities into account. The specific texts used are listed in the Appendix.

Entities were extracted using Python, and invalid Unicode characters in the diplomatic transcriptions were substituted to their closest representation in the Medieval Unicode Font Initiative (MUFI) (Haugen, 2015), or, when not possible, we used their normalised Menota equivalent. Words split up due to line breaks were merged to ensure complete words and entities. All in all, we ended up with two Menota datasets, one consisting of diplomatic entities, and one consisting of normalised entities.

To illustrate the transcription levels used in this study, the below example shows a parallel text using different transcription levels from *Alexander saga* from Menota, where person names are marked in *italic* and place names in underline.

Diplomatic: eınn a�Ꝑ hans ſcaττkonungom er neꝑnðꝛ *pнllıppuſ*. hann reð ꝑyr grıklanðı. ðroττnıng kona hanſ heτ *olımpıaſ*. Son aττo þau þann er *alexanðer* heτ. Sa maðꝛ ꝓar með hırð konungſenſ er *nepτanabuſ* heꝑer heıτıð.

Normalised: Einn af hans skattkonungum er nefndr *Philippus*. Hann réð fyr Grikklandi. Drótning kona hans hét *Olympias*. Son áttu þau þann, er *Alexander* hét. Sá maðr var með hirð konungsins, er *Neptanabus* hefir heitit.

The diplomatic transcription retains letter forms from the original manuscript, such as τ(t), ꝑ(f), ð(d), ꝛ(r), and long-s(ſ), while the normalised transcription substitutes these to their modern equivalents. Furthermore, the normalised transcription capitalises entities, which the diplomatic crucially does not, meaning that the diplomatic texts not only exhibit a greater variation of characters and orthography, but also showcase a lack of an often important textual clue for identifying a named entity for both humans and machines alike. For instance, compare the diplomatic grıklanðı with the normalised Grikklandi, or *nepτanabuſ* with *Neptanabus*.

3.2. Icelandic Parsed Historical Corpus

The Icelandic Parsed Historical Corpus (IcePaHC) (Wallenberg et al., 2024), presented and discussed in detail by Rögnvaldsson et al. (2012), is an open and freely available corpus of parsed historical texts written in Icelandic, ranging from the 12th century to modern day, in normalised form. While not explicitly entity-annotated, entities can be extracted from parsing annotations, where proper nouns are tagged with *NPR* (Noun Proper) together with its case suffix, and each token is represented with both surface form and lemma (e.g., ”NPR-A Þorlák-þorlákur”). Named entities were extracted through a script that takes all *NPR* tags and its context, for instance extracting adjacent proper noun sequences such as the person ”Hallur Gissurarson” from the sequence ”(NPR-N Hallur-hallur) (NPR-N Gissurarson-gissurarson)”.

This approach, while effective for a high-recall extraction of entities, does require manual verification, both to correct missed or mistaken entity annotations, but also to distinguish between names of persons and locations in ambiguous cases. To this end, six texts from IcePaHC, written in the 13th century and not present in the Menota corpora, were automatically extracted using the script, and then manually verified by the lead author of this paper. This process resulted in an extension of the normalised Old Icelandic NER data, comparable to normalised versions of the Menota texts.

3.3. MIM-GOLD-NER

To further augment our Old Icelandic data, we use the MIM-GOLD-NER dataset (Ingólfsdóttir et al., 2020), a contemporary Icelandic corpus of over 1 million tokens with over 48k annotated entities. This dataset has been used in previous studies to

benchmark Icelandic NER performance (Guðjónsson et al., 2021; Snæbjarnarson et al., 2022), and significantly extends our training data. To meet our needs, we filter the corpus to only contain our target entities, i.e., persons and locations, whereas other entity types are converted to non-entities.

3.4. Corpus Statistics

As seen in Table 1, our resulting Old Icelandic corpora (Menota normalised, IcePaHC, and Menota diplomatic) are relatively small with limited entity counts. The entity density (the percentage of tokens that represent entities) ranges between 3.3% and 5.5%, meaning that around 95% of the tokens are non-entities. There is also a significant class imbalance, where around 85% of the entities in the Old Icelandic corpora are person entities. It is also worth noting, that the diplomatic Menota corpus is almost twice the size of the normalised Menota corpus, and seeing as there is a general lack of entity-annotated transcriptions of Old Icelandic, this diplomatic corpus forms the basis of our diplomatic training data. For normalised data on the other hand, IcePaHC provides additional data comparable to Menota's normalised texts, and these two corpora together form the basis of our normalised training data. MIM-GOLD-NER adds a substantial amount of entities along with a more balanced entity distribution (63% persons, 37% locations), but is only used for training since our target domain is Old Icelandic. The Old Icelandic texts used in this study span from ∼1210 to 1700, with the vast majority of texts dating to the 13th and 14th century. All Old Icelandic texts used are written between ∼1210–1399 except for one, Íslendingabók AM 113 b fol., which is a later 17th century copy of a medieval manuscript which has since been lost. The full name and details about each text and the corpus it belongs to is available in the Appendix.

Corpus	Tokens	Per.	Loc.	Dens.
Normalised				
Menota	73,258	2,197	368	3.6%
IcePaHC	91,383	4,119	702	5.5%
Diplomatic				
Menota	138,404	5,803	913	5.1%
Modern (augmentation)				
MIM-GOLD	>1M	15,587	9,002	3.3%

Table 1: Corpus statistics. Per. = Person, Loc. = Location, Dens. = entity density, the percentage of tokens representing entities.

4. Methods

We use the aforementioned *IceBERT* by Snæbjarnarson et al. (2022), a RoBERTa-based model (Liu et al., 2019) trained on modern Icelandic, as our base language model for the NER task. This base model is already pre-trained on contemporary Icelandic texts and has achieved competitive performance for Icelandic NER, serving as a suitable base for our Old Icelandic NER models. However, we hypothesise that IceBERT may struggle with Old Icelandic diplomatic texts due to their orthographic variation, and that domain adaptive pre-training may improve performance since domain adaptation has shown improved performance in previous studies (Gururangan et al., 2020; Schweter and Baiter, 2019).

4.1. Domain Adaptation

We fine-tune IceBERT using Masked Language Modelling (MLM) on an unlabelled corpus of diplomatic Old Icelandic texts derived from 15 works in Menota consisting of ∼34k sentences and ∼385k words. This corpus consists of our nine entity-annotated diplomatic texts without their annotations, i.e. just the texts themselves, as well as six additional unannotated diplomatic texts from the same period from Menota. The 15 diplomatic texts used for MLM domain-adaptation are listed in the Appendix. Following Gururangan et al. (2020), we use task-adaptive pre-training (TAPT) to facilitate model familiarity with Old Icelandic diplomatic texts prior to being fine-tuned for NER. Around 12% of characters in the diplomatic texts are characters that appear in neither our modern nor normalised Old Icelandic NER corpus, featuring characters such as: ꝛ, ꝑ, ꝺ, ꝓ.

We train the model for 8 epochs (Liu et al., 2019), batch size of 32, 256 maximum sequence length, 3e-5 learning rate (Devlin et al., 2019), and 6% warm up (Gururangan et al., 2020). The MLM masking probability is set to the standard 15% (Devlin et al., 2019; Gururangan et al., 2020; Liu et al., 2019). Training was performed on a Quadro RTX 5000 GPU using FP16 mixed precision. This resulted in a fine-tuned IceBERT model created for diplomatic Old Icelandic texts. For diplomatic NER, we thus use two variations: IceBERT off-the-shelf; and our domain adapted version of IceBERT. For normalised texts, we use the original IceBERT as our base model seeing as the normalised transcriptions follow more modern editorial standards.

4.2. Sentence-Level Resampling

Following Wang and Wang (2022), we apply sentence-level resampling to address the data

scarcity and class imbalance in our Old Icelandic corpora. We leverage *sentence-level Count weighted by Rareness* (sCR), which calculates an expansion factor for each sentence based on the number of entity tokens and the rareness weight for each entity type, where under-represented entity types are given a higher weight. We perform sCR on the normalised Old Icelandic data (Menota + IcePaHC), but only on the diplomatic texts derived from Menota. We strictly use sCR on our training data, not on our development or test set. Table 2 shows the sCR on our training sets, along with our other training data configurations.

4.3. Training Data Configurations

We experiment with several training data configurations to investigate the effect of each component. For normalised NER, we combine Menota (M), IcePaHC (I), and MIM-GOLD-NER (MIM), with additional sCR resampling of Menota and IcePaHC (marked by superscript R, ^{R}). For diplomatic NER, we use the same combinations as above, but only Menota is resampled since it is the only diplomatic source available. We also train models using only MIM-GOLD-NER to evaluate potential transfer learning from modern to Old Icelandic. Our training data configurations are presented in Table 2.

Config.	Per.	Loc.	Total
Normalised			
M	1,486	180	1,666
M^{R}	7,421	1,041	8,462
M + I	4,283	542	4,825
$(M + I)^{R}$	22,330	2,929	25,259
M + I + MIM	19,870	9,544	29,414
$(M + I)^{R}$ + MIM	37,917	11,931	49,848
Diplomatic			
M	3,969	472	4,441
M^{R}	20,485	2,622	23,107
M + I	8,088	1,174	9,262
M^{R} + I	24,604	3,324	27,928
M + MIM	19,556	9,474	29,030
M^{R} + MIM	36,072	11,624	47,696
M + I + MIM	23,675	10,176	33,851
M^{R} + I + MIM	40,191	12,326	52,517
Modern			
MIM	15,587	9,002	24,589

Table 2: Training data configurations. M = Menota, I = IcePaHC, MIM = MIM-GOLD-NER. ^{R} = sCR resampling. Per. = Person, Loc. = Location.

4.4. NER Training

We view NER as a token classification task, adding a token classification head on top of IceBERT. We use the BIO tagging scheme with two entity types; person and location. We therefore have five labels: B-person, I-person, B-location, I-location, and O. Using the recommendations of RoBERTa (Liu et al., 2019) and BERT (Devlin et al., 2019) for fine-tuning along with what performed the best during hyperparameter experimentation based on F1 scores achieved on the development set, we train all NER models for 5 epochs with 2e-5 learning rate, batch size of 16, and maximum sequence length of 256 tokens. We leverage a 10% warm-up ratio and 0.01 weight decay. To address class imbalance, we use class weights for weighted cross-entropy loss (He and Garcia, 2009); 0.1 for non-entities (O) and 30.0 for entity classes. We employ early stopping with a patience of 3 and select the checkpoint with the highest development F1 score for evaluation.

4.5. Evaluation

We use a stratified split in our evaluation due to our entity-class imbalance; a naive random split could result in evaluation sets with too few location entities, which could hinder a generalisable and robust evaluation of the NER models' performance. Sentences containing location entities are thus split 50/50 between training and evaluation, and the evaluation set is split equally between development and test set. Sentences without location entities use a standard 70/15/15 split. This stratified approach allows for both the development and test set to contain a sufficient amount of location entities to enable reliable evaluation of model performance while still retaining location entities in the training data.

To reflect our target domain, development and test sets are derived exclusively from Old Icelandic data (normalised: Menota + IcePaHC; diplomatic: Menota). The test sets are strictly held out on the sentence level, and thus never seen during training. They are, however, derived from the same manuscripts as the training data. While this is due to the lack of available entity-annotated Old Icelandic texts, this also implies that the reported scores may benefit from scribal conventions within these manuscripts, which means that the models' generalisability to completely unseen manuscripts remains to be fully evaluated (discussed further in Limitations, section 8). We evaluate our models on the test sets using the standard metrics of precision, recall, and micro-F1 score using the seqeval library (Nakayama, 2018), which evaluates on entity-level and adheres to the evaluation of Tjong Kim Sang (2002), where both entity boundary span and entity type needs to be correct in order for the prediction to be viewed as correct. As an example, the entire multi-token entity "Björn Leifsson" needs to be entirely and correctly matched to give a correct result. There are two diplomatic NER model vari-

ations: one using the original IceBERT language model, the other using the domain-adapted IceBERT model fine-tuned through MLM for diplomatic Old Icelandic texts; normalised NER uses the original IceBERT as language model. Our evaluation sets are presented in Table 3.

Split	Person	Location	Total	Density
Normalised				
Dev	1,036	265	1,301	5.1%
Test	997	263	1,260	5.0%
Diplomatic				
Dev	912	210	1,122	5.5%
Test	922	231	1,153	5.4%

Table 3: Evaluation sets. Dev = Development set, Test = Test set.

5. Results

In the sections below, we present our results for normalised and diplomatic NER, followed by ablation studies examining the contributions of domain adaptation, resampling, and modern Icelandic data augmentation.

5.1. Normalised NER Results

Table 4 presents the results for normalised Old Icelandic NER. Our best model achieves an F1 score of 0.93 (93) by combining resampled Old Icelandic data from Menota and IcePaHC, and augmented with modern Icelandic NER data from MIM-GOLD-NER. All training data configurations maintain a high recall, while precision is the most varied. Each component added to the configuration contributes positively to model performance. With a baseline of an F1 score at 0.87 using only Menota, adding IcePaHC yields a minor improvement to 0.90, whereas sCR resampling results in a greater gain, achieving an F1 of 0.92. The best performing model uses the maximum amount of available data of Old Icelandic resampling and modern Icelandic data augmentation with an F1 score of 0.93. Unsurprisingly, using only modern Icelandic data has the lowest performance in terms of F1 score.

5.2. Diplomatic NER Results

Table 5 displays the results for diplomatic Old Icelandic NER, comparing the results of the original IceBERT and our domain-adapted IceBERT. The difficulty of the diplomatic NER task is reflected in the models' performance, with the best model, the domain-adapted model using the maximum available data, achieving an F1 score of 0.79 (14 points lower than the normalised model). This showcases the genuine difficulty of diplomatic texts and its frequent orthographic variations which are significantly more diverse compared to the normalised Old Icelandic texts. While the recall is relatively consistent across data configurations, between 0.77 and 0.85 (excluding the MIM-only configuration), the precision changes drastically. The lowest-resource configuration, using the Menota-only training configuration with the original IceBERT model, achieves an F1 score of 0.17, indicating that the model over-predicts entities mistakenly when dealing with unfamiliar medieval orthographic forms. Precision improves progressively with increased data, and the domain-adapted IceBERT model outperforms the original across all configurations.

Training Data	P	R	F1	Loc.	Per.
M	0.82	0.94	0.87	0.84	0.88
M^R	0.85	**0.95**	0.90	0.87	0.90
M + I	0.86	**0.95**	0.90	0.87	0.90
$(M + I)^R$	0.89	**0.95**	0.92	**0.90**	**0.93**
M + I + MIM	0.89	0.94	0.91	0.89	0.92
$(M + I)^R$ + MIM	**0.90**	**0.95**	**0.93**	0.89	**0.93**
MIM	0.84	0.84	0.84	0.86	0.83

Table 4: NER results on normalised Old Icelandic. P = Precision, R = Recall, M = Menota, I = IcePaHC, MIM = MIM-GOLD-NER. Superscript R indicates sCR resampling. Loc. and Per. show F1 scores for Location and Person entities respectively.

Training Data	P	R	F1	Loc.	Per.
Original IceBERT					
M	0.17	0.80	0.28	0.24	0.29
M^R	0.45	0.84	0.58	0.54	0.59
M + I	0.33	0.80	0.47	0.36	0.50
M^R + I	0.55	0.83	0.66	0.63	0.67
M + MIM	0.57	0.77	0.65	0.54	0.68
M^R + MIM	0.73	0.81	0.77	0.70	0.79
M^R + I + MIM	0.72	0.82	0.77	0.69	0.78
MIM	0.38	0.16	0.22	0.17	0.23
Domain-adapted IceBERT					
M	0.28	0.83	0.42	0.39	0.42
M^R	0.54	**0.85**	0.66	0.65	0.66
M + I	0.42	0.84	0.56	0.51	0.57
M^R + I	0.61	0.84	0.71	0.68	0.72
M + MIM	0.59	0.82	0.69	0.59	0.71
M^R + MIM	0.75	0.81	0.78	**0.73**	0.79
M^R + I + MIM	**0.77**	0.81	**0.79**	**0.73**	**0.80**
MIM	0.37	0.29	0.32	0.21	0.34

Table 5: NER results on diplomatic Old Icelandic. P = Precision, R = Recall, M = Menota, I = IcePaHC, MIM = MIM-GOLD-NER. Superscript R indicates sCR resampling. Loc. and Per. show F1 scores for Location and Person entities respectively.

5.3. Ablation Study

In this section, we present how our training configurations impacted the performance of NER on normalised and diplomatic Old Icelandic texts.

5.3.1. Effects of Resampling

Figure 1 illustrates how the sCR resampling consistently improves performance across all the training data configurations for both normalised and diplomatic texts. While resampling provides marginal gains to the normalised configurations (+2–3 F1), it does provide significant improvements in the low-resource diplomatic configurations where data is scarce, notably with the Menota-only diplomatic configurations gaining a +24–30 F1 score through resampling. sCR helps address class imbalance and performance, and this is especially prominent when training data is limited.

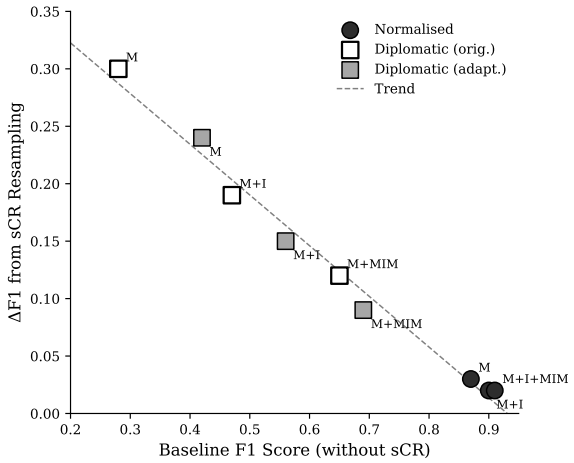


Figure 1: Relationship between baseline F1 and F1 improvement (Δ) from sCR resampling. Diplomatic texts benefit the most from resampling.

5.3.2. Effects of Domain Adaptation

Figure 2 illustrates how the domain-adapted IceBERT model outperforms the original IceBERT model on every training data configuration in terms of F1 score, where the difference in performance generally decreases as the amount of training data provided increases. The domain adaptation is most valuable in low-resource scenarios, where the domain-adapted language model's familiarity with the conventions of Old Icelandic diplomatic texts compensates for its limited NER training data, while the original IceBERT model progressively catches up as the amount of training data increases; the NER training itself successfully acts as implicit domain adaptation when given sufficient data.

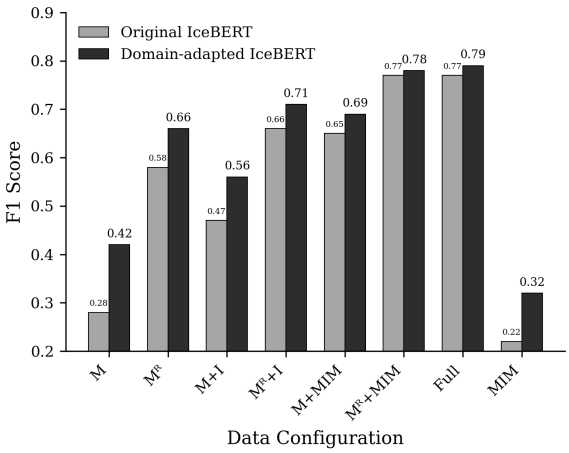


Figure 2: Effect of domain adaptation on diplomatic NER. "Full" relates to the data configuration of M^R + I + MIM which is the largest configuration.

5.3.3. Effects of Modern Icelandic Data Augmentation

Figure 3 illustrates the effect of adding modern Icelandic training data through the MIM-GOLD-NER dataset, where the lowest-resource diplomatic setting gains the biggest jump in performance, increasing F1 between +37 and +27 respectively, while the normalised settings sees an increase of a mere +0.01 F1. Similar to the results displayed in both Figure 1 and Figure 2 above, the impact of added data configurations, in this case the addition of modern Icelandic, diminishes as other, more domain-specific Old Icelandic training data increase. Interestingly, the big jump in performance for the lowest-resource diplomatic configurations suggests that there is some transfer learning from modern Icelandic to diplomatic Old Icelandic texts.

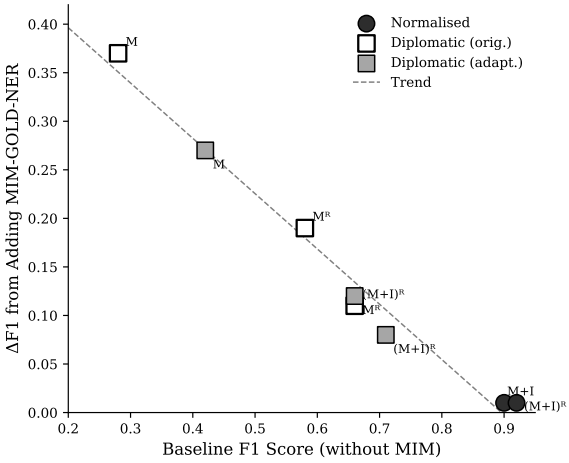


Figure 3: Relationship between baseline F1 and Δ improvement from adding modern Icelandic training data; diplomatic, low-resource, configurations benefits the most.

6. Discussion

Our experiments show competent NER performance for Old Icelandic despite its low-resource nature of NER-annotated data, with F1 scores of 93 for normalised and 79 for diplomatic texts. Our results are in line with previous medieval NER work, with F1 scores in the range of 74–94+ (Torres Aguilar, 2022; Torres Aguilar and Stutzmann, 2021; Díez Platas et al., 2021), and modern Icelandic F1 scores in the range 91–93 (Snæbjarnarson et al., 2022; Guðjónsson et al., 2021).

Looking at the difference between the F1 scores for person and location entities, person consistently has a higher F1 score than locations. Due to the class imbalance, this is expected, where around 85% of entities are persons in the Old Icelandic training data; the model simply sees far more person examples than locations during training. The resampling strategy of sCR is more effective for location entities compared to person entities, highlighting that the sCR strategy is working as intended since it privileges the resampling of under-represented entities (e.g., the domain-adapted diplomatic M vs. M^R gains +26 F1 points for location compared to +24 for person). Similarly, modern data augmentation aids the F1 scores of location in particular, which is likely due to an increase in location entities in the training data, as well as the MIM-GOLD-NER dataset having a more balanced entity distribution with around 37% entities being locations, balancing the person-focused Old Icelandic training data (e.g., domain-adapted M^R location F1 increases from 0.65 to 0.73 when adding MIM-GOLD-NER). However, the location F1 for normalised texts actually peaks at 0.90 with $(M + I)^R$ and slightly decreased to 0.89 when MIM-GOLD-NER is added. This suggests that the modern location names, often differing a lot to medieval place names, may dilute the model's performance ever so slightly when training data is abundant.

Across the ablation studies, the impact of training configurations are inverse to the baseline performance; low-resource configurations benefit greatly from each added configuration, while high-resource configurations show diminishing returns as the NER fine-tuning itself provides implicit domain adaptation. Conversely, when training data is scarce, each technique listed above compensates for the lack of data in complementary ways; resampling data helps address class imbalance, domain adaptation familiarises the model with Old Icelandic diplomatic orthography, and data augmentation of modern Icelandic provides additional entity examples to learn from. This is particularly evident in the precision-recall scores, where recall remains stable but precision varies considerably, where low-resource diplomatic training data configurations show an over-prediction on unfamiliar medieval orthographic forms.

6.1. Challenge of Diplomatic Texts

The 14-point performance gap between the best performing normalised and diplomatic model represents not only the fundamental challenge of working with medieval texts using different transcription levels, but also the linguistic distance between diplomatic Old Icelandic and modern Icelandic. Approximately 12% of all characters in the diplomatic corpus do not occur in the normalised Old Icelandic or modern Icelandic corpora, and the zero-shot results of models trained on only the modern Icelandic data are clearly indicative of this difficulty; MIM-only training achieves an F1 of 0.84 for normalised texts, but only 0.22-0.32 on diplomatic texts, where the domain-adapted model outperforms the original. This stark difference demonstrates how normalisation bridges the gap between medieval and modern texts through standardisation, while diplomatic texts preserve greater complexity and nuance, with the two different transcription levels likely appealing to different users. Our domain adaptation approach for diplomatic texts showcases consistent improvements, and the best F1 performance, across all configurations. However, while domain adaptation is valuable, it is not a complete substitute for domain-specific annotated NER training data, especially as the fine-tuning NER process itself acts as a way of adapting the model to its target domain in our experiments.

6.2. Error Analysis

To better understand our models' behaviour, we conducted a qualitative analysis on the predictions of our best-performing normalised and diplomatic NER models on their respective test set. The diplomatic model, unsurprisingly, showcases more errors than the normalised, which relates to their F1 score gap of 14 points. Both models primarily struggle with detection errors of false positives and negatives rather than entity span or classification errors; false positives and negatives constitute 67% of the normalised model's errors and 90% of the diplomatic errors. For both models, false positives are the most frequent, accounting for over half of all errors. The normalised model often predicts demonyms, such as "Gautar" (Geats), as entities (which is not necessarily wrong, but it is not what it is trained for), while the diplomatic model frequently mistakenly predicts common words as entities, like "ðalí". False negatives are relatively rare for the normalised model with 11%, while the diplomatic model's second largest error category is false negatives with 38%. This likely relates to the increased orthographic variation in the diplomatic texts, as

well as the lack of capitalised entities which might make it more difficult for the model to latch onto textual clues of an entity. In terms of classification errors, the normalised model errors account for 12% of correct entity span but wrong type, with only 4% for the diplomatic model. Span boundary errors constitute 21% of normalised errors and mainly include missed patronyms, with 6% for diplomatic. Full per entity precision and recall scores are presented in the Appendix.

6.3. Implications

Our results enable researchers working with normalised Old Icelandic texts to apply our NER model to extract person and place names, for instance for social network analysis of Icelandic saga literature, or mapping the spatial settings of medieval Icelandic texts. The diplomatic NER model, while requiring more manual verification, can allow for NER on texts that are only available in diplomatic digital editions, reducing the potential preprocessing burden of scholars needing NER for Old Icelandic while simultaneously preserving the texts' diplomatic textual form. The successful cross-temporal transfer of modern Icelandic NER data to normalised Old Icelandic texts demonstrates that modern language resources can be used to aid historical ones, suggesting that cross-temporal transfer for historical NLP development can be useful to augment low-resource historical training data. Finally, our work contributes to a growing ecosystem of NER tools crafted for medieval texts (Torres Aguilar and Stutzmann, 2021; Torres Aguilar, 2022; Besnier and Mattingly, 2021; Díez Platas et al., 2021), where we build on this to include capabilities for Old Icelandic.

6.4. Future Work

Future work could involve expanding the corpus of NER-annotated Old Icelandic texts, where a systematic annotation of Menota manuscripts, potentially involving our NER models and manual verification to speed up the annotation process, would provide a significant amount of additional training and evaluation data. Leveraging the many medieval Scandinavian manuscripts stored in Menota could also be used to explore the multilingual use of medieval NER models for Scandinavian languages. Old Norwegian, featuring prominently in Menota, shares significant linguistic similarities with Old Icelandic, and combining the two could make our NER models more robust. Finally, integrating the NER models and their output with entity linking could transform them from entities into linked data, where person names could be linked to name registries and place names to geographical gazetteers. Authority lists could be created in order to link an entity and its many spelling variations to one singular form through the use of clustering. The entities from this work could, for instance, be integrated as part of the Norse World project, a database of the Scandinavian world and spatiality sourced from medieval literature (Petrulevich et al., 2020). Similarly, investigating transfer learning from other Germanic languages, or indeed medieval Latin, could potentially provide a beneficial way of improving our models and their robustness, especially considering that medieval Latin was often used alongside vernacular languages in the Middle Ages. Finally, future work could include facsimile transcriptions since they are the most orthographically diverse and thereby more challenging, and can help shed further light on how orthographic variation affects NER performance.

7. Conclusion

We have presented OldBERTur, a NER system for Old Icelandic, achieving F1 scores of 93 for normalised and 79 for diplomatic transcriptions. Our experiments showcase competitive performance despite limited training data through domain adaptation, sentence-level resampling, and modern Icelandic data augmentation. Across ablation studies, we find that the effectiveness of additional data configurations is the inverse of the baseline performance; low-resource configurations show substantial benefit to added training configurations, higher-resource configurations show diminishing returns as the NER fine-tuning itself acts as an implicit domain adaptation. The 14-point F1 gap between our best performing normalised and diplomatic model reflects the difficulty of working with medieval orthographic conventions, especially for diplomatic texts where many characters do not appear in normalised Old Icelandic or modern Icelandic. We release our models, code, and data as open resources to contribute and support further research into medieval Scandinavian NLP and beyond.

8. Limitations

Our range of entities is restricted to only person and location names, and we perform flat rather than nested NER. For instance, "Björn Leifsson úr Ási" is captured as two entities rather than one; "Björn Leifsson" (person) and "Ási" (place). Furthermore, we evaluate our models on held-out test sets derived from the same manuscripts that we used for training. While this is due to the low-resource nature of Old Icelandic NER data, and we use stratified splitting to ensure a robust evaluation, we do not evaluate on completely unseen manuscripts or textual genres. While this reflects our target domain, it does imply that the generalisability of our models to other Old Icelandic texts, especially those with different scribal standards or from different time periods, remains to be evaluated upon. The Old Icelandic diplomatic training data is sourced exclusively from Menota, while the normalised data combines both Menota and IcePaHC. The asymmetry in training data means that the diplomatic model might be less diverse and generalisable.

9. Ethical Considerations

We work with publicly available historical texts with no modern personal data involved. As with all annotation, it is interpretative, and there may be potential biases in the NER annotations stemming from the annotators individual interpretation of how or what constitutes an entity and/or how it should be annotated according to an annotation guideline. However, this is reflective of the nature of annotation, and we consider this to not be an ethical risk.

10. Acknowledgements

This work was supported by Riksbankens Jubileumsfond (grant IN21-0003). We are also deeply grateful to Alexandra Petrulevich for providing the key conceptual starting point.

11. Bibliographical References

Paul Acker and Carolyne Larrington. 2002. *The Poetic Edda: Essays on Old Norse Mythology*. Routledge.

Sergio Torres Aguilar. 2025. TRIDIS: A Comprehensive Medieval and Early Modern Corpus for HTR and NER. *arXiv preprint arXiv:2503.22714*.

Kristján Árnason and Stephen Pax Leonard. 2011. Language ideology and standardisation in iceland. In *Standard Languages and Language Standards in a Changing Europe*, pages 91–96, Noregur. Novus Forlag.

Clément Besnier and William Mattingly. 2021. Named-Entity Dataset for Medieval Latin, Middle High German and Old Norse. *Journal of Open Humanities Data*, 7(23):1–5.

David J Birnbaum, Sheila Bonde, and Mike Kestemont. 2017. The Digital Middle Ages: An Introduction. *Speculum*, 92(S1):S1–S38.

Thibault Clérice, Ariane Pinche, Malamatenia Vlachou-Efstathiou, Alix Chagué, Jean-Baptiste Camps, Matthias Gille Levenson, Olivier Brisville-Fertin, Federico Boschetti, Franz Fischer, Michael Gervers, et al. 2024. CATMuS Medieval: A Multilingual Large-scale Cross-century Dataset in Latin Script for Handwritten Text Recognition and Beyond. In *International Conference on Document Analysis and Recognition*, pages 174–194. Springer.

Carol J. Clover and John Lindow. 2019. *Old Norse-Icelandic Literature: A Critical Guide*. Cornell University Press, Ithaca, NY.

Margaret Clunies Ross. 2000. *Old Icelandic Literature and Society*. Cambridge: Cambridge University Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mª Luisa Díez Platas, Salvador Ros Muñoz, Elena González-Blanco, Pablo Ruiz Fabo, and Elena Álvarez Mellado. 2021. Medieval Spanish (12th–15th centuries) Named Entity Recognition and Attribute Annotation System Based on Contextual Information. *Journal of the Association for Information Science and Technology*, 72(2):224–238.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named Entity Recognition and Classification in Historical Documents: A Survey. *ACM Computing Surveys*, 56(2):1–47.

Maud Ehrmann, Matteo Romanello, Alex Flückiger, and Simon Clematide. 2020. Extended Overview of CLEF HIPE 2020: Named Entity Processing on Historical Newspapers. In *CEUR Workshop Proceedings*, (No. 2696). CEUR-WS.

Maud Ehrmann, Matteo Romanello, Sven Najem-Meyer, Antoine Doucet, and Simon Clematide. 2022. Overview of HIPE-2022: Named Entity Recognition and Linking in Multilingual Historical Documents. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 423–446, Cham. Springer International Publishing.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Odd Einar Haugen. 2004. Parallel Views: Multilevel Encoding of Medieval Nordic Primary Sources. *Literary and Linguistic Computing*, 19(1):73–91.

Haibo He and Edwardo A. Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284.

Amanda Hilmarsson-Dunn and Ari Páll Kristinsson. 2010. The language situation in Iceland. *Current Issues in Language Planning*, 11(3):207–276.

Even Hovdhaugen, Carol Henriksen, Fred Karlsson, and Bengt Sigurd. 2000. *The History of Linguistics in the Nordic Countries*. Societas Scientiarum Fennica.

Frédéric Kaplan and Isabella di Lenardo. 2017. Big Data of the Past. *Frontiers in Digital Humanities*, 4:12.

Dick Kasperowski, Karl-Magnus Johansson, and Olof Karsvall. 2024. Temporalities and Values in an Epistemic Culture: Citizen Humanities, Local Knowledge, and AI-Supported Transcription of Archives. *Archives and Manuscripts*, 51(2):3–22.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Hrafn Loftsson and Eiríkur Rögnvaldsson. 2007. IceNLP: A Natural Language Processing Toolkit for Icelandic. In *Interspeech 2007*, pages 1533–1536.

Joseph Nockels, Paul Gooding, and Melissa Terras. 2024. The Implications of Handwritten Text Recognition for Accessing the Past at Scale. *Journal of Documentation*, 80(7):148–167.

Alexandra Petrulevich, Agnieszka Backman, and Jonathan Adams. 2020. Medieval macrospace through gis: the norse world project approach. *The Cartographic Journal*, 57(1):18–27.

Eiríkur Rögnvaldsson, Anton Karl Ingason, Einar Freyr Sigurðsson, and Joel Wallenberg. 2012. The Icelandic Parsed Historical Corpus (IcePaHC). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1977–1984. European Language Resources Association (ELRA).

Jenna Schoen and Gianmarco E Saretto. 2022. Optical Character Recognition (OCR) and medieval Manuscripts: Reconsidering Transcriptions in the Digital Age. *Digital Philology: A Journal of Medieval Cultures*, 11(1):174–206.

Stefan Schweter and Johannes Baiter. 2019. Towards Robust Named Entity Recognition for Historic German. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 96–103, Florence, Italy. Association for Computational Linguistics.

Vésteinn Snæbjarnarson, Haukur Barri Símonarson, Pétur Orri Ragnarsson, Svanhvít Lilja Ingólfsdóttir, Haukur Jónsson, Vilhjalmur Thorsteinsson, and Hafsteinn Einarsson. 2022. A Warm Start and a Clean Crawled Corpus - A Recipe for Good Language Models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4356–4366, Marseille, France. European Language Resources Association.

Merve Tekgürler. 2025. LLMs for Translation: Historical, Low-Resourced Languages and Contemporary AI Models. In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, pages 227–237, Albuquerque, New Mexico. Association for Computational Linguistics.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*, pages 1–4. Association for Computational Linguistics.

Sergio Torres Aguilar. 2022. Multilingual Named Entity Recognition for Medieval Charters Using Stacked Embeddings and BERT-based Models. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 119–128, Marseille, France. European Language Resources Association.

Sergio Torres Aguilar and Dominique Stutzmann. 2021. Named Entity Recognition for French Medieval Charters. In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, pages 37–46, NIT Silchar, India. NLP Association of India (NLPAI).

Sverrir Tómasson. 2006. Old Icelandic Prose. In Daisy Neijmann, editor, *A History of Icelandic Literature*, volume 5 of *Histories of Scandinavian Literature*, pages 64–173. University of Nebraska Press and The American-Scandinavian Foundation, Lincoln, Nebraska and London.

Xiaochen Wang and Yue Wang. 2022. Sentence-Level Resampling for Named Entity Recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2151–2165, Seattle, United States. Association for Computational Linguistics.

12. Language Resource References

Ásmundur Alma Guðjónsson, Hrafn Loftsson, and Jón Friðrik Daðason. 2021. Icelandic NER API - ensemble model (21.09). CLARIN-IS.

Odd Einar Haugen. 2015. MUFI character recommendation, version 4.0. Medieval Unicode Font Initiative.

Odd Einar Haugen. 2019. The Menota Handbook: Guidelines for the Electronic Encoding of Medieval Nordic Primary Sources. Gen. ed. Odd Einar Haugen. Version 3.0. Bergen: Medieval Nordic Text Archive.

Svanhvít Lilja Ingólfsdóttir, Ásmundur Alma Guðjónsson, and Hrafn Loftsson. 2020. MIM-GOLD-NER – named entity recognition corpus (21.09). CLARIN-IS.

Hrafn Loftsson. 2019. IceNLP Natural Language Processing Toolkit. CLARIN-IS.

Menota. 2025. Medieval Nordic Text Archive. Menota Consortium, hosted at CLARINO Bergen Centre, University of Bergen.

Hiroki Nakayama. 2018. seqeval: A Python Framework for Sequence Labeling Evaluation. GitHub.

TEI Consortium. 2025. TEI P5: Guidelines for electronic text encoding and interchange. Version 4.10.2. Text Encoding Initiative Consortium.

Joel C. Wallenberg, Anton Karl Ingason, Einar Freyr Sigurðsson, and Eiríkur Rögnvaldsson. 2024. Icelandic parsed historical corpus (IcePaHC) 2024.03. CLARIN-IS.

Appendix

Text	Date	Transcription
Alexanders saga (AM 519a 4to)	∼1280	Dipl., Norm.
Njáls saga fragm. (AM 162B θ fol)	1300–1350	Dipl., Norm.
Njáls saga fragm. (AM 162B κ fol)	1325–1375	Dipl., Norm.
Drauma-Jóns saga (AM 657a-b 4to)	1350–1399	Dipl., Norm.
Íslendingabók (AM 113b fol)	1650–1700	Dipl., Norm.
Læknísbók (AM 655 XXX 4to)	1250–1300	Dipl., Norm.
Óláfs saga helga (Lbs fragm. 82)	1258–1264	Dipl., Norm.
Vǫluspá (AM 544 4to)	1290–1360	Dipl. only
Codex Wormianus (AM 242 fol)	∼1350	Dipl. only

Table 6: Menota texts used for NER training and evaluation. Dipl. = Diplomatic, Norm. = Normalised. Seven texts have both diplomatic and normalised transcriptions; two are diplomatic only. The normalised texts were used for the normalised NER model, and diplomatic for the diplomatic NER model.

Text	Date
Jarteinabók	1200-1220
Þorláks saga helga	1200-1220
Sturlunga saga	1271-1284
Þetubrot Egils Sögu	∼1250
Jómsvíkinga saga	1250-1275
Morkinskinna	1200-1275

Table 7: IcePaHC texts used for normalised NER training and evaluation. All texts are in normalised form and date to the 13th century.

Text	Date
Alexanders saga (AM 519a 4to)	∼1280
Njáls saga fragm. (AM 162B θ fol)	1300–1350
Njáls saga fragm. (AM 162B κ fol)	1325–1375
Brennu-Njáls saga (AM 132 fol)	1320–1350
Drauma-Jóns saga (AM 657a-b 4to)	1350–1399
Egils saga (AM 132 fol)	1320–1350
Finnboga saga (AM 132 fol)	1330–1370
Íslendingabók (AM 113b fol)	1650–1700
Kormáks saga (AM 132 fol)	1330–1370
Læknísbók (AM 655 XXX 4to)	1250–1300
Laxdæla saga (AM 132 fol)	1320–1350
Óláfs saga helga (Lbs fragm. 82)	1258–1264
Víga-Glúms saga (AM 132 fol)	1330–1370
Vǫluspá (AM 544 4to)	1290–1360
Codex Wormianus (AM 242 fol)	∼1350

Table 8: Menota texts used for MLM domain adaptation. This corpus consists of the nine entity-annotated diplomatic texts (without their annotations) and six additional unannotated diplomatic texts.

Model	Entity	P	R	F1	Support
Normalised: (M + I)^R + MIM					
	Location	0.83	0.96	0.89	263
	Person	0.92	0.95	0.93	997
	micro avg	0.90	0.95	0.93	1,260
Diplomatic: M^R + I + MIM (domain-adapted)					
	Location	0.79	0.68	0.73	232
	Person	0.76	0.85	0.80	922
	micro avg	0.77	0.81	0.79	1,154

Table 9: Per-entity precision (P), recall (R), and F1 scores for the best-performing normalised and diplomatic NER models on their respective test sets. Support shows the number of entities in the test set.

Neural Machine Translation for Coptic-French: Strategies for Low-Resource Ancient Languages

Nasma Chaoui, Richard Khoury
Université Laval, Department of Computer Science and Software Engineering
Québec City, Canada
nasma.chaoui.1@ulaval.ca, richard.khoury@ift.ulaval.ca

Abstract

This paper presents the first systematic study of strategies for translating Coptic into French. Our comprehensive pipeline systematically evaluates: pivot versus direct translation, the impact of pre-training, the benefits of multi-version fine-tuning, and model robustness to noise. Utilizing aligned biblical corpora, we demonstrate that fine-tuning with a stylistically-varied and noise-aware training corpus significantly enhances translation quality. Our findings provide crucial practical insights for developing translation tools for historical languages in general.

Keywords: Machine Translation, Statistical and Machine Learning Methods, Low-resource languages

1. Introduction

Neural Machine Translations (NMT), with its steady growth, has profoundly impacted the field of computational linguistics. Transformer models and large-scale multilingual systems have enabled translation coverage across hundreds of language pairs. However, many historical languages — especially ancient ones with scarce data — have yet to benefit from these advancements. Coptic is one such language. As the last stage of the Egyptian language family, understanding and translating it is crucial for advancing fields such as Egyptology and Early-Christian studies. As a result, the language is studied in universities around the world; according to one listing[1], over 50 universities in 13 countries offer Coptic courses. But despite this international interest, there are currently few NMT tools for Coptic, and only to translate to English or Arabic.

This paper aims to fill that gap by investigating a wide range of NMT strategies to translate Coptic into French. In the process, we will focus on four important questions: (1) should we translate directly to our target language or use English as a pivot language; (2) how should we pick a language model to fine-tune for this task; (3) should we provide the model with a diverse set of possible translations for a passage; and (4) how can we make the model robust to noise in the original manuscript. While our focus is solely on Coptic→French NMT, these questions are of general interest, and our results will help guide the development of NMT tools between any ancient-modern language pair.

The remainder of this paper is organized as follows: Section 2 reviews relevant work on natural language processing for Coptic. Section 3 details the methodology we use to explore our four questions. Section 4 presents our experimental setup: our dataset, NMT models and metrics, and Section 5 presents the results we obtain on the experiments about each of the four questions. Finally, Section 6 has some concluding remarks. Our dataset and code are available on our GitHub repository[2], while the best Coptic→French NMT models we trained are distributed on HuggingFace [3] [4].

2. Background

Coptic occupies a unique position as the final stage of the Egyptian language family, bridging thousands of years from Old Egyptian through Demotic. It holds critical historical and cultural significance for Egyptology, biblical and early monasticism studies, as many Coptic manuscripts illuminate these fields.

In recent years, substantial progress has been made in building digital resources for Coptic. The Coptic SCRIPTORIUM project (Schroeder and Zeldes, 2016) stands out as a foundational effort, providing annotated corpora with multiple linguistic layers, including part-of-speech tagging, lemmatization, morphological analysis, and syntactic parsing (Zeldes and Schroeder, 2016). Similarly, tools such as the BabyLemmatizer have extended lemmatization and POS tagging capabilities across Coptic, Demotic, and Earlier Egyptian, demonstrating the shared linguistic complexities within the Egyptian language family (Sahala and Lincke, 2025).

Recent work has also explored NMT within the Egyptian language family, particularly focusing on

[1] Listing assembled by https://www.cmcl.it/~iacs/course.htm.

[2] https://github.com/chaouin/coptic-french-nmt

[3] https://huggingface.co/chaouin/coptic-french-translation-helsinki

[4] https://huggingface.co/chaouin/coptic-french-translation-hiero

Proceedings of LT4HALA 2026 @ LREC 2026, pages 482–490.
11 May 2026. *©ELRA Language Resources Association (ELRA), 2026*

hieroglyphic texts. Notably, (Cao et al., 2024) developed a Hieroglyphic Transformer by adapting the multilingual M2M-100 model to translate ancient Egyptian hieroglyphs into both German and English. Their approach involved constructing a carefully curated dataset from the *Thesaurus Linguae Aegyptiae*, which includes hieroglyphic inscriptions paired with transliterations, POS tags, and translations. They relied on transliteration (romanization) to render hieroglyphic sequences compatible with modern NLP model vocabularies, and demonstrated that transfer learning from existing multilingual architectures could achieve promising results even on low-resource ancient scripts.

The authors of (Saeed et al., 2024) introduced the largest parallel corpus to date for Coptic, comprising approximately 24,000 Coptic-English sentence pairs and 8,000 Coptic-Arabic pairs, and used it to train the first NMT system from Coptic to Arabic and English. They used both basic Seq2Seq models and fine-tuned multilingual models such as mT5 and M2M-100. Their work emphasized the necessity of romanization due to Coptic characters being absent from standard vocabularies. Similarly, (Enis and Megalaa, 2024) implemented Coptic→English NMT by fine-tuning multilingual models on normalized and romanized texts from the Coptic SCRIPTORIUM. They employed back-translation and explored multiple data representations, demonstrating the critical impact of morphological normalization and tokenization for handling Coptic's agglutinative structure.

3. Methodology

While implementing our Coptic→French NMT model, we seek to study four fundamental design questions that are common to most if not all ancient-modern language translation projects. Thus, our methodology is divided into four stages, each building on the findings of the previous.

3.1. Pertinence of Fine-Tuning

To begin, we seek to answer a fundamental question: is it even worthwhile to fine-tune a NMT model for Coptic→French translation? After all, while no such model exists, there are alternatives. The first one would be to use a more general multilingual NMT model that can work out of the box without any fine-tuning. Another alternative could use a NMT model that can translate the ancient language into another pivot language, and prompt it to answer in our target modern language. In our case, this will mean using a readily-available Coptic→English NMT model and prompting it to answer in French. Finally, since NMT models trained to translate between pairs of modern languages abound today, we could again start from an NMT model that can translate the ancient language into a pivot language, and use a second NMT model to translate that content into our target modern language. In our case, we could daisy-chain a Coptic→English and an English→French NMT model to accomplish this. Consequently, the first stage of our research consists in comparing the translations produced by a fine-tuned Coptic→French NMT model and these three alternatives, to see the differences between these four options.

3.2. Pre-Trained Model Choice

While there are not a lot of ancient→modern NMT models, there are a handful of them available, and therefore a question arises: how to pick which one to fine-tune? There are technical and linguistic considerations in this decision. From a technical standpoint, the different models available may have different numbers of parameters and be trained using different quantities of data (but given the scarcity of ancient-language data, they will typically not have different data sources). And linguistically, when no model exists for our ancient and modern language pair, three options are available. The first is to take a model trained to translate the ancient language to a different modern language and fine-tune it to our target language, or in our case fine-tune a Coptic-English model to translate to French. The second option is to take a model trained to translate into the modern language and fine-tune it to the ancient language. In our case this would mean taking any of the available multilingual NMT models trained for French and fine-tune it to translate from Coptic. Finally, a NMT translation model may exist for a different ancient language in the same family, giving us a third option. As mentioned previously, Coptic is the final stage of the Egyptian language, so a NMT model trained for an earlier stage of the language could be fine-tuned for Coptic.

3.3. Fine-Tuning Dataset Choice

If we want to fine-tune an existing model, we need a corpus of parallel text in the ancient and modern languages. However, for a variety of reasons, an ancient-language text may be translated to a modern language in several different ways. We thus conduct an experiment to see whether or not fine-tuning can benefit from this stylistic diversity. Our hypothesis is that having multiple different translations of a same original text will make it possible to build a larger training corpus from the scarce ancient language resource, while also allowing the NMT model to better learn concepts of syntactic variation and word synonymy.

3.4. Training Robustness to Noise

Ancient-language texts often suffer from corruption due to the passage of time. In our final experiment, we consider the benefit of training a NMT model using noisy data as a way of building in robustness to noise. We consider two sources of noise. The first are lacunae that result from missing or damaged fragments of the text (Levine et al., 2024). We can introduce this problem in our training data by randomly replacing characters in the text with a "missing character" symbol. We acknowledge that this is a simplification of real manuscript damage, which typically affects entire lines or sections of a page rather than isolated characters; more realistic corruption models are a direction for future work.The second source of noise in our study is letter substitutions. This problem results from two different sources: on the one hand certain characters in a language can look similar and hand-writing can be ambiguous, and on the other optical character recognition (OCR) systems used to digitize manuscripts are prone to misrecognize characters, with one study finding that character recognition accuracy in ancient text OCR could be as low as 62% (Tzogka et al., 2021). We can simulate this problem by creating a list of similar-looking characters, and randomly replacing a character by a similar one from the list. A final source of noise is spelling variations and mistakes, which we simulate by randomly switching the order of two adjacent characters. By controlling the rate of these three types of random replacements, we can further experiment with the impact of different proportions of noise in the data.

4. Experimental Setup

4.1. Dataset and Preprocessing

To construct our parallel Coptic–French corpus, we first extracted Coptic texts from the *Coptic SCRIPTORIUM* project (Schroeder and Zeldes, 2016) , which provides richly annotated PAULA XML files. We focused on the Sahidic New and Old Testament Bibles, the most complete and standardized Coptic corpus available. We automatically parsed the XML files to extract verse-level Coptic content along with their verse ID (the biblical book, chapter and verse number).

Since most pre-trained NMT models are not compatible with the original Coptic Unicode script, we romanized the Coptic data using the `uroman` tool[5], an idea proposed by (Amrhein and Sennrich, 2020) and used for Coptic by (Enis and Megalaa, 2024). Although Coptic uses only a small number of characters beyond the Greek alphabet, romanization ensures full compatibility with the tokenizers and vocabularies of all four models used in our study, and is consistent with prior NMT work on Coptic and related ancient languages (Cao et al., 2024).

For the French text, we collected three public-domain Bible translations : the Louis Segond, Crampon, and Darby versions. These were chosen for their stylistic diversity, modern relevance, and free availability. Alignment with the Coptic text was performed automatically using verse ID.

Next, we performed a cleaning step to ensure a high-quality parallel corpus. Any verse pair missing either the Coptic or French translation was eliminated. We also removed structurally mismatched pairs by filtering out entries where the Coptic segment consisted only of ellipses (e.g., "[...]"), indicating incomplete or missing verses. Additionally, we discarded verses whose French translation was missing or blank, and stripped leading inline annotations (e.g., (1.2)) from the French verses to preserve only the canonical content. All removed entries were logged separately to ensure reproducibility. After cleaning, we retained 23,561 aligned verses for both the Segond and Darby Bibles, and 23,718 verses for the Crampon version, for a total of 70,840 aligned pairs from 63 biblical books.

We split this dataset into a training and testing corpus. We opted for a book-level split, keeping the books of *1 Corinthians*, *Mark*, *Galatians* and *Hebrews* as test data and the other 59 books as training data. This represents a total of 1,460 Coptic test verses, each with 3 French translations. This setup prevents data contamination, where similar verses from a same book could be found in the training and testing corpora. It also simulates a real-world use case, of a new never-before-seen Coptic document being discovered.

4.2. Pre-trained NMT Models

We selected four pre-trained NMT models for our study. They were the most related ones to our challenge of Coptic→French translation.

- **Enis**: The first-ever Coptic→English NMT model, by Enis and Megalaa (2024)[6].
- **Helsinki**: A multilingual model from the University of Helsinki Tiedemann et al. (2023) trained on over 100 languages using a massive parallel Bible corpus, including Coptic (using a New Testament text different from the SCRIPTORIUM version) and French[7].
- **T5**: A general-purpose encoder–decoder model trained by Google on the C4 corpus

[5] https://github.com/isi-nlp/uroman

[6] https://huggingface.co/megalaa/coptic-english-translator

[7] https://huggingface.co/Helsinki-NLP/opus-mt-tc-bible-big-mul-mul

(Raffel et al., 2020), with exposure to French. It does not include Coptic. [8]

- **Hiero**: A NMT model fine-tuned by (Cao et al., 2024) for Hieroglyphic→English and Hieroglyphic→German translation[9].

Unless specified differently, all models were fine-tuned for 15 epochs on our training corpus.

4.3. Metrics

As our dataset clearly demonstrates, a single Coptic statement can have several different but equivalent French translations. Thus, a good model is one that correctly translates the intended meaning of the original statement, even if the wording differs from the gold-standard reference. To evaluate translation quality, we focus on metrics that measure semantic similarity rather than lexical overlap.

- **BERTScore** (Zhang et al., 2020) computes token-level similarity using contextual embeddings from a pre-trained BERT model. Scores range from 0 to 1, where higher values indicate greater semantic similarity between the candidate and reference sentences. It is well-suited to evaluate semantic equivalence, particularly when multiple valid versions of a statement exist.
- **BLEURT** (Sellam et al., 2020) is a learned evaluation metric that fine-tunes BERT-based models to predict human judgments of text quality. It combines semantic similarity, fluency, and grammaticality into a single score, and has been widely used to evaluate machine translation outputs (Pu et al., 2021). BLEURT scores are unbounded but typically center around [0, 1], with higher values indicating better quality.
- **COMET** (Rei et al., 2020) is a learned evaluation metric specifically designed for machine translation. It trains a multilingual BERT model to predict human judgment of translation quality. COMET scores are in the range of 0 to 1, with higher scores correlating with higher human appreciation of the translation.
- **METEOR** (Banerjee and Lavie, 2005) is a machine translation metric based on unigram matching, but that takes into account synonymy, stemming, and paraphrases, and that has been adapted to reward semantically-valid phrasings rather than exact lexical matches. METEOR scores are normalized between 0 and 1, with 1 indicating a perfect match.

[8] https://huggingface.co/t5-base

[9] https://huggingface.co/mattiadc/hiero-transformer

By considering these metrics together, we can obtain a more holistic evaluation of translation quality, one that accounts for meaning preservation and text quality and is tolerant of semantically-equivalent lexical changes.

5. Results and Analysis

5.1. Baseline

We begin by establishing a set of baselines using the models of section 4.2. To do this, each model is evaluated by translating our test corpus books in the most relevant language pair they are trained for: Coptic→English for Enis, Coptic→French for Helsinki, and English→French for T5. As no Biblical texts exist in hieroglyphic form to evaluate the Hiero model, we tested it using its own testing corpus sentences, which is composed of 50 Hieroglyph→English pairs and 50 Hieroglyph→German pairs.

Model	BERT Score	BLEURT	COMET	METEOR
Enis	0.632	0.344	0.328	0.119
Helsinki	0.595	0.136	0.248	0.082
T5	0.809	0.501	0.673	0.429
Hiero	0.833	0.555	0.664	0.496

Table 1: Baseline pre-trained model performances.

5.2. Pertinence of Fine-Tuning

Our first experiment compares the four different strategies to achieve Coptic→French translation described in Section 3.1. First is the strategy of fine-tuning an existing Coptic translation model, the Enis model, to a French target. The second strategy is to use a multilingual model that includes both Coptic and French out of the box, namely the Helsinki model. The third strategy is to prompt the Enis model to translate to French without fine-tuning, in a zero-shot manner. Finally, we will use English as a pivot language, by using Enis to translate from Coptic to English then translating that output to French using T5.

Approach	BERT Score	BLEURT	COMET	METEOR
Fine-tuning	**0.820**	**0.402**	**0.613**	**0.467**
Multilingual	0.595	0.136	0.248	0.082
Prompting	0.606	0.149	0.238	0.041
Pivot language	0.620	0.115	0.274	0.100

Table 2: Comparison of translation strategies.

The results in Table 2 clearly demonstrates that the benefits of fine-tuning justify the extra labor,

compared to the other three methods. The Enis model generates better translations in French after being fine-tuned than it did in English in our baseline results in Table 1, thanks to the extra training. Prompting Enis to translate to French without fine-tuning gives the worst results of the three Enis tests, due to mismatched decoding expectations. The model fine-tuned to a single pair of languages also outperforms the multilingual model, a result coherent with the literature, which finds that multiligual models perform suboptimally for low-resource languages (such as Coptic) (Joshi et al., 2025). Finally, it is interesting to see that the Enis-T5 pair achieves worse results on all metrics than either Enis or T5 did in the baseline test of Table 1. This indicates that translation errors in each step are compounded by the sequence of translations, reducing overall accuracy. This matches the observations of (Paul et al., 2013).

5.3. Pre-Trained Model Choice

Having confirmed that fine-tuning a model for Coptic→French is worth the effort, our second experiment explores the question of Section 3.2, namely which model should be fine-tuned. For this experiment, we fine-tune each model on our entire training corpus. This enables Enis to learn to translate to French, T5 to translate from Coptic, and Hiero to translate from Coptic to French. While Helsinki was already trained on Coptic and French among its 100 languages, it was also fine-tuned to become specialized to those two languages.

Model	BERT Score	BLEURT	COMET	METEOR
Enis	0.820	0.402	0.613	0.467
Helsinki	**0.850**	**0.564**	**0.731**	**0.563**
T5	0.763	0.240	0.474	0.325
Hiero	0.832	0.498	0.681	0.501

Table 3: Comparison of fine-tuned models.

The results translating our test corpus, presented in Table 3, show that Helsinki outperforms the other models on all four metrics. This performance can be attributed to two key advantages. First, its pre-training corpus includes both the source and target languages, unlike the other models that lacked one or both of the languages and had to learn them from scratch. Second, it's the model whose pre-training corpus is most aligned with our task, being also a Bible corpus. The Helsinki model is also relatively large, with 248M parameters, which may contribute to its strong performance. However, we cannot exclude the possibility that it is becoming over-fitted to translating biblical texts; since we do not have non-biblical Coptic-French translations to test on, we cannot check that possibility.

The next best model on all metrics is Hiero. Interestingly, this one has the opposite properties from Helsinki: it is trained on neither Coptic, French, nor biblical text. It is however trained for Hieroglyphics, a writing system related to Coptic. This seems to indicate that knowledge learned on a language transfers well to related languages. A similar observation was made by (Babych et al., 2007) in the case of NMT using pivot languages. The Enis model is a close third place. While that model benefits from having been pre-trained on Coptic, it is smaller than the Hiero model, with 77.1M against 484M parameters for Hiero, and thus doesn't learn the new language as well. Finally, the T5 model, with 223M parameters, which does not have any Coptic nor related language in its pre-training data, achieves the worst fine-tuning results. It should be noted that given the significant differences in model size, not all models may have fully converged after 15 epochs; a systematic study of training hyperparameters such as the number of epochs and early stopping criteria is left for future work.

5.4. Fine-Tuning Dataset Choice

To test our hypothesis of Section 3.3, we fine-tune the two best models of Section 5.3, Helsinki and Hiero, on each of our three French translations individually. For this experiment, since each single-source model trains on roughly one-third of the complete training dataset, we train them for 45 epochs instead of 15 for the model that trains on the entire dataset. We tested each model on each translation of the test corpus separately, to observe the impact each different training has.

Test set	Training set	BERT Score	BLEURT	COMET	METEOR
Crampon	All datasets	0.836	**0.549**	**0.723**	0.513
Crampon	Crampon only	**0.849**	0.537	0.715	**0.583**
Crampon	Darby only	0.821	0.494	0.690	0.460
Crampon	Segond only	0.825	0.519	0.703	0.472
Darby	All datasets	0.858	**0.576**	**0.737**	0.587
Darby	Crampon only	0.825	0.511	0.696	0.493
Darby	Darby only	**0.868**	0.558	0.729	**0.615**
Darby	Segond only	0.846	0.535	0.710	0.546
Segond	All datasets	0.856	**0.569**	**0.732**	0.587
Segond	Crampon only	0.828	0.517	0.699	0.514
Segond	Darby only	0.844	0.515	0.701	0.550
Segond	Segond only	**0.865**	0.564	0.727	**0.611**

Table 4: Comparison of dataset choice on Helsinki.

The results, given in Tables 4 and 5, show that the models trained on a specific dataset translation perform best when tested on that same translation in two of our four metrics, BERTScore and METEOR. These are the two metrics that value seman-

Test set	Training set	BERT Score	BLEURT	COMET	METEOR
Crampon	All datasets	0.815	**0.480**	**0.672**	0.445
Crampon	Crampon only	**0.825**	0.458	0.655	**0.493**
Crampon	Darby only	0.805	0.433	0.642	0.410
Crampon	Segond only	0.806	0.448	0.646	0.413
Darby	All datasets	0.844	**0.517**	**0.692**	0.535
Darby	Crampon only	0.812	0.450	0.648	0.444
Darby	Darby only	**0.847**	0.493	0.678	**0.547**
Darby	Segond only	0.830	0.472	0.660	0.489
Segond	All datasets	0.837	**0.495**	**0.680**	0.519
Segond	Crampon only	0.810	0.442	0.642	0.443
Segond	Darby only	0.826	0.451	0.648	0.487
Segond	Segond only	**0.839**	0.479	0.663	**0.521**

Table 5: Comparison of dataset choice on Hiero.

tic similarity to the correct translation. By training and testing on the same translation, these models learn the correct choice of wording and translation decisions, and thus can generate a translation that is very similar to the expected one. By contrast, when training on one translation and testing on another, these systems all yield much weaker BERTScore and METEOR performances, showing they are generating translations worded differently from the expected result.

On the other hand, the model trained on all three dataset translations always achieves the top BLEURT and COMET scores, no matter which dataset translation it is tested on. These two metrics are tuned to replicate human appreciation of the generated text's quality. By training on a variety of possible translations for each Coptic sentence, that model has learned to generate better translations overall. In addition, this model always achieves the second-best BERTScore and METEOR scores, behind the model trained on the same test corpus translation but ahead of the two trained on other translations. So while the wording generated by this model is not optimal compared to the correct translation, it is always closer than any other model not specialized to generate that specific translation. Overall, this shows that stylistic diversity in the training data improves generation quality at inference time, especially when multiple valid translations are acceptable.

5.5. Training Robustness to Noise

As explained in Section 3.4, we simulate real-world degradation of the manuscripts by injecting noise in our training and testing data. To simulate textual corruption, each character in a Coptic verse is subjected to three independent noise processes: a 2% probability of deletion (simulating lacunae), a 2% probability of switch with an adjacent character (simulating spelling errors), and a 10% probability of replacement by a visually similar character (simulating OCR errors), using the visual confusion mapping of Figure 1, which we constructed manually based on the visual similarity of Coptic character shapes. These rates were chosen heuristically to introduce meaningful variation while preserving overall text readability, with the higher rate for visual substitution reflecting the greater prevalence of OCR errors compared to physical lacunae in digitized manuscripts (Tzogka et al., 2021). We do not alter the French translations of the verses. We experiment with different levels of simulated degradation by creating 5 versions of our corpora, setting the probability that a verse is corrupted to 0%, 10%, 30%, 50%, and 100%.

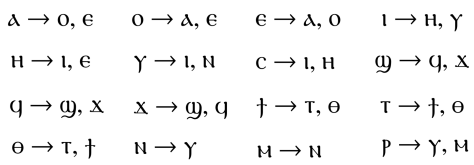


Figure 1: Visual confusion of Coptic characters.

Test noise	Train noise	BERT Score	BLEURT	COMET	METEOR
0%	0%	0.850	0.564	0.731	0.563
0%	10%	**0.852**	**0.568**	**0.733**	**0.564**
0%	30%	0.849	0.560	0.727	0.560
0%	50%	0.848	0.553	0.723	0.559
0%	100%	0.846	0.547	0.719	0.546
10%	0%	0.844	0.545	0.717	0.549
10%	10%	**0.849**	**0.557**	**0.725**	**0.557**
10%	30%	0.847	0.553	0.722	0.555
10%	50%	0.846	0.545	0.718	0.554
10%	100%	0.844	0.544	0.717	0.543
30%	0%	0.829	0.492	0.679	0.511
30%	10%	0.842	0.534	0.708	0.538
30%	30%	0.842	**0.539**	0.711	0.545
30%	50%	**0.843**	0.536	**0.712**	**0.546**
30%	100%	0.841	0.533	0.709	0.536
50%	0%	0.817	0.457	0.650	0.478
50%	10%	0.837	0.514	0.695	0.522
50%	30%	**0.838**	0.522	0.701	**0.533**
50%	50%	**0.838**	0.519	0.698	**0.533**
50%	100%	**0.838**	**0.524**	**0.702**	0.529
100%	0%	0.788	0.366	0.579	0.411
100%	10%	0.822	0.471	0.665	0.488
100%	30%	0.826	0.491	0.675	0.502
100%	50%	0.829	0.496	0.683	0.509
100%	100%	**0.831**	**0.505**	**0.688**	**0.513**

Table 6: Comparison of robustness to noise for Helsinki.

The results in Tables 6 and 7 show several clear trends. First, both models trained with any level of noise perform better on clean test data, and their scores drop steadily as noise increases. This shows that texts with errors or gaps are naturally harder to translate than clean ones. Likewise, both the Helsinki and Hiero models trained with low noise do better on low-noise data, while those

Test noise	Train noise	BERT Score	BLEURT	COMET	METEOR
0%	0%	**0.832**	0.497	**0.682**	**0.500**
0%	10%	**0.832**	0.496	0.679	0.496
0%	30%	0.831	**0.498**	0.679	0.494
0%	50%	0.830	0.492	0.674	0.492
0%	100%	0.827	0.482	0.668	0.482
10%	0%	0.825	0.474	0.665	0.485
10%	10%	0.828	0.484	0.670	0.488
10%	30%	**0.829**	**0.489**	**0.674**	**0.489**
10%	50%	**0.829**	0.485	0.671	0.488
10%	100%	0.826	0.478	0.665	0.480
30%	0%	0.809	0.412	0.619	0.445
30%	10%	0.822	0.452	0.649	0.470
30%	30%	**0.825**	0.465	**0.658**	**0.476**
30%	50%	0.824	**0.466**	0.656	**0.476**
30%	100%	0.824	**0.466**	0.657	0.473
50%	0%	0.795	0.364	0.582	0.416
50%	10%	0.815	0.426	0.630	0.453
50%	30%	0.819	0.447	0.643	0.463
50%	50%	**0.821**	0.452	0.647	**0.466**
50%	100%	**0.821**	**0.456**	**0.650**	**0.466**
100%	0%	0.762	0.232	0.486	0.335
100%	10%	0.800	0.366	0.592	0.411
100%	30%	0.810	0.405	0.616	0.436
100%	50%	0.813	0.419	0.627	0.445
100%	100%	**0.815**	**0.426**	**0.629**	**0.451**

Table 7: Comparison of robustness to noise for Hiero.

trained with more noise perform better on noisy data. For Helsinki, the model trained with 10% noise leads early on, then the 50% model takes over in mid-noise settings, and the 100% model does best at the end. For Hiero, the 0% model performs best early, the 30% model leads in the middle, and the 100% model dominates in the last tests.

What's really happening is that models trained with less noise do better on clean data, but they're also more fragile: their scores drop sharply as noise rises. To quantify this fragility, Table 8 shows, for each model and training noise level, the relative drop in score between the *cleanest* test condition (0% test noise) and the *noisiest* one (100% test noise) for each metric. Each row corresponds to one model under one training noise setting, and the columns report the relative score drops for BERTScore, BLEURT, COMET, and METEOR. Models trained with 0% and 10% noise lose around 20% on average—Hiero's BLEURT score even drops by 53%. In contrast, models trained with 50% or 100% noise lose less than 15%, and as little as 6% on average. This confirms that controlled noise injection creates NMT models more resilient to real-world degradation, such as OCR errors or fragmentary manuscripts, at the cost of lower performance on clean manuscripts. While the optimal choice thus depends on the characteristics of the manuscript being translated, a good general-case compromise value seems to be 50% training noise.

Model	Training noise	BERT Score	BLEURT	COMET	METEOR
Helsinki	0%	7.3%	35.1%	20.8%	27.0%
Helsinki	10%	3.5%	17.1%	9.3%	13.5%
Helsinki	30%	2.7%	12.3%	7.2%	10.4%
Helsinki	50%	2.2%	10.3%	5.5%	8.9%
Helsinki	100%	1.8%	7.7%	4.3%	6.0%
Hiero	0%	8.4%	53.3%	28.7%	33.0%
Hiero	10%	3.8%	26.2%	12.8%	17.1%
Hiero	30%	2.5%	18.7%	9.3%	11.7%
Hiero	50%	2.0%	14.8%	7.0%	9.6%
Hiero	100%	2.0%	11.6%	5.8%	6.4%

Table 8: Relative drop in score between 0% and 100% *test* noise, per model and training noise level. Lower values indicate greater robustness to noisy input.

6. Conclusion

This paper presents the first systematic study of neural machine translation for the Coptic→French language pair. Through four experiments, we outlined best practices for building a translation model for this new pair. First, we showed that task-specific fine-tuning clearly improves translation quality compared to using pre-trained multilingual models or translating through a pivot language. Next, we found that fine-tuning a model already trained on the ancient language or a related one in the same family (such as Hieroglyphics) yields better results. Third, we examined the effect of having multiple translations of the same text, and found that including them all enhanced translation quality. Finally, we tested the impact of noise in the dataset and found that the best trade-off between translation quality and robustness was achieved when training with 50% corrupted verses. This led to the release of two strong NMT models for this task, one based on the multilingual model of (Tiedemann et al., 2023) and the other on the Hieroglyphic translator of (Cao et al., 2024), both fine-tuned on three French versions of the training corpus with 50% noise.

An important direction for future work is the evaluation of our models on non-biblical Coptic texts. Our study is necessarily limited to biblical material, as no parallel Coptic→French corpus exists outside this domain. Evaluating generalization to other genres, could be achieved through backtranslation and constitutes a key objective for future research.

Together, our findings extend beyond Coptic→French translation, providing practical guidance for building robust NMT systems in other low-resource or ancient language settings.

7. Acknowledgements

This research was made possible by the support of the Natural Sciences and Engineering Research

Council of Canada (NSERC) USRA and Discovery Grant programs, and the Fonds de recherche du Québec (FRQ).

8. Bibliographical References

Chantal Amrhein and Rico Sennrich. 2020. On romanization for model transfer between scripts in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2461–2469. Association for Computational Linguistics.

Bogdan Babych, Anthony Hartley, and Serge Sharoff. 2007. Translating from under-resourced languages: Comparing direct transfer against pivot translation. In *Proceedings of Machine Translation Summit XI*, Copenhagen, Denmark.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures*, pages 65–72.

Mattia Cao, Nicola De Cao, Angelo Colonna, and Alessandro Lenci. 2024. Deep learning meets egyptology: A hieroglyphic transformer for translating ancient egyptian. In *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages*, pages 71–86.

Maxim Enis and Andrew Megalaa. 2024. Ancient voices, modern technology: Low-resource neural machine translation for coptic texts. *Coptic Translator*, pages 1–15.

Raviraj Joshi, Kanishk Singla, Anusha Kamath, Raunak Kalani, Rakesh Paul, Utkarsh Vaidya, Sanjay Singh Chauhan, Niranjan Wartikar, and Eileen Long. 2025. Adapting multilingual LLMs to low-resource languages using continued pretraining and synthetic corpus. In *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*, pages 50–57. Association for Computational Linguistics.

Lauren Levine, Cindy Li, Lydia Bremer-McCollum, Nicholas Wagner, and Amir Zeldes. 2024. Lacuna language learning: Leveraging RNNs for ranked text completion in digitized coptic manuscripts. In *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages*, pages 61–70. Association for Computational Linguistics.

Michael Paul, Andrew Finch, and Eiichiro Sumita. 2013. How to choose the best pivot language for automatic translation of low-resource languages. *ACM Transactions on Asian Language Information Processing*, 12(4):1–17.

Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for machine translation. In *Proceedings of EMNLP*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of EMNLP*, pages 2685–2702. Association for Computational Linguistics.

Muhammed Saeed, Asim Mohamed, Mukhtar Mohamed, Shady Shehata, and Muhammad Abdul-Mageed. 2024. From nile sands to digital hands: Machine translation of coptic texts. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 298–308.

Aleksi Sahala and Eliese-Sophia Lincke. 2025. Neural models for lemmatization and POS-tagging of earlier and late egyptian. In *Proceedings of the Second Workshop on Ancient Language Processing*, pages 99–104. Association for Computational Linguistics.

Caroline T. Schroeder and Amir Zeldes. 2016. Raiders of the lost corpus. *Digital Humanities Quarterly*, 10(2).

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of ACL*. Association for Computational Linguistics.

Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato, Yves Scherrer, Raul Vazquez, and Sami Virpioja. 2023. Democratizing neural machine translation with OPUS-MT. *Language Resources and Evaluation*, 58:713–755.

Christina Tzogka, Fotini Koidaki, Stavros Doropoulos, Ioannis Papastergiou, Efthymios Agrafiotis, Katerina Tiktopoulou, and Stavros Vologiannidis. 2021. Ocr workflow: Facing printed texts of ancient, medieval and modern greek literature. *Journal of Documentation*.

Amir Zeldes and Caroline T. Schroeder. 2016. An NLP pipeline for coptic. In *Proceedings of the*

10th SIGHUM Workshop on Language Technology for Cultural Heritage, pages 146–155.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Author Index

Abela, Charlie, 407
Amblard, Maxime, 72
Andersen, Nicklas Sindlev, 139
Asahara, Masayuki, 89, 448
Atzori, Denise, 339

Bawden, Rachel, 453
Beersmans, Marijke, 152
Ben-Gigi, Nati, 96
Bernstein, Tal, 197
Biffi, Marco, 49
Bin, Li, 244
Bin, Liu, 305
Bizais-Lillig, Marie, 339
Boano, Valeria Irene, 152, 226
Borg, Claudia, 380, 407
Brigada Villa, Luca, 368
Brisson, Colin, 311
Buhagiar, Vanessa, 407
Bui, Marc, 311

Cai, Dongheng, 299
Caine, Daniel, 58
Caraffa, Edoardo, 361
Casagrande, Gianluca, 38
Cerqueglini, Letizia, 197
Chan, Callum, 234
Changshun, Quan, 305
Chaoui, Nasma, 482
Chengfei, Li, 305
Cinkova, Silvie, 1
Clérice, Thibault, 453
Colavizza, Giovanni, 361
Constant, Frédéric, 311
Cui, Wenhui, 112

Darwin, Gregory, 125
de Daele, Ine, 437
de Graaf, Evelien, 152
Depauw, Mark, 152, 437
Deret, Jules, 219
Dezotti, Lucas Consolin, 353

Ellul, Charlene, 407

Fan, Qi, 277
Fantoli, Margherita, 152
Favaro, Manuel, 49
Fratellini, Erica, 368

Gao, Xin, 244
Garnier, Mathias, 339
Gershuni, Hillel, 58
Ginevra, Riccardo, 368
Glaise, Anthony, 453
Gordin, Shai, 197
Guadagnini, Elisa, 49

He, Liqi, 322
Héder, Mihály, 184
Henningsson, Pontus, 469
Hsieh, Shu-Kai, 163
Hu, Jieming, 277

Iurescia, Federica, 211, 353

Kaalep, Heiki-Jaan, 400
Kahfy, Ayoub, 311
Kato, Sachi, 89
Katzoff, Binyamin, 96
Keersmaekers, Alek, 152
Khoury, Richard, 482
Kikuchi, Rei, 89
Komiya, Kanako, 89
Koolen, Marijn, 389
Koppens, Thomas, 380
Korovina, Evgeniya, 177

Láng, Benedek, 184
Lenas, Erik, 469
Létoffé, Maxime, 339
Li, Jieqiong, 244
Li, Qiwei, 322
Li, Zuchao, 322
Liang, KeYan, 263
Llebeskind, Shmuel, 96
Lim, Chunki, 163
Lin, Chih-Han, 163
Lindenbaum, Yitzchak, 58
Lindqvist, Ellinor, 19

Litta, Eleonora, 226, 368
Liu, Chang, 244
Liu, Liu, 244
Liu, Meiling, 263
Liu, Yulong, 299

MacDougall, Byron, 139
Mambrini, Francesco, 106
Marini, Costanza, 38
Marongiu, Paola, 418
Matsuda, Hiroshi, 448
Megyesi, Beata, 184
Meng, Yuchun, 294
Miani, Irene, 125
Modesti, Maddalena, 361
Montemagni, Simonetta, 49
Moretti, Giovanni, 353
Muñoz Acebes, Francisco Javier, 443

Nicolás-Flores, Fernando, 443
Nijs, Julie, 152
Nikolova-Stoupak, Iglika, 72
Nivre, Joakim, 19

Ohara, Mitoki, 89

Paccosi, Teresa, 389
Palmero Aprosio, Alessio, 38
Pan, Tung-Le, 163
Papáček, Aleš Manuel Manuel, 7
Passarotti, Marco, 106, 211, 353, 368
Pettersson, Eva, 19, 469
Pinche, Ariane, 453
Pizzone, Aglae, 139
Planque, Charles, 339
Polomac, Vladimir, 1
Pommeret, Luc, 219
Principe, Claudia, 38

Rattenborg, Rune, 184
Raziel-Kretzmer, Vered, 58
Rey, Frédérique, 72
Ripoll-Alberola, Luisa, 443
Romanello, Matteo, 226
Rosensweig, Elisha, 58

Sassolini, Eva, 49, 418
Schler, Jonathan, 96
Shmidman, Avi, 58
Shu, Yuhan, 330
Smith, David, 453
Soen, Violet, 437
Sprugnoli, Rachele, 211
Ströbel, Phillip Benjamin, 112
Stymne, Sara, 125, 429
Swaelens, Colin, 106

Taihe, Cao, 305
Tan, Yuyang, 299
Trusca, Maria Mihaela, 437
Tseng, Guan-Yu, 163

Van de Cruys, Tim, 152, 437
Verbruggen, Kevin, 437
Vidal-Gorène, Chahan, 339

Wagret, Thibault, 219
Waldispühl, Michelle, 184
Wang, Dongbo, 244
Wang, Yilin, 299
Wang, Yu-Chieh, 163
Wen, Xin, 268
Wu, Ruifeng, 244
Wu, Zhongzheng, 244

Xia, Tian, 299
XiaoYi, Li, 305

Yang, Fan, 244
Yang, Menghui, 299
Yang, Yumeng, 299
Yang, Ziye, 322
Ye, Chen, 277
Yeh, Lang-Ching, 163
Yin, Tianjie, 339
Yin, Yihuan, 273
Yousef, Tariq, 139
Yue, Zhu, 244
Yunjie, Zhang, 305

Žabokrtský, Zdeněk, 7
Zanchi, Chiara, 368
Zhang, Chaokun, 268
Zhang, Shibingfeng, 361
Zhao, Qian, 273
Zhao, Xue, 244
Zheng, Chenrui, 288
Zhitomirsky-Geffet, Maayan, 96
Zhixiao, Zhao, 244
Zhou, Huizi, 330
Zhou, Tongtong, 268
Zhu, Dongmei, 244
Zhu, Hang, 89
Zuffrano, Annafelicia, 361