

VarDial 2026

**VarDial 2026 - The Thirteenth Workshop on
NLP for Similar Languages, Varieties and Dialects**

Proceedings of the Workshop

March 29, 2026

©2026 The Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-372-2

Preface

These proceedings include the 32 papers presented at the Thirteenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2026), co-located with the 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL). VarDial was held in Rabat, Morocco.

This year, VarDial saw a record number of submissions, with 44 research papers submitted to the workshop, 25 of which were accepted to appear in these proceedings. We extended the programme committee to be able to handle the increased reviewing load and would like to thank all new and returning PC members for being an important part of the workshop's success!

The accepted papers focus on a large number of language varieties, including Tamil, Finnish, Arabic, Ukrainian, Greek, Galician, and Piedmontese, among others. Similarly, the NLP tasks addressed in the papers are very diverse, including topics such as syllabification and quantifying linguistic distances as well as sentiment analysis and generating texts with LLMs. Notably, six of the accepted papers focus on spoken language input.

As in previous editions, VarDial 2026 features an evaluation campaign. This year, the AMIYA shared task (Arabic Modeling In Your Accent) attracted six participating teams whose system description papers are included in these proceedings, along with a report by the shared task organizers. We thank all the shared task organizers and the participants for their hard work!

The VarDial workshop organizers:

Yves Scherrer, Noëmi Aeppli, Verena Blaschke, Tommi Jauhiainen,
Nikola Ljubešić, Preslav Nakov, Jörg Tiedemann, and Marcos Zampieri

<http://sites.google.com/view/wardial-2026/>

Organizing Committee

Organizers:

Yves Scherrer, University of Oslo (Norway)

Noëmi Aepli, University of Pennsylvania (USA)

Verena Blaschke, LMU Munich and Munich Center for Machine Learning (Germany)

Tommi Jauhiainen, University of Helsinki (Finland)

Nikola Ljubešić, Jožef Stefan Institute and University of Ljubljana (Slovenia)

Preslav Nakov, Mohamed bin Zayed University of Artificial Intelligence (UAE)

Jörg Tiedemann, University of Helsinki (Finland)

Marcos Zampieri, George Mason University (USA)

Program Committee

Program Committee:

César Aguilar (Universidad Veracruzana, Mexico)
Sina Ahmadi (George Mason University, United States)
Jorge Baptista (University of Algarve and INESC-ID, Portugal)
Delphine Bernhard (University of Strasbourg, France)
Gabriel Bernier-Colborne (National Research Council, Canada)
David Chiang (University of Notre Dame, United States)
Steven Coats (University of Oulu, Finland)
Çağrı Çöltekin (University of Tübingen, Germany)
Stefanie Dipper (Ruhr University Bochum, Germany)
Sascha Diwersy (University of Montpellier, France)
Mark Dras (Macquarie University, Australia)
Jonathan Dunn (University of Illinois Urbana-Champaign, United States)
Fahim Faisal (George Mason University, USA)
Pablo Gamallo (University of Santiago de Compostela, Spain)
Ona de Gibert Bonet (University of Helsinki, Finland)
Rob van der Goot (IT University Copenhagen, Denmark)
Cyril Goutte (National Research Council, Canada)
Jindřich Helcl (University of Oslo, Norway)
Radu Ionescu (University of Bucharest, Romania)
Aditya Joshi (UNSW, Australia)
Hour Kaing (NICT, Japan)
Anjali Kantharuban (Carnegie Mellon University, United States)
Amr Keleg (MBZUAI, UAE)
Olli Kuparinen (Tampere University, Finland)
Taja Kuzman Pungeršek (Jožef Stefan Institute, Slovenia)
John McCrae (University of Galway, Ireland)
Aleksandra Miletic (CNRS, France)
Filip Miletic (University of Stuttgart, Germany)
John Nerbonne (University of Groningen, Netherlands and University of Freiburg, Germany)
Ekaterina Lapshinova-Koltunski (University of Hildesheim, Germany)
Lung-Hao Lee (National Yang Ming Chiao Tung University, Taiwan)
Maciej Ogrodniczuk (Institute of Computer Science, Polish Academy of Sciences, Poland)
Petya Osenova (Bulgarian Academy of Sciences, Bulgaria)
Siyao (Logan) Peng (LMU Munich, Germany)
Alistair Plum (University of Luxembourg, Luxembourg)
Jelena Prokic (Leiden University, Netherlands)
Christoph Purschke (University of Luxembourg, Luxembourg)
Alan Ramponi (FBK Trento, Italy)
Francisco Rangel (Symanto Research, Spain)
Reinhard Rapp (University of Mainz, Germany)
Tanja Samardžić (IDSIA, Switzerland)
Serge Sharoff (University of Leeds, United Kingdom)
Milena Slavcheva (Bulgarian Academy of Sciences, Bulgaria)
Aarohi Srivastava (University of Notre Dame, United States)
Joel Tetreault (Dataminr, United States)
Samia Touileb (University of Bergen, Norway)
Taro Watanabe (Google Inc., Japan)

Table of Contents

<i>AMIYA Shared Task: Arabic Modeling In Your Accent at VarDial 2026</i> Nathaniel R. Robinson, Shahd Abdelmoneim, Anjali Kantharuban, Otba Alsboul, Salima Lamsiyah, Kelly Marchisio and Kenton Murray	1
<i>Far Out: Evaluating Language Models on Slang in Australian and Indian English</i> Deniz Kaya Dilsiz, Dipankar Srirag and Aditya Joshi	18
<i>Effects of Speaker Bias in Dialect Identification and Automatic Transcription with Self-Supervised Speech Models</i> Olli Kuparinen	32
<i>OcWikiDialects: A Wikipedia Dataset With Rich Metadata for Occitan Dialect Identification</i> Oriane Nédey, Rachel Bawden, Thibault Clérice and Benoît Sagot	45
<i>Language Mixture to Develop Accurate Galician Dependency Parsers: An Exploration of Its Effects</i> Xabier Irastortza-Urbieta, José M. García-Miguel and Marcos Garcia	58
<i>Crowdsourcing Piedmontese to Test LLMs on Non-Standard Orthography</i> Gianluca Vico and Jindřch Libovický	70
<i>German-English Code-Switching in Large Language Models</i> Firat Cem Aksüt, Stefan Hillmann, Pia Knoeferle and Sebastian Möller	87
<i>Perplexity as a Metric for Dialectal Distance: A Computational Study of Greek Varieties</i> Stergios Chatzikyriakidis, Erofilis Psaltaki, Dimitrios Papadakis, Erik Henriksson and Veronika Laippala	101
<i>A Subword Embedding Approach for Variation Detection in Luxembourgish User Comments</i> Anne-Marie Lutgen, Alistair Plum and Christoph Purschke	113
<i>Onomasiological Sense Alignment Across Dialect Dictionaries. A Taxonomy-Constrained LLM Classifi- cation</i> Nathalie Mederake, Nico Urbach, Hanna Fischer and Alfred Lameli	123
<i>On the Intelligibility of Romance Language Varieties: Spanish and Portuguese in Europe and America</i> Liviu P. Dinu, Ana Sabina Uban, Teodor-George Marchitan, Ioan-Bogdan Iordache and Simona Georgescu	139
<i>Dialect Matters: Cross-Lingual ASR Transfer for Low-Resource Indic Language Varieties</i> Akriti Dhasmana, Aarohi Srivastava and David Chiang	145
<i>Ara-HOPE: Human-Centric Post-Editing Evaluation for Dialectal Arabic to Modern Standard Arabic Translation</i> Abdullah Alabdullah, Lifeng Han and Chenghua Lin	157
<i>Indic-TunedLens: Interpreting Multilingual Models in Indian Languages</i> Mihir Panchal, Deeksha Varshney, Mamta . and Asif Ekbal	172
<i>Building ASR Resources for the Hutsul Dialect of Ukrainian</i> Roman Kyslyi, Artem Orlovskyi, Pavlo Khomenko, Bohdan Onyshchenko and Zakhar Guzii . .	186

<i>From FusHa to Folk: Exploring Cross-Lingual Transfer in Arabic Language Models</i> Abdulmuizz Khalak, Abderrahmane Issam and Gerasimos Spanakis	196
<i>Extending ASR Evaluation Resources for Modern Greek Dialects</i> Chara Tsoukala, Stavros Bompolas, Antigoni Margariti, Konstantina Panagiotou, Maria Elisavet Plaiti, Nefeli Tzanakaki, Petros Karatsareas, Angela Ralli, Antonios Anastasopoulos and Stella Markantonatou	210
<i>How Should We Model the Probability of a Language?</i> Rasul Dent, Pedro Ortiz Suarez, Thibault Cl�rice and Beno�t Sagot	223
<i>Bridging Dialectal Variation: A Phonetic Transcription Tool for Tamil</i> Ahrane Mahaganapathy, Sumirtha Karunakaran, Kavitha Navakulan and Kengatharaiyer Sarveswaran	234
<i>Regional Variation in the Performance of ASR Models on Croatian and Serbian</i> Tanja Samard�i�, Peter Rupnik and Nikola Ljube�i�	242
<i>Syllable Structures Across Arabic Varieties</i> Abdelrahim Qaddoumi, Jordan Kodner, Salam Khalifa, Ellen Broselow and Owen Rambow ...	250
<i>Curriculum Learning and Pseudo-Labeling Improve the Generalization of Multi-Label Arabic Dialect Identification Models</i> Ali Mekky, Mohamed El Zeftawy, Lara Hassan, Amr Keleg and Preslav Nakov	261
<i>OpenLID-v3: Improving the Precision of Closely Related Language Identification – An Experience Report</i> Mariia Fedorova, Nikolay Arefyev, Maja Buljan, Jindřich Helcl, Stephan Oepen, Egil R�nningstad and Yves Scherrer	275
<i>Improving Dialect Robustness in Large Language Models via LoRA and Mixture-of-Experts</i> Sanjh Maheshwari, Aniket Singh Rajpoot, Oana Cocarascu and Mamta	293
<i>Evaluation Framework for Transfer Learning between Closely Related Lects: A Case Study of Lemko</i> Iliia Afanasev	304
<i>Do Large Language Models Adapt to Language Variation across Socioeconomic Status?</i> Elisa Bassignana, Mike Zhang, Dirk Hovy and Amanda Cercas Curry	317
<i>Aladdin-FTI @ AMIYA Three Wishes for Arabic NLP: Fidelity, Diglossia, and Multidialectal Generation</i> Jonathan Mutal, Perla Al Almaoui, Simon Hengchen and Pierrette Bouillon	339
<i>Maastricht University at AMIYA: Adapting LLMs for Dialectal Arabic using Fine-tuning and MBR Decoding</i> Abdulhai Alali and Abderrahmane Issam	352
<i>SDNLP at AMIYA 2026: Syrian Arabic Dialect Modeling with LoRA</i> Hasan Alkhder and Mohammad Abboush	359
<i>NUS-IDS at AMIYA/VarDial 2026: Improving Arabic Dialectness in LLMs with Reinforcement Learning</i> Sujatha Das Gollapalli, Mouad Hakam, Mingzhe Du and See-Kiong Ng	365
<i>MBZUAI at AMIYA Shared Task 2026: Adapting Open-Source LLMs for Dialectal Arabic</i> Rana Gaber, Yara Allam, Serag Amin, Ranwa Aly and Bashar Alhafni	373

A Closed-Track System for Palestinian Arabic in the AMIYA Shared Task

Khaleel Hamad and Ahmad Al-Najjar 385

AMIYA Shared Task: Arabic Modeling In Your Accent at VarDial 2026

Nathaniel R. Robinson¹, Shahd Abdelmoneim², Anjali Kantharuban³, Otba Alsoul,
Salima Lamsiyah⁴, Kelly Marchisio⁵, Kenton Murray¹

¹Johns Hopkins University, ²Cohere Labs Community, ³Carnegie Mellon University,
⁴University of Luxembourg, ⁵Cohere

Correspondence: nrobin38@jhu.edu, amiyasharedtask@gmail.com

Abstract

Arabic, often considered a single language, actually describes a wide variety of sometimes mutually unintelligible language varieties. While large language models (LLMs) have revolutionized natural language processing (NLP) with rapid advances, these models still best serve speakers of high-resource and standard language varieties. One particular deficiency of theirs is in dialectal Arabic. We present the first ever shared task for dialectal Arabic language modeling: **Arabic Modeling In Your Accent**, or AMIYA. The goal of the shared task was to develop LLMs that could (1) respond in the correct dialectal variety when explicitly or implicitly prompted to, (2) translate between dialectal Arabic and standard Arabic or English, (3) adhere to LLM instructions in dialectal Arabic, and (4) produce fluent Arabic outputs. We called for submissions in the dialectal varieties of five countries: Morocco, Egypt, Palestine, Syria, and Saudi Arabia. We received 45 submitted systems from six participating teams. We saw positive results from supervised fine-tuning on a translation objective, and reinforcement learning to improve dialectness. Manual evaluation also showed that some systems had learned to output dialectal words or phrases, but at the expense of actual fluency or coherence. Overall the most effective system involved continual pre-training and supervised fine-tuning of 12 candidate LLMs, followed by selection of the best performing models.

1 Introduction and Related Work

Recent advances in LLMs have demonstrated improvements in text generation in Modern Standard Arabic (MSA). However, performance remains limited when it comes to generating content in dialectal Arabic (Keleg et al., 2023). LLMs are typically proficient in high-resource language varieties, including Modern Standard Arabic (MSA) (Singh et al., 2024; Robinson et al., 2023), but they typi-

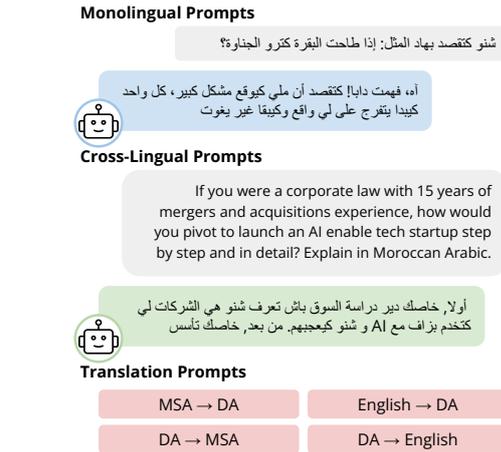


Figure 1: Evaluation across three task settings: monolingual instructions, cross-lingual instructions, and translation (in four directions). Example outputs are Moroccan Arabic from Maastricht University's (UM) primary submission, which achieved the highest human evaluation scores.

cally have poor or unreliable proficiency in dialectal Arabic (DA) language varieties.

DA varieties often lack support in NLP applications, in part because of researchers' tendency to view Arabic as a "monolith" (Bergman and Diab, 2022). Contrary to this common perception, there are 28 different ISO 639-3 specifications¹ for distinct Arabic language varieties.

Robinson et al. (2025) found that while LLMs demonstrate a growing capacity to comprehend text in DA, they continue to struggle with generating dialectal text fluently. Specifically, many LLMs understand some DA varieties and are able to model them, but that they frequently fail to do so because of a strong preference for MSA. This is problematic for a number of reasons. While MSA is useful in many contexts, it is not used in many others.

¹https://wikipedia.org/wiki/ISO_639_macrolanguage

According to the Ethnologue,² MSA does not have native speakers. Comparatively, DA varieties like Egyptian and Moroccan Arabic (i.e. Moroccan Darija) have tens of millions of speakers. Many native speakers of DA varieties are not proficient in MSA, and those who are tend to be more educated and socially advantaged (Bergman and Diab, 2022). Hence LLMs’ singular proficiency in MSA may exacerbate, rather than ameliorate, existing inequalities. Robinson et al. (2025) suggest that different pre-training, post-training, and prompting methods may be used to mitigate LLMs’ difficulty with DA.

To address this gap, we designed the first shared task to evaluate and improve dialectal Arabic generation: **Arabic Modeling In Your Accent**, or **AMIYA**.³ For AMIYA we adopt the *AL-QASIDA* evaluation methodology (Robinson et al., 2025), which evaluates the dialectal fidelity, understanding, quality, and diglossia of LLMs when processing and generating DA. Our goal is to develop LLMs proficient in five major Arabic varieties, and provide the research community with tools and insights to bring about future developments.

2 Shared Task Overview

The AMIYA shared task evaluates LLM proficiency in Dialectal Arabic for varieties from five countries: Morocco, Egypt, Palestine, Syria, and Saudi Arabia. This evaluation is conducted using the *AL-QASIDA* methodology: both automatically and with human scoring on a subset of samples (Robinson et al., 2025). There are three available tracks: closed data, closed model, and open.

2.1 Closed Data Track

For the closed data track, the only permitted training data is provided by the organizers, but any open-source LLM can be used (along with their pre-trained weights). The training set for this submission track contains data from 17 different data sources, which we describe here.

MADAR 26 The MADAR 26 bitext contains 12k sentences of multi-way parallel training data across eight Arabic dialects, English, and Modern Standard Arabic (MSA) (Bouamor et al., 2018). It additionally includes 2k sentences for each of 26 city-level Arabic dialects. The corpus is high-quality, and is composed of translations of the En-

glish Basic Traveling Expressions Corpus (BTEC) (Takezawa et al., 2007). Its sentences are conversational and concise. The most recent test split⁴ was off limits to participants.

SauDial The Saudi Arabic Dialects Game Dataset is a multi-parallel bitext of 1001 lines from 4 Saudi Arabian dialects (Hejazi, Najdi, Eastern, and Southern), English, and MSA (Alanazi et al., 2025). The dataset focuses on topics related to localization for gaming and includes information on each games’ age rating and any relevant cultural context.

ASR-EgArbCSC The Egyptian Arabic Speech Corpus⁵ contains 5.5 hours of conversational Egyptian Arabic transcribed into 3.2k lines.

MASC Corpus The Multi-Arabic Sentiment Corpus includes 6k lines in three divisions: product reviews, political comments, and software comments (Al-Moslmi et al., 2018). Each sentence is labeled by two native speakers with positive or negative sentiment. Each annotator’s native country is included as well.

Maknuune Maknunne is an open lexicon for Palestinian Arabic with 36k entries from 17k lemmas and 3.7k roots (Dibas et al., 2022). Entries include the diacritic Arabic orthographic form, phonological transcription, and English gloss.

DODa The Darija Open Dataset contains ~50k sentences in Moroccan Arabic aligned with English translations (Outchakoucht and Es-Samaali, 2024). The dataset contains entries written both in Arabic- and Latin-based writing systems to reflect real-world usage patterns.

Shami Corpus The Shami Corpus contains approximately 10k tweets in each of four varieties: Lebanese, Palestinian, Jordanian, and Syrian (Abu Kwaik et al., 2018).

Atlaset The Atlaset dataset is a curated collection of Moroccan text data (Bounhar, 2025). The training set includes 155M tokens from sources such as social media posts and news articles. The dataset encompasses both formal and informal registers across domains.

²<https://www.ethnologue.com/>

³"Amiya," or "عامية" is the Arabic word for dialectal Arabic itself.

⁴corpus-6-test-corpus-26-test

⁵<https://magichub.com/datasets/egyptian-arabic-conversational-speech-corpus/>

SDC The Saudi Dialect Corpus contains social media data spanning various dialects of Arabic from Saudi Arabia (Tarmom et al., 2020). In total, the dataset includes 200k words.

Saudi Tweets Corpus The Saudi Tweets Corpus includes 200k tweets from Saudi Arabian users (Alruily, 2020). The dataset is cleaned of Twitter artifacts such as emojis, retweet markers, and hashtags.

SADSLyC The Saudi Arabian Dialects Song Lyrics Corpus contains 31k lines of dialectal song lyrics from the five major Saudi varieties: Najdi, Hijazi, Shamali, Janoubi, and Shargawi (Alahmari, 2025). The dialect of each song is ascertained using the hometown or birthplace of its author.

EDGAD The Egyptian Dialect Gender Annotated Dataset contains about 200k tweets, labeled with the gender of the writer (Hussein et al., 2019). Each tweet comes from a user with at least 1k tweets, and gender is determined via manual annotation, using profile information, tweet content, and external information (for public figures).

EDC The Egyptian Dialect Corpus (EDC) includes 200k words, across 13k lines, of Egyptian Arabic collected from Facebook (Tarmom et al., 2020).

Casablanca Casablanca is an Arabic speech corpus with transcriptions for samples in dialects from eight countries: Algeria, Egypt, United Arab Emirates, Jordan, Mauritania, Morocco, Palestine, and Yemen (Talafha et al., 2024). The sentences also include annotations on gender, dialect, and code switching.

JODA The Jordanian Dialect Arabic (JODA) dataset contains social media text, transcriptions of films, and existing dialect corpora (Abandah et al., 2025). These 50k sentences have been manually annotated by expert linguists and translated into MSA.

UFAL The UFAL parallel corpus includes 120k sentences from OpenSubtitles (Sellat et al., 2023). These sentences are translated into MSA and Northern Levantine Arabic, in addition to English, French, German, Greek, and Spanish.

Goud-Sum The Goud-Sum dataset is a set of 189k articles, with both headlines and categories (Aftiss et al., 2025). These articles are in Moroccan

Arabic and/or MSA, scraped from the GOUD news website.⁶

2.2 Closed Model Track

For this track, teams must train their LLMs from scratch—that is, pre-trained weights are not allowed. They can use any data except data that has been explicitly restricted for testing.

2.3 Open Track

For this track, teams may use any pre-trained, open-source LLM and any available data except that which is restricted for the test set.

2.4 Evaluation Methodology

Our evaluation set was composed of subsets of four datasets: Palm (Alwajih et al., 2025), FLORES-200 (Goyal et al., 2022; NLLB Team et al., 2022), Arena-Hard (Li et al., 2024), and MADAR-26 (Bouamor et al., 2018; Takezawa et al., 2007).

Palm is a dialectally diverse Arabic dataset composed of LLM prompts. We initially intended to use its designated test set for our evaluation. However, the test portion did not have enough data in each of our dialects of interest. Hence, we used portions of the Palm train set and repurposed them for evaluation. The plurality of data in the set are MSA, and we found through manual inspection that even many of the sentences labeled as a DA variety were also MSA. To mitigate this we filtered the sentences in our dialects of focus by using ALDi (Keleg et al., 2023) to measure their dialectness. We only kept sentences scoring above 0.5 (indicating that they are more dialectal than standard). A native Arabic speaker manually inspected a portion of the filtered sentences to verify that the automatic filtering was effective. This filtering yielded a total of 649 sentences for Moroccan Arabic, 283 for Egyptian, 419 for Palestinian, 181 for Syrian, and 141 for Saudi.

As DA LLM instructions, the Palm sentences were meant to evaluate LLMs’ monolingual generation ability to produce the same DA variety as the input. For additional data of this same kind, we used the first 200 sentences of FLORES-200 (NLLB Team et al., 2022) and the first 200 sentences of the MADAR-26 corpus⁶-train split (Bouamor et al., 2018) in each dialect. We converted these generic sentences to LLM instructions via AL-QASIDA, exactly as Robinson et al. (2025).

⁶<https://www.goud.ma>

To evaluate LLMs’ ability in cross-lingual generation—i.e., producing an explicitly requested DA variety to respond to an English prompt, we used 220 of the structured English LLM prompts from Arena-Hard (Li et al., 2024). We converted 110 of these into LLM instructions requesting a DA output by hand. We used Gemini⁷ to do the same to another 110 and manually verified that they were correct (with only minor error corrections needed). We made copies of these 220 total sentences, requesting each of our five DA varieties by simple string replacement.

Just as Robinson et al. (2025), we also availed the multi-dialectal multi-parallel nature of the FLORES-200 and MADAR-26 data to create evaluation sets for English→DA, DA→English, MSA→DA, and DA→MSA translation (with each input formatted to be a translation instruction). Figure 1 illustrates the three types of LLM prompts included in our evaluation.

We compute ADI2 dialectal fidelity score (Robinson et al., 2025) to evaluate all outputs that we expect to be DA, and chrF++ (Popović, 2017, 2015) to evaluate all translations.

In addition to these automatic metrics, native Arabic speakers manually evaluated a total of 100 outputs from each team’s primary system (25 from each of the data sources, excluding data intended for translation). Our evaluator for Palestinian and Syrian Arabic is a native speaker of Levantine Arabic from Jordan, who has lived many years in Syria. Our evaluator of Egyptian and Saudi outputs is a native speaker of both Egyptian and Gulf Arabic, and our evaluator of Moroccan outputs is a native speaker of Moroccan Arabic.

Our evaluation is patterned after the AL-QASIDA methodology (Robinson et al., 2025). We measure dialectal fidelity with ADI2 score for both monolingual, cross-lingual, and translation prompts. We measure understanding by chrF++ score for DA→English MT and by human-given adequacy scores. We measure generation quality by English→DA chrF++ and by human-given fluency scores. And we measure diglossia by chrF++ for MSA↔DA.

3 Shared Task Submissions

We received a total of 45 submitted systems from six teams.

⁷<https://gemini.google.com>

3.1 Aladdin-FTI

Aladdin-FTI (or Aladdin) (Mutal et al., 2026) participated in the closed data track with a system designed to jointly model dialectal fidelity and diglossic control. Their approach fine-tuned Hugging Face’s SmolLM3-3B⁸ and Llama-3-8B-Instruct⁹ using a multi-objective training method that combines (1) instruction-conditioned next-token dialectal generation and (2) machine translation between dialects, MSA, and English. The translation objective encourages semantic adequacy and awareness of MSA–DA boundaries, while the instruction-following generation objective promotes natural dialectal realization. By interpolating the two losses, the team aimed to balance dialectal fidelity and meaning preservation (e.g. in translation). Their system supports Moroccan, Egyptian, Palestinian, Syrian, and Saudi dialects. In their experiments they compared fine-tuning (1) purely for MT, (2) purely for instruction following, and (3) their combination. They found that training with the joint objective, by combining MT instruction data and standard instruction data, provides the best trade-off between dialect fidelity and translation quality. They submitted for all five dialects, using their system based on Llama-3-8B-Instruct as their primary submission and their system based on SmolLM as contrastive1.

3.2 Brigham Young University

BYU (Hamad and Al-Najjar, 2026) made a closed data track submission for Palestinian Arabic by fine-tuning on a combination of the provided Maknuune, Shami, Casablanca, MASC, and JODA corpora (Dibas et al., 2022; Abu Kwaik et al., 2018; Talafha et al., 2024; Al-Moslemi et al., 2018; Abandah et al., 2025). Their submitted system was based on Qwen2.5-1.5B-Instruct.¹⁰ They employed parameter-efficient fine-tuning, updating approximately 0.28% of the model’s parameters, and demonstrated improved performance on both translation and generation.

3.3 Maastricht University

Maastricht University (UM) (Alali and Issam, 2026) participated in the closed data track. Their

⁸<https://huggingface.co/HuggingFaceTB/SmolLM3-3B>

⁹<https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct>

¹⁰<https://huggingface.co/Qwen/Qwen2.5-1.5B-Instruct>

system development involved three processes: Low Rank Adaptation (LoRA), adapter merging, and Minimum Bayes Risk (MBR) decoding (Hu et al., 2021; Bapna and Firat, 2019; Houlsby et al., 2019; Bickel and Doksum, 1977; Kumar and Byrne, 2004). They used LoRA to train layer adapters on two contrastive objectives: (1) unsupervised fine-tuning with a language modeling objective on DA data, and (2) supervised fine-tuning with a translation objective. They found that merging monolingual and translation-based adapters improved the balance between dialectal fidelity (measured by ADI2) and semantic fidelity (measured by chrF++). They compared use of Llama 3.2 (Meta, 2024) and JAIS-2 (Sengupta et al., 2023) as their base models and found that while Llama 3.2 had superior ADI2, JAIS-2 was a stronger translator and had more balanced scores overall. As a final step they used MBR to ensemble different checkpoints, using ADI2 and chrF++ as objectives. They submitted systems based on JAIS-2 for Syrian, Moroccan, and Saudi dialects.

3.4 Mohammed Bin Zayed University of Artificial Intelligence

The MBZUAI team (Gaber et al., 2026) explored fine-tuning methods for the closed data track. They explored 12 different LLMs to use as an initialization, including some instruction-tuned models and some base models. The methodology was to fine-tune the base models using continual pre-training (CPT), then to fine-tune both base and instruct models with instruction fine-tuning. They explored both curriculum and mixed data approaches to handling DA, MSA, and English in the provided data for CPT. They created instruction tuning data from MADAR-26 data with an MT instruction template, and then took a subset of the same data and synthetically created instructions for monolingual and cross-lingual generation (using provided templates for the former and Command series models from Cohere for the latter). They tested which models performed best for each dialect off the shelf, then explored which models showed promise in fine-tuning on data subset. The six best overall candidate models were selected for full fine-tuning. They selected the three best performing models for each of the five dialects as their primary, contrastive1, and contrastive2 submissions.

3.5 National University of Singapore - Institute of Data Science

NUS-IDS (Das Gollapalli et al., 2026) made a closed data track submission by employing a unique training approach. Their initial step was to train LLMs for Arabic dialect identification and inter-dialectal translation. They then pass prompts from various data categories (including translation, summarization, and question-answering) through an LLM to construct a dataset for reinforcement learning (RL), i.e. preference tuning. If the trained dialect classifier finds the output to be the right dialect, it is labeled as a positive example. Otherwise, it is labeled as negative, and the trained translator creates a second output by translating into the correct dialect, which is labeled as positive. The team used this preference data for RL and found improvements in ADI2, at the expense of worsened translation metrics. Their primary submission for all five dialects was Llama-3.1-8B-Instruct¹¹ fine-tuned in this way. Their contrastive submissions used Qwen-2.5-7B-Instruct¹² (Yang et al., 2024; Team, 2024) post-trained as a translator; contrastive1 was trained with RL as well afterwards, while contrastive2 was not.

3.6 Syrian Dialect Arabic NLP

SDNLP (Alkhder and Abboush, 2026) was the only team to submit to the open track. They focused on Syrian Arabic adaptation based on the integration of parameter-efficient fine-tuning with prompt-guided inference. They used their own 30k-utterance dataset of caption data from Turkish television shows dubbed in Syrian Arabic. They employed Low-Rank Adaptation (LoRA) (Hu et al., 2022) to adapt a pretrained instance of Llama-3.1-8B to model Syrian Arabic.

4 Evaluation Results

We discuss results from both automatic and human evaluations across systems. Because all but one submission were in the closed data track, we compare all results together.

4.1 Automatic Scores

Results are in Tables 1 - 5. Best system results are **bold**, as are baseline results when they remained unbested. Table 1 shows overall ADI2 scores

¹¹<https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>

¹²<https://huggingface.co/Qwen/Qwen2.5-7B-Instruct>

	submission	mar	egy	pse	syr	sau
Aladdin-FTI	primary	37.8%	15.2%	7.4%	21.5%	4.9%
Aladdin-FTI	contrastive1	37.5%	15.1%	7.1%	24.5%	3.6%
BYU	primary	-	-	5.4%	-	-
MBZUAI	primary	56.9%	44.5%	9.5%	17.8%	10.1%
MBZUAI	contrastive1	54.0%	39.0%	8.6%	17.3%	8.5%
MBZUAI	contrastive2	57.5%	45.2%	10.1%	21.3%	14.2%
NUS-IDS	primary	70.1%	62.9%	12.2%	10.2%	7.3%
NUS-IDS	contrastive1	68.5%	56.0%	13.3%	36.2%	14.6%
NUS-IDS	contrastive2	32.0%	0.6%	1.3%	0.6%	0.7%
SDNLP	primary	-	-	-	28.0%	-
UM	primary	67.9%	-	-	38.9%	46.4%
Baseline		10.2%	11.1%	1.9%	1.3%	0.9%

Table 1: Average ADI2 scores across all datasets and tasks asking for DA responses. As [Robinson et al. \(2025\)](#), we use an LID filter that automatically assigns a score of 0 to sentences not identified as Arabic.

	submission	mar	egy	pse	syr	sau
Aladdin-FTI	primary	32.3	34.2	37.6	22.6	36.3
Aladdin-FTI	contrastive1	32.0	33.2	36.3	20.1	36.2
BYU	primary	-	-	13.8	-	-
MBZUAI	primary	22.7	33.6	36.9	31.0	34.1
MBZUAI	contrastive1	26.8	34.3	34.0	31.0	36.2
MBZUAI	contrastive2	25.8	31.2	30.5	25.7	32.6
NUS-IDS	primary	8.6	11.7	12.1	11.9	13.2
NUS-IDS	contrastive1	11.9	15.8	15.3	14.0	15.8
NUS-IDS	contrastive2	24.0	20.6	22.9	20.2	32.9
SDNLP	primary	-	-	-	14.8	-
UM	primary	30.0	-	-	34.4	19.8
Baseline		23.9	30.3	32.4	31.6	35.5

Table 2: Aggregate chrF++ scores for ENG→DA MT

for each submitted system. These are computed across all combined eval data subsets that expect DA responses: both monolingual prompts (Palm, FLORES, MADAR); cross-lingual (Arena-Hard); and MT (FLORES and MADAR English→DA and MSA→DA). In the Moroccan (mar), Egyptian (egy), and Palestinian (pse) dialects, NUS-IDS’ RL-tuned systems scored best, casting their unique preference tuning approach as a promising method for better dialectal fidelity. In Syrian (syr) and Saudi (sau), UM performed best. This was somewhat surprising since the only supervised data their model was fine-tuned on were MT prompts specifically, and could be an indication of the effectiveness of the unsupervised post-training they conducted prior to adapter merging.

The baseline model for automatic evaluations is Llama-3.1-8B-Instruct, run with each test prompt placed entirely in the "user" field of its input. We selected this baseline because it relatively performed

	submission	mar	egy	pse	syr	sau
Aladdin-FTI	primary	46.2	50.0	53.0	39.8	54.0
Aladdin-FTI	contrastive1	45.6	49.6	53.5	39.8	54.5
BYU	primary	-	-	11.2	-	-
MBZUAI	primary	44.9	47.7	50.4	54.1	58.4
MBZUAI	contrastive1	51.0	53.4	58.0	52.7	57.9
MBZUAI	contrastive2	44.6	43.3	50.0	48.2	52.2
NUS-IDS	primary	19.2	19.0	20.4	18.9	20.1
NUS-IDS	contrastive1	17.9	21.3	22.3	21.1	23.2
NUS-IDS	contrastive2	33.9	6.2	47.2	9.4	54.1
SDNLP	primary	-	-	-	36.2	-
UM	primary	49.9	-	-	51.9	0.0
Baseline		45.9	49.7	52.5	49.5	56.4

Table 3: Aggregate chrF++ scores for DA→ENG MT

	submission	mar	egy	pse	syr	sau
Aladdin-FTI	primary	26.6	28.8	31.2	23.2	43.2
Aladdin-FTI	contrastive1	25.1	27.5	30.7	19.4	32.1
BYU	primary	-	-	17.6	-	-
MBZUAI	primary	34.9	39.5	40.5	37.6	55.3
MBZUAI	contrastive1	35.3	43.7	40.1	36.8	56.9
MBZUAI	contrastive2	33.1	36.4	42.4	37.4	42.8
NUS-IDS	primary	11.6	17.3	16.0	16.6	19.7
NUS-IDS	contrastive1	11.2	16.5	15.5	15.7	17.5
NUS-IDS	contrastive2	32.4	12.5	30.9	9.7	60.1
SDNLP	primary	-	-	-	15.6	-
UM	primary	33.8	-	-	40.3	24.2
Baseline		30.8	42.1	40.4	39.8	61.4

Table 4: Aggregate chrF++ scores for MSA→DA MT

well on [Robinson et al.’s \(2025\)](#) evaluation. However, with this eval set and prompting method its dialectal fidelity is far below that of the submitted systems.

Table 2 shows chrF++ scores for MT of English (ENG) into DA. In this setting, Aladdin’s Llama-based system performed best on Moroccan, Palestinian, and Syrian; MBZUAI did best on Egyptian, and UM won out on Syrian. (Note also that best scores from each of these three teams are fairly close to one another across all five dialects.) Aladdin, MBZUAI, and UM (as well as NUS-IDS with their contrastive2 system) all trained on MT-specific instruction fine-tuning data and all scored fairly well on this task.

In the following tables, 3 for DA→ENG, 4 for MSA→DA, and 5 for DA→MSA, we see a common trend: MBZUAI’s systems perform best. The only exceptions to this are MSA-to-Syrian MT, which UM excelled at, and MSA-to-Saudi MT, which NUS-IDS did best (with their model

	submission	mar	egy	pse	syr	sau
Aladdin-FTI	primary	27.4	29.4	31.2	24.6	42.1
Aladdin-FTI	contrastive1	27.1	28.0	30.7	21.3	34.2
BYU	primary	–	–	17.3	–	–
MBZUAI	primary	39.3	43.9	41.6	40.3	65.3
MBZUAI	contrastive1	44.1	50.6	42.9	44.4	66.3
MBZUAI	contrastive2	37.9	39.1	41.8	34.6	53.2
NUS-IDS	primary	13.3	17.5	15.9	14.9	21.6
NUS-IDS	contrastive1	16.0	19.7	18.6	20.8	32.7
NUS-IDS	contrastive2	35.1	11.2	32.8	12.1	59.1
SDNLP	primary	–	–	–	6.7	–
UM	primary	39.5	–	–	43.4	37.2
Baseline		39.3	47.3	44.4	42.1	68.5

Table 5: Aggregate chrF++ scores for DA→MSA MT

Adherence level	Rating
The response fulfills user request completely	3
The response fulfills half or part of the user request	2
The response does not fulfill the user request at all	1

Table 6: Adherence: *did the model response fulfill all that was asked in the prompt?*

trained on an MT objective only). Specifically for the out-of-DA translation directions (Tables 3 and 5), MBZUAI’s Aya-based systems performed best; these include all of their contrastive1 systems and their primary systems for Syrian and Saudi DA. This could be in part because Aya (Dang et al., 2024; Üstün et al., 2024) was trained on a highly multilingual dataset, including explicit support for English and MSA (the target languages here).

Notice that across MT evaluations, many systems did not outperform the baseline, and even the best systems typically only outperformed it by small margins (if at all). However, we wish to point out that the baseline model’s outputs for into-DA translation, while they may achieve high chrF++, have extremely low ADI2 scores. And for into-MSA translation it may benefit from the unfortunate proximity between MSA and the DA subsets of FLORES. All of this said, we hope that in future years participants will be able to beat the baseline more convincingly in all metrics.

See Appendix B for more more detailed, genre-specific score breakdowns.

Fluency level	Rating
The response is indistinguishable from native Arabic text	5
The response is understandable, but likely not native Arabic text	4
The response is clearly not native Arabic text, recurrent disfluencies are glaring or inhibit understanding (or includes copied text from the input prompt alongside newly generated text)	3
The response contains some fluent elements, but mostly not fluent (or copies the input prompt without innovating)	2
The response is not fluent, or not Arabic	1

Table 7: Fluency: *did the model generate a response that could have been written by a native Arabic speaker?*

4.2 Human Evaluation Scores

We detail our human evaluation process and present its results.

4.2.1 Evaluation Criteria

To assess the quality of dialectal Arabic generation, we conducted a comprehensive human evaluation with three native Arabic speakers, each proficient in multiple dialectal varieties. Our evaluators included a native Gulf/Egyptian speaker, a Levantine speaker, and a Moroccan speaker, providing coverage across all five dialects included in our study. The annotators, who are also authors of this work, evaluated each team’s primary submission on a sample of 25 prompt completions each from Palm, MADAR, FLORES, and Arena-Hard; or 100 samples total per dialect per team. Each completion was assessed along two dimensions that are difficult to measure automatically: model *Adherence* and *Fluency*.

Adherence refers to whether the model response fulfilled the original user request, using a 3-point scale, displayed in Table 6. *Fluency* refers to whether the model generated a response that appears to have been written by a native Arabic speaker, with coherent sentence structure (in either DA or MSA). We used a 5-point scale to capture the various degrees of fluency, shown in Table 7.

4.2.2 Annotator Results Overview

The resulting average scores of our human evaluation are shown in tables 8 and 9. One thing is fairly simple about the results trends: MBZUAI’s submission performs best in every measurement, with the exception of UM in Moroccan fluency. For *Adherence*, most models average between 1 (“does not

	mar	egy	pse	syr	sau
Aladdin	1.78	1.80	1.17	1.12	1.66
BYU	–	–	1.26	–	–
MBZUAI	2.04	2.07	1.61	1.62	2.10
NUS-IDS	1.40	1.92	1.17	1.18	2.05
SDNLP	–	–	–	1.12	–
UM	1.97	–	–	1.15	1.12

Table 8: Adherence: *average human score per team and dialect*

	mar	egy	pse	syr	sau
Aladdin	3.15	2.77	2.68	2.44	3.15
BYU	–	–	2.99	–	–
MBZUAI	3.20	3.59	3.48	3.43	3.37
NUS-IDS	1.43	2.22	2.40	2.46	2.83
SDNLP	–	–	–	2.93	–
UM	3.37	–	–	2.63	2.38

Table 9: Fluency: *average human score per team and dialect*

fulfill the user request at all") and 2 ("fulfills half or part of the user request"); and the best performing models average between 2 and 3 ("fulfills user request completely"). In terms of *Fluency*, most models average between 2 ("contains some fluent elements, but mostly not fluent") and 3 ("clearly not native Arabic text"); and the best models perform between 3 and 4 ("understandable, but likely not native Arabic text").

Our findings suggest that knowledge of dialectal lexicon does not necessarily translate to conversational competence in a particular dialect. Even when models demonstrate facility with dialectal vocabulary, they lack the cultural and contextual knowledge necessary for accurate and disambiguous discourse. High dialectal scores alone do not necessarily result in coherent and fluent responses, as models can parrot the original prompt or use a pool of dialectal words without meaningful sentence structure. For example, Aladdin’s system achieved high ADI2 scores (shown in Table 1), but scored poorly on *Adherence* in Table 8. This is a limitation that extends beyond purely linguistic competence to encompass broader understanding of DA varieties. This pattern aligns with Robinson et al. (2025), who report that LLMs understand Dialectal Arabic better at than they generate it, and reveal a reluctance to produce dialectal output along with insufficient depth of dialect-specific knowledge.

We include detailed observations from the annotators regarding each dialect in Appendix A.

5 Conclusions

One of our conclusions from our human evaluation was that command of dialectal vocabulary does not equate to conversational fluency, as models often lack the cultural and contextual structure required to use words appropriately in the native dialect. Models can achieve high ADI2 scores by parroting prompts or inserting dialectal terms without constructing meaningful, coherent sentences, which could be misleading as a signal of fluency.

Looking at our evaluation campaign comprehensively, it seems that the MBZUAI team performed best overall. Their systems, produced by selecting the best performing of 12 candidate model initialization after CPT and SFT with an MT objective, dominated most of the MT evaluations and nearly all the human evaluations, and they performed reasonably well in dialectal fidelity as well. This speaks to the effectiveness of a simple approach combining CPT and SFT, and the importance of casting a wide net in terms of candidate models. The MBZUAI teams’ thoroughness in exploring a wide variety of LLMs paid off in their performance.

We also point out that Aladdin’s systems performed well on into-DA translation, though they performed less well in adherence to generic user requests. This indicates that the model may have become an MT specialist, which is conceivable given that it was trained on a mixture including MT supervised data. Other submissions that occasionally stood out were UM and NUS-IDS. UM performed best or nearly best on all translation directions involving Syrian Arabic. They also achieved best or nearly best ADI2 scores across all three dialects they submitted for, and their model performed nearly as well as or better than MBZUAI’s for both *Adequacy* and *Fluency* in Moroccan. The RL-trained systems of NUS-IDS generally performed poorly in MT and human evaluations, but they scored best or nearly best in ADI2 score for every dialect except Saudi. This indicates that their novel method of preference tuning with synthetic correct-dialect and incorrect-dialect completions has promise for increasing ADI2 score. In this case, it appears to have degraded other abilities, but perhaps future researchers will find a way to combine it with other objectives and minimize this loss.

Limitations

Our evaluation campaign this year left a few things to be desired. To start, our suggested method for

participants to evaluate their own systems internally was by running [Robinson et al.'s \(2025\) AL-QASIDA](#). We found that this was less than practical for three reasons. First, participants struggled to set up the proper environments to run AL-QASIDA, and the code base was less adaptable to different kinds of models than we would have liked. For future evaluations it may be wise to set up a "lite" version of AL-QASIDA that simply computes ADI2 and chrF++ scores. The original implementation has model inference built in as an intermediate step, but perhaps it would be easier for teams to do their own inference separately. This would allow for more versatility. Second, AL-QASIDA is time-consuming to run (and essentially not runnable without access to GPUs). This is more difficult to address, but could also be alleviated somewhat by a "lite" version of the benchmark. Third, different conventions about how LLMs are prompted (whether with plain text input or a structured dictionary with roles) can vary. The original AL-QASIDA evaluation process was not equipped to account for this variation, which can lead to results that seem inconsistent with previously published scores.

Another shortcoming of our campaign this year lay in our use of the FLORES dataset. Unfortunately, some of the dialectal Arabic subsets of FLORES-200 are too close to MSA. The Saudi subset, for example, is closer to the Flores-200 MSA subset itself than even a reasonable set of MSA paraphrases would be. This makes results using FLORES somewhat questionable (and virtually ensures that models scoring high on Saudi ADI2 will score poorly on into-Saudi MT, and vice versa).

Acknowledgments

We acknowledge Amr Keleg, Sanjeev Khudanpur, and Samar Magdy for their feedback and advice for this shared task.

References

Gheith A Abandah, Moath R Khaleel, Iyad F Jafar, Mohammad R Abdel-Majeed, Yousef H Hamdan, Ashraf E Suyyagh, Asma A Abdel-Karim, and Shorouq M AlAwawdeh. 2025. Jordanian arabic to modern standard arabic translation using a large model tuned on a purpose-built dataset and synthetic error injection. *Jordanian Journal of Computers & Information Technology*, 11(3).

Kathrein Abu Kwaik, Motaz Saad, Stergios Chatzikyriakidis, and Simon Dobnik. 2018. [Shami: A corpus of Levantine Arabic dialects](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Azzedine Aftiss, Salima Lamsiyah, Christoph Schommer, and Said Ouatic El Alaoui. 2025. Empirical evaluation of pre-trained language models for summarizing moroccan darija news articles. In *Proceedings of the 4th Workshop on Arabic Corpus Linguistics (WACL-4)*, pages 77–85.

Tareq Al-Moslmi, Mohammed Albared, Adel Al-Shabi, Nazlia Omar, and Salwani Abdullah. 2018. [Arabic senti-lexicon: Constructing publicly available language resources for arabic sentiment analysis](#). *Journal of Information Science*, 44(3):345–362.

Salwa Saad Alahmari. 2025. [SADSLyC: A corpus for saudi Arabian multi-dialect identification through song lyrics](#). In *Proceedings of the 4th Workshop on Arabic Corpus Linguistics (WACL-4)*, pages 38–43, Abu Dhabi, UAE. Association for Computational Linguistics.

Abdulhai Alali and Abderrahmane Issam. 2026. Maas-tricht university at AMIYA: Adapting llms for dialectal arabic using fine-tuning and mbr decoding. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, Rabat, Morocco. Association for Computational Linguistics.

Naif Alanazi, Mohammed Al-Batineh, and Hussein Abu-Rayyash. 2025. Saudial: The saudi arabic dialects game localization dataset. *Data in Brief*, page 111906.

Hasan Alkhder and Mohammad Abboush. 2026. SDNLP at AMIYA 2026: Syrian arabic dialect modeling with lora. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, Rabat, Morocco. Association for Computational Linguistics.

Meshrif Alruily. 2020. [Issues of dialectal saudi twitter corpus](#). *The International Arab Journal of Informational Technology*, 17:367–374.

Fakhraddin Alwajih, Abdellah El Mekki, Samar Mohamed Magdy, AbdelRahim A. Elmadany, Omer Nacar, El Moatez Billah Nagoudi, Reem Abdel-Salam, Hanin Atwany, Youssef Nafea, Abdulfattah Mohammed Yahya, Rahaf Alhamouri, Hamzah A. Alsayadi, Hiba Zayed, Sara Shatnawi, Serry Sibae, Yasir Ech-chammakhy, Walid Al-Dhabyani, Marwa Mohamed Ali, Imen Jarraya, and 25 others. 2025. [Palm: A culturally inclusive and linguistically diverse dataset for Arabic LLMs](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32871–32894, Vienna, Austria. Association for Computational Linguistics.

- Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. 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 1538–1548.
- A. Bergman and Mona Diab. 2022. [Towards responsible natural language annotation for the varieties of Arabic](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 364–371, Dublin, Ireland. Association for Computational Linguistics.
- Peter J Bickel and Kjell A Doksum. 1977. *Mathematical statistics: basic ideas and selected topics*. Holden-Day Inc.
- Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouni, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. [The MADAR Arabic dialect corpus and lexicon](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Abdelaziz Bounhar. 2025. Atlaset: A moroccan darija text corpus. <https://huggingface.co/datasets/atlasia/Atlaset/>.
- John Dang, Shivalika Singh, Daniel D’souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. [Aya expand: Combining research breakthroughs for a new multilingual frontier](#). Preprint, arXiv:2412.04261.
- Sujatha Das Gollapalli, Mouad Hakam, Mingzhe Du, and See-Kiong Ng. 2026. NUS-IDS at AMIYA/VarDial 2026: Improving arabic dialectness in llms with reinforcement learning. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, Rabat, Morocco. Association for Computational Linguistics.
- Shahd Salah Uddin Dibas, Christian Khairallah, Nizar Habash, Omar Fayez Sadi, Tariq Sairafy, Karmel Sarabta, and Abrar Ardah. 2022. [Maknuune: A large open palestinian Arabic lexicon](#). In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 131–141, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Rana Gaber, Yara Allam, Serag Amin, Ranwa Aly, and Bashar Alhafni. 2026. MBZUAI at AMIYA shared task 2026: Adapting open-source llms for dialectal arabic. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, Rabat, Morocco. Association for Computational Linguistics.
- Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc’Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. [The Flores-101 evaluation benchmark for low-resource and multilingual machine translation](#). *Transactions of the Association for Computational Linguistics*, 10:522–538.
- Khaleel Hamad and Ahmad Al-Najjar. 2026. A closed-track system for palestinian arabic in the AMIYA shared task. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, Rabat, Morocco. Association for Computational Linguistics.
- 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, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.
- Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.
- Shereen Hussein, Mona Farouk, and ElSayed Hemayed. 2019. [Gender identification of egyptian dialect in twitter](#). *Egyptian Informatics Journal*, 20(2):109–116.
- Amr Keleg, Sharon Goldwater, and Walid Magdy. 2023. [ALDi: Quantifying the Arabic level of dialectness of text](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10597–10611, Singapore. Association for Computational Linguistics.
- Shankar Kumar and Bill Byrne. 2004. Minimum bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176.
- Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. 2024. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. In *Forty-second International Conference on Machine Learning*.
- AI Meta. 2024. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. *Meta AI Blog*. Retrieved December, 20:2024.
- Jonathan Mutal, Perla Al Almaoui, Simon Hengchen, and Pierrette Bouillon. 2026. AladdinFTI @ AMIYA - three wishes for arabic nlp: Fidelity,

- diglossia, and multidialectal generation. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, Rabat, Morocco. Association for Computational Linguistics.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. [No language left behind: Scaling human-centered machine translation](#). *Preprint*, arXiv:2207.04672.
- Aissam Outchakoucht and Hamza Es-Samaali. 2024. [The evolution of darija open dataset: Introducing version 2](#). *Preprint*, arXiv:2405.13016.
- Maja Popović. 2015. [chrF: character n-gram F-score for automatic MT evaluation](#). In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.
- Maja Popović. 2017. [chrF++: words helping character n-grams](#). In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.
- Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. [ChatGPT MT: Competitive for high- \(but not low-\) resource languages](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.
- Nathaniel Romney Robinson, Shahd Abdelmoneim, Kelly Marchisio, and Sebastian Ruder. 2025. [AL-QASIDA: Analyzing LLM quality and accuracy systematically in dialectal Arabic](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22048–22065, Vienna, Austria. Association for Computational Linguistics.
- Hashem Sellat, Shadi Saleh, Mateusz Krubiński, Adam Pospíšil, Petr Zemánek, and Pavel Pecina. 2023. [UFAL parallel corpus of north levantine 1.0](#). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).
- Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, and 1 others. 2023. [Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models](#). *arXiv preprint arXiv:2308.16149*.
- Shivalika Singh, Freddie Vargus, Daniel D’souza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O’Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, and 14 others. 2024. [Aya dataset: An open-access collection for multilingual instruction tuning](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.
- Toshiyuki Takezawa, Genichiro Kikui, Masahide Mizushima, and Eiichiro Sumita. 2007. [Multilingual spoken language corpus development for communication research](#). pages 303–324.
- Bashar Talafha, Karima Kadaoui, Samar Mohamed Magdy, Mariem Habiboullah, Chafei Mohamed Chafei, Ahmed Oumar El-Shangiti, Hiba Zayed, Mohamedou Cheikh Tourad, Rahaf Alhamouri, Rwa Assi, Aisha Alraeesi, Hour Mohamed, Fakhraddin Alwajih, Abdelrahman Mohamed, Abdellah El Mekki, El Moatez Billah Nagoudi, Benelhadj Djelloul Mama Saadia, Hamzah A. Alsayadi, Walid Al-Dhabyani, and 8 others. 2024. [Casablanca: Data and models for multidialectal Arabic speech recognition](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21745–21758, Miami, Florida, USA. Association for Computational Linguistics.
- Taghreed Tarmom, William Teahan, Eric Atwell, and Mohammad Ammar Alsalka. 2020. [Compression versus traditional machine learning classifiers to detect code-switching in varieties and dialects: Arabic as a case study](#). *Natural Language Engineering*, 26(6):663–676.
- Qwen Team. 2024. [Qwen2.5: A party of foundation models](#).
- Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D’souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. [Aya model: An instruction finetuned open-access multilingual language model](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.
- 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, and 40 others. 2024. [Qwen2 technical report](#). *arXiv preprint arXiv:2407.10671*.

A Annotator Observations

In this section we discuss some of the annotators’ qualitative observations as they conducted the human evaluation.

Prompt	[..] dataset which contains 2D images, [..] how to find the closest image? Explain in Egyptian Arabic.	Flu.	Adh.
NUS-IDS	موصلة إني، إني إمتلك بيانات بتحتوى على قاعدة صور في ابعاد ٢، بس، إني محتاج أقدر أسوي حاجة هتكون كده: إني عندني صورة جديدة، وكنت مش هتستنى الدنيا أزورق منها، بس هتفكر إن أهم صورة بنقدر نحياها هي الي قدامنا، أو اللي هيوازيها مش هتفرقش بينا و بينها، هتكون كيف كدة؟ ...بالظبط هتستخدم الفنيقة الي هيحسها	2	3
Aladdin	Step 1: Understand the problem statement: The problem is that I am trying to find the best match for an image from a dataset containing a list of images. The goal is to output the index of the best-matching image in the list. Step 2: Identify the input and output formats. Input format: A list of images. Output format: The index of the image in the input list that matches the query image...	1	3
MBZUAI	-	1	1

Table 10: Example prompt for a technical request in Egyptian Arabic, responses shortened for display

A.1 Egyptian

The LLMs evaluated for Egyptian Arabic demonstrate comparable *Adherence* levels, with most responses partially fulfilling the original prompt requests. However, notable differences emerged with respect to Fluency. The NUS-IDS model drew from a mix of dialectal terms spanning multiple varieties, while maintaining only partial sentence coherence. Aladdin’s model generated dialectal phrases characteristic of social media registers, with less alignment to the request and repetition of words within the prompt (i.e. parroting). The MBZUAI model demonstrated stronger comprehension relative to other submissions, though responses contained grammatical errors or fell short of native-like production.

Additionally, we observed instances where NUS-IDS and Aladdin models demonstrated understanding and high adherence of the prompt but failed to maintain the Egyptian dialect, instead responding in a mixed dialect or entirely in English. This is especially evident in prompts requiring code, or technical or highly structured outputs, where responses could be thorough, albeit in English or other dialects.

Table 10 presents a representative example where the prompt asks to explain a technical concept in Egyptian Arabic. NUS-IDS produced mixed dialectal lexical items with inconsistent structure, Aladdin generated dialectal phrasing but failed to address prompt content, and MBZUAI

defaulted entirely to English.

A.2 Moroccan

Across the 100 Moroccan prompts, the four systems can be divided into two distinct performance tiers. NUS-IDS consistently underperformed on both *Fluency* and *Adherence*, producing responses that were often difficult to comprehend and drifted from the target Moroccan variety.

Among the remaining three systems, the performance trade-offs demonstrate more nuance. MBZUAI achieves the highest average Adherence, suggesting it most reliably maintains alignment with Moroccan dialectal features. UM achieves the highest average Fluency, indicating more consistently readable and well-formed responses. Aladdin demonstrates competitive performance on Arabic prompts but exhibits high variance, performing well when prompts are already in Arabic, but declining sharply on English or technical prompts that still require Moroccan Arabic output.

This gap is illustrated in Table 11, where an English prompt explicitly requesting Moroccan Arabic asks how a corporate lawyer might pivot to launching an AI-enabled startup, UM produces a structured, actionable response in the Moroccan dialect demonstrating stability. MBZUAI similarly maintains dialectal output but with slightly more variability in phrasing. In contrast, Aladdin tends to drift on such cross-lingual instructions, often resorting to English or generating unrelated con-

tent, while NUS-IDS outputs incoherent or off-topic text.

A.3 Palestinian

The results indicate that none of the evaluated models consistently adhered well to the prompts, with average Adherence scores remaining low across all submissions. MBZUAI performed best overall, achieving the highest Adherence score and notably stronger Fluency (3.48), suggesting that it has better command of the Palestinian dialect. The remaining teams (NUS-IDS, Aladdin, and BYU) demonstrated similar levels of Adherence, while Fluency scores varied from limited to moderately understandable.

These findings indicate that while some progress has been made in Fluency, particularly by MBZUAI, accurate fulfillment of user requests in Palestinian Arabic remains a significant challenge across models.

Additionally, prompts requiring paraphrasing of specific statements while preserving meaning in Palestinian Arabic revealed notable performance variation across teams. Several models struggled to produce genuine paraphrases without using the original wording, implying limited lexical flexibility. See Table 12.

A.4 Saudi Arabian

Overall, the models demonstrate greater Fluency in the Saudi dialect compared to the other dialects, in addition to marginally higher Adherence scores. Although responses contained highly dialectal vocabulary, there are instances where sentence structure did not form fully coherent answers or failed to completely address the original user request.

It is interesting to note that where prompts incorporated culturally relevant content using local idioms, the models did not produce meaningful responses despite using terminology used frequently in Saudi conversational style.

NUS-IDS produced instances of mixed dialects within single responses alongside highly dialectal but incoherent outputs. Team Aladdin generated phrases in non-Saudi varieties and failed to provide responses with substantial meaning, largely parroting the original prompt without meaningful elaboration. UM produced dialectal phrases typically used in conversation but without coherent sentences.

MBZUAI demonstrated stronger comprehension and Fluency relative to other submissions,

though generated responses contained grammatical errors or fell short of native-like production. It is also notable that this model produced hallucinated responses with a confident tone, particularly relating to idioms and phrases. See Table 13.

A.5 Syrian

With regards to Syrian Arabic, Adherence to prompts remained limited across most submissions. MBZUAI clearly outperformed other systems and demonstrated stronger Syrian dialect generation, achieving the highest Adherence (1.62) and Fluency (3.43).

For the NUS-IDS, Aladdin, UM, and SDNLP teams, the results yielded comparably low Adherence scores, reflecting partial or irrelevant responses to prompts. With respect to Fluency, SDNLP and UM produced moderately more natural responses than NUS-IDS and Aladdin. However, responses did not sound completely native, reinforcing the idea of gaps in dialectal naturalness and accuracy.

Models exhibited unexpected difficulties with what should have been a relatively straightforward vocabulary word: *صفحة* (sfha), a well-known Syrian and Levantine dish. Only UM partially identified the term within a food context, while SDNLP largely reproduced the prompt without elaboration. NUS-IDS and Aladdin misinterpreted the term entirely, conflating *صفحة* with *صفحة تكتونية* (tectonic plate) and producing responses that were off-tangent. See Table 14.

B Genre-specific Automatic Scores

In this section we display the ADI2 and chrF++ scores across all different evaluation subsets mentioned in §2.4.

See tables 15, 16, 17, 18, 19, 20, 21, and 18, 22 for ADI2 scores across eval subsets; and see tables 23, 24, 25, 26, 27, 28, 29, and 30 for chrF++ scores across eval subsets.

Prompt	وشو شكل علم السعودية؟	Flu.	Adh.
NUS-IDS	علم المملكة العربية السعودية هو علم يتبنى بتركيبة من الصفراء والمينا. وشملت التباينات الفنية للوحة المينا زخارف هندسية متوسطة الدقة وعلى شكل حلقة. وتتوزع الزخارف في جميع الأقسام والزوايا في لوحة المينا، مع احتفاظ لوحة الصفراء بدواخلها السمكية والضخمة، وتباين الألوان في اللوحة بأشكال هندسية دقيقة.	4	2
Aladdin	هذا هو علم المملكة العربية السعودية	3	4
UM	ايش لونها علمكم ايش لونها علمكم الله يستر عليك من الاشكال هذي الي تبطح الناس تحت اقدامها ياليتي انا الي ابوه يوه ياليت	3	2
MBZUAI	علم المملكة العربية السعودية يتكون من اربعة ارباع متساوية، الجزء العلوي باللون الازرق والجزء السفلي باللون الابيض، وفي الوسط بلون الاخضر مع ثلاث نجوم بيضاء في اليسار.	4	1

Table 13: Example prompt asking to describe the Saudi flag in Saudi Arabic; models generate made up answers with moderate fluency. Note: responses have been shortened for formatting purposes.

Prompt	شو هي الصفيحة السورية؟	Flu.	Adh.
NUS-IDS	وكم توزع عليها الجبال والمرتفعات؟ المساحة الكلية للصفيحة السورية: ١٢ مليون كم. المساحة الكلية للصفيحة السورية تشكل حوالي ٣٥% من المساحة الكلية للنصف الكرة الغربي ويضم الكتلة الجبلية السورية الممتدة عبر الجزء الجنوبي الشرقي من أوراسيا... ويحدها من الشمال قارتا أوروبا وأقصى شمال إفريقيا،	3	1
Aladdin	الصفيحة السورية هي منطقة من الأرض تحت المحيطات.	3	1
UM	انا مايعرف الا صرره ازمه قلبه من ريحه طبخك اديش بتاكل من هالصفيحه امجد انا ماني جوع	1	3
MBZUAI	؟	1	1
SDNLP	أنت شو عم تقولي صفيحة سوريا؟	3	1

Table 14: Example prompt asking about a well-known Syrian dish; models generate mixed responses with little adherence. Note: responses have been shortened for formatting purposes.

	submission	mar	egy	pse	syr	sau
Aladdin-FTI	primary	35.4%	13.3%	8.9%	11.8%	3.2%
Aladdin-FTI	contrastive1	33.8%	12.8%	9.8%	14.1%	4.5%
BYU	primary	–	–	7.0%	–	–
MBZUAI	primary	53.3%	42.0%	13.7%	9.8%	10.8%
MBZUAI	contrastive1	54.8%	33.2%	16.2%	12.2%	8.3%
MBZUAI	contrastive2	56.8%	51.5%	21.5%	27.5%	7.2%
NUS-IDS	primary	73.9%	66.6%	16.3%	8.0%	3.9%
NUS-IDS	contrastive1	43.4%	19.3%	12.8%	26.8%	2.9%
NUS-IDS	contrastive2	25.4%	0.6%	3.6%	0.7%	0.9%
SDNLP	primary	–	–	–	30.4%	–
UM	primary	87.8%	–	–	53.7%	53.9%
Baseline		5.7%	5.4%	4.3%	1.4%	0.4%

Table 15: ADI2 on Palm prompts

	submission	mar	egy	pse	syr	sau
Aladdin-FTI	primary	39.2%	8.2%	4.4%	8.4%	0.3%
Aladdin-FTI	contrastive1	39.3%	7.7%	4.8%	11.8%	0.6%
BYU	primary	–	–	6.8%	–	–
MBZUAI	primary	69.6%	59.2%	12.6%	13.0%	2.5%
MBZUAI	contrastive1	56.8%	47.8%	9.2%	13.5%	3.8%
MBZUAI	contrastive2	71.6%	58.6%	11.6%	11.4%	9.2%
NUS-IDS	primary	87.1%	66.8%	6.9%	2.1%	1.7%
NUS-IDS	contrastive1	73.3%	17.1%	6.4%	32.2%	0.3%
NUS-IDS	contrastive2	39.7%	0.2%	0.9%	0.0%	0.1%
SDNLP	primary	–	–	–	29.7%	–
UM	primary	77.5%	–	–	45.5%	42.8%
Baseline		13.1%	12.3%	0.7%	1.2%	0.0%

Table 16: ADI2 on FLORES monolingual prompts

	submission	mar	egy	pse	syr	sau
Aladdin-FTI	primary	44.1%	17.0%	11.3%	15.3%	4.9%
Aladdin-FTI	contrastive1	49.7%	12.4%	11.9%	24.1%	2.2%
BYU	primary	–	–	11.1%	–	–
MBZUAI	primary	72.8%	68.2%	19.3%	24.7%	16.0%
MBZUAI	contrastive1	74.7%	61.2%	16.6%	22.2%	14.3%
MBZUAI	contrastive2	77.3%	66.6%	20.1%	20.1%	21.1%
NUS-IDS	primary	85.2%	75.2%	9.3%	4.7%	4.8%
NUS-IDS	contrastive1	58.6%	25.3%	8.0%	38.6%	0.5%
NUS-IDS	contrastive2	52.6%	1.4%	2.8%	0.6%	2.7%
SDNLP	primary	–	–	–	36.3%	–
UM	primary	82.3%	–	–	57.5%	47.9%
Baseline		20.3%	27.1%	3.1%	4.8%	1.3%

Table 17: ADI2 on MADAR monolingual prompts

	submission	mar	egy	pse	syr	sau
Aladdin-FTI	primary	38.8%	10.3%	4.6%	29.2%	1.9%
Aladdin-FTI	contrastive1	36.9%	11.3%	3.7%	32.0%	1.7%
BYU	primary	–	–	1.4%	–	–
MBZUAI	primary	64.2%	33.9%	4.4%	14.3%	4.9%
MBZUAI	contrastive1	52.3%	25.0%	0.9%	16.5%	2.8%
MBZUAI	contrastive2	53.6%	30.1%	3.1%	20.3%	8.0%
NUS-IDS	primary	71.8%	59.4%	6.6%	6.9%	3.9%
NUS-IDS	contrastive1	78.2%	83.1%	15.0%	36.9%	15.0%
NUS-IDS	contrastive2	24.5%	0.4%	0.1%	0.2%	0.2%
SDNLP	primary	–	–	–	25.5%	–
UM	primary	47.5%	–	–	31.1%	37.7%
Baseline		7.9%	1.6%	0.1%	0.1%	0.1%

Table 18: ADI2 on FLORES ENG→DA MT outputs

	submission	mar	egy	pse	syr	sau
Aladdin-FTI	primary	21.1%	8.0%	3.3%	20.1%	1.2%
Aladdin-FTI	contrastive1	17.9%	10.2%	3.1%	22.8%	0.6%
BYU	primary	–	–	1.0%	–	–
MBZUAI	primary	38.2%	25.2%	0.3%	14.7%	4.3%
MBZUAI	contrastive1	34.2%	21.6%	0.6%	14.5%	3.0%
MBZUAI	contrastive2	38.5%	21.1%	0.1%	15.4%	6.2%
NUS-IDS	primary	86.1%	75.0%	10.6%	12.1%	8.2%
NUS-IDS	contrastive1	83.0%	79.3%	13.3%	34.9%	20.5%
NUS-IDS	contrastive2	20.4%	0.2%	0.1%	0.1%	0.2%
SDNLP	primary	–	–	–	10.5%	–
UM	primary	31.5%	–	–	29.1%	37.6%
Baseline		18.3%	16.6%	1.2%	0.5%	1.0%

Table 19: ADI2 on FLORES MSA→DA MT outputs

	submission	mar	egy	pse	syr	sau
Aladdin-FTI	primary	66.7%	33.3%	12.0%	41.9%	16.7%
Aladdin-FTI	contrastive1	65.7%	30.6%	9.5%	46.3%	12.2%
BYU	primary	–	–	5.6%	–	–
MBZUAI	primary	63.6%	48.0%	10.6%	22.9%	17.7%
MBZUAI	contrastive1	60.0%	47.2%	6.8%	21.4%	15.6%
MBZUAI	contrastive2	69.3%	49.2%	6.7%	25.1%	24.6%
NUS-IDS	primary	58.9%	51.9%	3.5%	9.3%	7.5%
NUS-IDS	contrastive1	89.8%	77.4%	17.3%	38.8%	29.1%
NUS-IDS	contrastive2	46.8%	1.3%	0.2%	1.2%	0.8%
SDNLP	primary	–	–	–	37.8%	–
UM	primary	66.0%	–	–	38.0%	55.7%
Baseline		10.2%	7.9%	1.0%	1.1%	0.6%

Table 20: ADI2 on MADAR ENG→DA MT outputs

	submission	mar	egy	pse	syr	sau
Aladdin-FTI	primary	61.0%	33.0%	12.2%	39.0%	10.5%
Aladdin-FTI	contrastive1	65.4%	37.8%	10.5%	39.9%	6.6%
BYU	primary	–	–	3.1%	–	–
MBZUAI	primary	57.2%	45.1%	4.2%	20.8%	16.9%
MBZUAI	contrastive1	51.2%	46.3%	5.6%	19.5%	16.2%
MBZUAI	contrastive2	50.6%	48.0%	1.2%	22.2%	21.6%
NUS-IDS	primary	70.0%	51.1%	32.0%	29.2%	15.3%
NUS-IDS	contrastive1	88.6%	82.7%	17.2%	27.5%	29.5%
NUS-IDS	contrastive2	38.8%	0.6%	0.3%	0.8%	0.2%
SDNLP	primary	–	–	–	22.3%	–
UM	primary	63.3%	–	–	36.0%	53.4%
Baseline		13.7%	19.6%	2.2%	1.0%	2.6%

Table 21: ADI2 on MADAR MSA→DA MT outputs

	submission	mar	egy	pse	syr	sau
Aladdin-FTI	primary	4.4%	0.5%	1.3%	6.7%	0.7%
Aladdin-FTI	contrastive1	3.0%	0.5%	1.1%	5.6%	0.6%
BYU	primary	–	–	5.0%	–	–
MBZUAI	primary	45.2%	35.8%	6.8%	21.3%	8.3%
MBZUAI	contrastive1	46.7%	33.1%	4.5%	18.2%	4.7%
MBZUAI	contrastive2	44.9%	34.7%	5.1%	28.4%	13.4%
NUS-IDS	primary	23.8%	56.3%	8.5%	9.2%	11.3%
NUS-IDS	contrastive1	87.6%	77.0%	16.5%	51.4%	15.6%
NUS-IDS	contrastive2	23.5%	0.4%	0.3%	0.8%	0.5%
SDNLP	primary	–	–	–	31.0%	–
UM	primary	44.8%	–	–	23.2%	44.5%
Baseline		2.9%	1.4%	0.5%	0.8%	0.8%

Table 22: ADI2 on Arena-Hard prompts

	submission	mar	egy	pse	syr	sau
Aladdin-FTI	primary	31.7	35.1	37.3	24.0	35.7
Aladdin-FTI	contrastive1	31.4	34.4	36.3	21.4	35.6
BYU	primary	-	-	17.7	-	-
MBZUAI	primary	20.7	31.9	36.4	30.3	31.8
MBZUAI	contrastive1	25.7	33.5	33.7	30.2	33.8
MBZUAI	contrastive2	25.0	29.1	29.0	24.1	29.4
NUS-IDS	primary	11.5	15.7	16.1	15.4	16.3
NUS-IDS	contrastive1	23.7	18.1	19.2	17.7	19.2
NUS-IDS	contrastive2	28.3	25.9	31.5	21.8	35.8
SDNLP	primary	-	-	-	13.7	-
UM	primary	29.1	-	-	36.4	21.1
Baseline		25.8	32.9	35.1	34.4	37.1

Table 23: chrF++ on FLORES ENG→DA

	submission	mar	egy	pse	syr	sau
Aladdin-FTI	primary	25.1	29.2	30.6	25.1	46.0
Aladdin-FTI	contrastive1	23.9	28.3	30.0	21.1	32.4
BYU	primary	-	-	22.7	-	-
MBZUAI	primary	35.3	39.9	43.0	38.6	58.5
MBZUAI	contrastive1	35.5	45.3	42.0	37.6	60.3
MBZUAI	contrastive2	32.9	36.5	46.0	37.8	43.1
NUS-IDS	primary	14.4	23.2	22.2	22.1	25.9
NUS-IDS	contrastive1	22.5	19.3	20.0	21.3	22.9
NUS-IDS	contrastive2	33.9	14.4	35.4	16.3	67.3
SDNLP	primary	-	-	-	16.6	-
UM	primary	35.0	-	-	43.6	27.1
Baseline		33.4	46.9	44.0	43.7	69.9

Table 24: chrF++ on FLORES MSA→DA

	submission	mar	egy	pse	syr	sau
Aladdin-FTI	primary	43.1	49.4	52.2	40.6	52.4
Aladdin-FTI	contrastive1	42.5	49.2	53.0	40.8	53.0
BYU	primary	-	-	15.0	-	-
MBZUAI	primary	40.7	45.1	48.4	53.6	56.9
MBZUAI	contrastive1	48.3	51.6	57.3	52.1	55.7
MBZUAI	contrastive2	41.0	41.0	47.2	46.8	49.7
NUS-IDS	primary	25.1	26.1	27.9	25.9	26.5
NUS-IDS	contrastive1	45.8	26.4	28.5	28.6	27.9
NUS-IDS	contrastive2	34.8	6.3	48.8	9.8	55.5
SDNLP	primary	-	-	-	34.7	-
UM	primary	46.7	-	-	52.4	0.0
Baseline		47.0	52.7	57.2	53.3	58.8

Table 25: chrF++ on FLORES DA→ENG

	submission	mar	egy	pse	syr	sau
Aladdin-FTI	primary	25.5	28.9	30.9	25.8	44.1
Aladdin-FTI	contrastive1	25.2	27.4	30.7	22.0	33.9
BYU	primary	-	-	22.6	-	-
MBZUAI	primary	39.1	45.0	44.1	41.8	72.2
MBZUAI	contrastive1	44.1	53.1	45.0	45.4	73.0
MBZUAI	contrastive2	36.9	39.3	43.7	37.4	55.9
NUS-IDS	primary	18.7	26.4	23.8	22.5	32.7
NUS-IDS	contrastive1	35.1	25.4	23.9	26.8	56.7
NUS-IDS	contrastive2	38.6	16.2	41.0	17.9	67.2
SDNLP	primary	-	-	-	7.5	-
UM	primary	42.6	-	-	49.7	43.8
Baseline		42.2	52.6	49.0	46.1	79.1

Table 26: chrF++ on FLORES DA→MSA

	submission	mar	egy	pse	syr	sau
Aladdin-FTI	primary	34.5	31.0	38.3	17.7	38.2
Aladdin-FTI	contrastive1	34.0	28.7	35.9	15.7	38.4
BYU	primary	-	-	6.6	-	-
MBZUAI	primary	29.2	39.5	38.4	33.2	42.9
MBZUAI	contrastive1	30.7	37.0	35.0	33.3	45.4
MBZUAI	contrastive2	28.5	38.5	36.0	31.3	44.5
NUS-IDS	primary	4.6	5.5	5.8	6.0	7.4
NUS-IDS	contrastive1	5.6	11.5	9.6	8.3	10.4
NUS-IDS	contrastive2	15.2	9.7	8.2	15.4	23.3
SDNLP	primary	-	-	-	18.2	-
UM	primary	32.6	-	-	28.9	16.7
Baseline		17.6	20.9	22.5	21.5	29.7

Table 27: chrF++ on MADAR ENG→DA

	submission	mar	egy	pse	syr	sau
Aladdin-FTI	primary	31.4	26.9	33.3	16.1	32.2
Aladdin-FTI	contrastive1	28.9	24.5	32.9	13.5	30.7
BYU	primary	-	-	7.6	-	-
MBZUAI	primary	33.6	38.1	31.0	33.5	42.8
MBZUAI	contrastive1	34.6	37.9	32.9	33.5	43.7
MBZUAI	contrastive2	33.9	35.8	29.0	35.5	41.6
NUS-IDS	primary	7.9	9.1	7.8	8.8	10.3
NUS-IDS	contrastive1	5.3	11.2	8.8	8.0	9.1
NUS-IDS	contrastive2	26.7	7.6	16.7	4.4	29.8
SDNLP	primary	-	-	-	13.8	-
UM	primary	30.0	-	-	30.8	17.2
Baseline		22.1	25.4	27.2	25.6	29.6

Table 28: chrF++ on MADAR MSA→DA

	submission	mar	egy	pse	syr	sau
Aladdin-FTI	primary	56.3	51.4	55.5	36.9	58.9
Aladdin-FTI	contrastive1	55.3	50.7	55.1	36.4	59.0
BYU	primary	-	-	4.3	-	-
MBZUAI	primary	58.1	55.9	56.6	55.7	63.0
MBZUAI	contrastive1	59.6	59.1	60.0	54.4	64.8
MBZUAI	contrastive2	56.2	50.8	59.1	52.8	60.1
NUS-IDS	primary	11.1	9.7	10.2	9.6	11.1
NUS-IDS	contrastive1	7.8	14.0	13.9	12.1	15.9
NUS-IDS	contrastive2	31.1	6.0	41.6	8.4	49.6
SDNLP	primary	-	-	-	40.6	-
UM	primary	59.4	-	-	50.2	0.0
Baseline		42.0	39.8	37.2	37.1	48.5

Table 29: chrF++ on MADAR DA→ENG

	submission	mar	egy	pse	syr	sau
Aladdin-FTI	primary	34.0	30.9	31.9	20.4	34.7
Aladdin-FTI	contrastive1	33.6	30.3	30.7	18.7	35.0
BYU	primary	-	-	7.9	-	-
MBZUAI	primary	39.7	40.1	33.3	35.1	42.7
MBZUAI	contrastive1	43.7	42.5	35.8	40.8	44.2
MBZUAI	contrastive2	40.8	38.1	35.3	25.5	44.1
NUS-IDS	primary	5.8	6.3	5.7	5.4	6.7
NUS-IDS	contrastive1	6.4	11.1	10.5	11.4	11.4
NUS-IDS	contrastive2	23.9	4.8	14.3	5.2	29.5
SDNLP	primary	-	-	-	4.4	-
UM	primary	31.7	-	-	26.2	22.6
Baseline		30.2	30.6	29.3	29.2	34.0

Table 30: chrF++ on MADAR DA→MSA

Far Out: Evaluating Language Models on Slang in Australian and Indian English

Deniz Kaya Dilsiz Dipankar Srirag Aditya Joshi
University of New South Wales, Sydney, Australia
{d.dilsiz, d.srirag, aditya.joshi}@unsw.edu.au

Abstract

Language models exhibit systematic performance gaps when processing text in non-standard language varieties, yet their ability to comprehend variety-specific slang remains underexplored for several languages. We present a comprehensive evaluation of slang awareness in Indian English (en-IN) and Australian English (en-AU) across seven state-of-the-art language models. We construct two complementary datasets: WEB, containing 377 web-sourced usage examples from Urban Dictionary, and GEN, featuring 1,492 synthetically generated usages of these slang terms, across diverse scenarios. We assess language models on three tasks: target word prediction (TWP), guided target word prediction (TWP*) and target word selection (TWS). Our results reveal four key findings: (1) Higher average model performance TWS versus TWP and TWP*, with average accuracy score increasing from 0.03 to 0.49 respectively (2) Stronger average model performance on WEB versus GEN datasets, with average similarity score increasing by 0.03 and 0.05 across TWP and TWP* tasks respectively (3) en-IN tasks outperform en-AU when averaged across all models and datasets, with TWS demonstrating the largest disparity, increasing average accuracy from 0.44 to 0.54. These findings underscore fundamental asymmetries between generative and discriminative competencies for variety-specific language, particularly in the context of slang expressions despite being in a technologically rich language such as English.

1 Introduction

Varieties of a language are considered to differ in syntax, vocabulary and pragmatics (Joshi et al., 2025). A particularly unique aspect in terms of vocabulary is colloquial terms or slang. Slang is defined as language used by a particular group of people, during the period of its popularity (Dumas and Lighter, 1978; Reves, 1926). In this paper,

we investigate the ability of large language models (LLMs) to interpret slang in the context of Australian and Indian English, examples of which are as follows:

Prepone (Verb): A colloquial Indian English term used to bring a scheduled event to an earlier time or date; the opposite of postpone. An example usage is ‘*Since the manager is leaving early, we have decided to prepone the meeting to 10 AM.*’

Far Out (Adj): A colloquial Australian English expression used to denote surprise, disbelief, or to describe something as being excellent or extreme. For example, ‘*Far out, I can’t believe how much that concert ticket cost!*’

Slang is inherently dynamic, community-specific, and culturally embedded, making it a critical starting point for evaluating whether contemporary language models can handle the full spectrum of linguistic diversity. Recent research has highlighted systematic performance gaps when large language models (LLMs) process non-standard English varieties (Wuraola et al. 2024, Khanuja et al. 2020, Deas et al. 2023), yet most work on slang interpretation has focused exclusively on Standard American English or has not distinguished between English varieties (Mei et al. 2024, Sun et al. 2024). This limitation overlooks the reality that slang terms are often variety-specific, carrying meanings and connotations that differ across geographical and cultural contexts. Being a global language, English has developed several colloquialisms in the geographies where it is spoken, highlighting the varied usage across cultures.

In this paper, we systematically evaluate how well state-of-the-art LLMs understand slang phrases from Indian English (en-IN) and Australian English (en-AU). The two are representative of two kinds of Englishes: en-IN may be spoken as an

additional language while en-AU may be spoken as a first language. With these English varieties in focus, we address the research question:

“How well do language models understand lexical variations in language varieties in terms of colloquial expressions or slang?”

In order to do so, we set up two downstream tasks, inspired by [Srirag et al. \(2025b\)](#): (a) **target word¹ prediction**: Predicting a word in the place of a blanked position in an example sentence from among all words in the vocabulary, and (b) **target word selection**: Predicting a word in the place of a blanked position in an example sentence from among a set of options. We use example sentences from two sources: (a) slang definitions and example usages from Urban Dictionary² that are manually curated by two native speakers; and (b) LLM-generated examples based on definitions from part (a). The sources result in two evaluative datasets: WEB and GEN. We then evaluate several LLMs on target word prediction and target word selection. **Our evaluation highlights the limitations of LLMs to identify language variety-specific colloquial words in context, and bear implications to culturally aware NLP systems.**

This paper makes the following contributions:

1. In the methodological sense, this is the first evaluative study that utilises target word selection and target word prediction using encoder and decoder models to evaluate if LLMs understand language variety-specific slangs, particularly for varieties of English.
2. Novel Evaluation Datasets: Our datasets, WEB, containing web-sourced usage examples from Urban Dictionary, and GEN, featuring synthetically generated diverse scenarios comprise 377 unique slang phrases across en-IN and en-AU varieties.
3. Systematic Performance Analysis: We evaluate 7 language models ranging from 110M to 8B parameters, revealing key findings related to model performance in target word prediction and target word selection settings.

Our findings demonstrate that while modern LLMs show promising capabilities compared to MLMs particularly for English, significant challenges remain in understanding variety-specific

slang, such as Australian English. These results underscore the need for more inclusive training data and evaluation benchmarks that account for linguistic diversity beyond standard language varieties.

2 Methodology

We present a detailed overview of our evaluation methodology in Figure 1. Using web-based sources, we collect candidate slang phrases from the two varieties alongside their definitions and usages (WEB). We then manually validate the correctness using native speakers. We then use Google Gemini Pro 2.5 ([Comanici et al., 2025](#)) to create scenarios and examples where such slang words will be used (GEN). With the two datasets in place, we mask the slang phrase and get encoder and decoder models to predict the masked phrase.

2.1 Dataset Creation

As described in Figure 2, we construct the dataset with English slang phrases used in two regions, hence covering two language varieties³: Australian English (en-AU) and Indian English (en-IN). We collect the slang phrases from a web-based source and validate the relevance using native speakers. We also augment the dataset to include diverse scenarios with the usage of slang phrases.

Source Collection We collate an initial list of the slang phrases, from both regions, using Wiktionary⁴. For each phrase, we extract corresponding definitions and usage examples from Urban Dictionary, a peer-contributed platform where users submit multiple definitions for individual phrases. Source Collection yields an initial list of 940 slang phrases for en-IN and 2540 slang phrases for en-AU.

Expert Validation We then employ one expert annotator from each region of interest to manually review and filter and remove irrelevant or incorrect entries. This process yields WEB, a high-quality subset of phrase, definition and usage example tuples from Urban Dictionary. Expert Validation removes 876 slang phrases from en-IN and 2227 slang phrases from en-AU.

Scenario Generation Using the phrases and their corresponding definitions from WEB, we prompt Google Gemini Pro 2.5 ([Comanici et al., 2025](#))

¹Some slangs are phrases and are treated as such.

²www.urbandictionary.com; Accessed on 31 December 2025.

³We do not consider the subdialects present in both regions, and acknowledge the same as a reasonable limitation.

⁴en.wiktionary.org; Accessed on 31 December 2025.

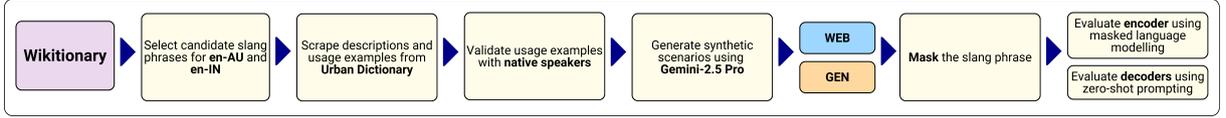


Figure 1: Methodological Overview of Our Approach.

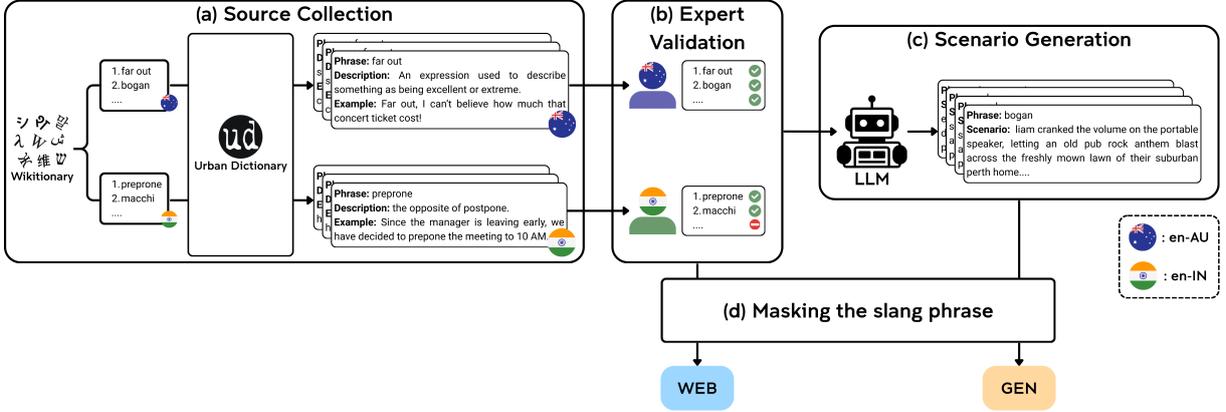


Figure 2: Dataset Creation.

Variety	Count		\bar{p}	\bar{d}	\bar{u}		Perp.		Latest Entry
	WEB	GEN			WEB	GEN	WEB	GEN	
en-AU	313	1244	7.7	138	90.4	443.0	701.5	48.9	Sept 10 2025
en-IN	64	248	8.0	197.7	114.1	468.2	406.5	47.1	Oct 15 2025

Table 1: Constructional statistics of WEB and GEN. \bar{p} , \bar{d} and \bar{u} is the average character length of a slang phrase, corresponding definition and usage/scenario respectively. *Perp.* represents average perplexity computed using GPT-2 (Radford et al., 2019).

Variety	R_1	R_2	R_L	<i>Sim.</i>
en-AU	0.25	0.03	0.16	0.39
en-IN	0.24	0.03	0.15	0.37

Table 2: Diversity analysis of generated scenarios in GEN. $R_{(1,2,L)}$ are average ROUGE scores and *Sim.* is the average cosine similarity between scenario embeddings, extracted using allMiniLM-L6-v2 (Reimers and Gurevych, 2019).

to generate four unique scenarios that naturally motivate the usage of each phrase. The generation prompt (provided in Appendix A) instructs the model to create scenarios with named characters, specific settings, and single quotations containing the target phrase.

2.2 Analysis

Table 1 provides a constructional statistics of both WEB and GEN. We notice a higher number of slang phrases extracted for en-AU compared to en-IN.

We also compute perplexity of the usage examples and generated scenarios, using GPT-2 (Radford et al., 2019) from WEB and GEN respectively. We observe a higher perplexity with the usage examples in WEB (en-AU: 701.5) as compared to the generated scenarios in GEN (en-IN: 47.1). This is due to the usage examples in WEB, being derived from peer-contributed web content. Furthermore, both subsets differ in terms of character lengths, with WEB having shorter examples (en-IN: 114.1) compared to GEN (en-AU: 443.0). Taken together, this disparity suggests that while WEB captures authentic naturalistic usage, it possesses lower overall linguistic fluency compared to the synthetically curated GEN. **This observation is pertinent to future research in creation or collection of datasets for language varieties, particularly for high-resource languages such as English.** We present example slang phrases and their corresponding usage examples/scenarios in Appendix B.

Scenario Diversity Analysis We also evaluate the diversity of the generated scenarios in GEN in terms of lexical overlap and semantic similarity. This evaluation is necessary because evaluating language models on similar scenarios, and usage examples does not yield any meaningful insights. For each slang phrase in GEN, we compute a pairwise n-gram overlap between all the generated scenarios using ROUGE (Lin, 2004) to evaluate lexical diversity. Similarly, we compute pairwise cosine similarity between the scenario embeddings, extracted using allMiniLM-L6-v2 (Reimers and Gurevych, 2019). Table 2 reports low average semantic similarity, evidencing that the scenarios remain semantically dissimilar to each other, while also diverging lexically.

2.3 Evaluative Tasks

Given a usage example with the slang phrase, we mask the slang phrase. For inputs with multiple instances of the slang phrase, we create multiple instances with each position being masked. We use pre-trained language models to (a) predict the masked slang phrase; and (b) select the slang phrase from a set of options with other distractors. The two tasks are described as follows.

Target Word Prediction (TWP) In this task, the model is given a sentence in which a slang phrase has been masked and is asked to predict the missing phrase. The model is free to generate any word or phrase from its vocabulary, and the output is considered correct if it exactly matches the original slang phrase. With encoder-only language models, we predict the phrase at the masked position, utilising masked language modeling. For decoder-only large language models, we convert the masked sentence into a cloze-style prompt. An indicative prompt used for this task is provided in Appendix C.

For decoder-only models, we additionally conduct a guided variant of the task (labeled TWP*), where the prompt is augmented with an explicit instruction directing the model to generate a slang phrase suitable for the specific language variety. An indicative prompt used for this task is provided in Appendix D.

Target Word Selection (TWS) We additionally evaluate models using a multiple-choice version of the task. The model is given a sentence with the slang phrase masked, along with a fixed set of candidate answers. The candidate set consists of

four options: the correct slang phrase and three randomly selected distractor phrases drawn from the same language variety. The model is instructed to select the option that best fits the masked context.

For decoder-only large language models, we add an explicit instruction: “*Fill in the blank with the best-fitting answer.*”. An indicative prompt used for this task is provided in Appendix E

3 Experiment Setup

We report the performance on *seven* pre-trained language models including *three* encoder models: BERT-Base (BERT; Devlin et al. 2019), RoBERTa-Large (ROBERTA; Liu et al. 2019), XLM-RoBERTa-Large (XLM; Conneau et al. 2020) and *four* decoder models: Granite-4.0-1B (GRANITE; Granite Team 2025), Llama-3.2-3B-Instruct (LLAMA; Grattafiori et al. 2024), Olmo-2-7B-Instruct (OLMO; OLMo et al. 2025), Qwen3-4b-Instruct (QWEN; Yang et al. 2025). We evaluate encoder models under masked language modelling, where we extract the top-1 predictions from the masked position. For decoder models, we formulate the evaluation as a multiple-choice cloze format task with temperature=0.8. All experiments were conducted on an Apple M1 Pro with 16GB RAM using 8-bit quantized versions of the models to optimize inference efficiency on local hardware.

We report our results on two metrics: *accuracy* and *similarity*. *Accuracy* is the proportion of instances where the model predicted the ground truth slang phrase. As *similarity*, we report the cosine similarity between the Sentence-BERT embeddings (Reimers and Gurevych, 2019) of the reference slang phrase and the predicted slang phrase. To verify robustness of Sentence-BERT embeddings, we additionally compute cosine similarity between the Granite-embedding-125m-english embeddings (Awasthy et al., 2025) and the two embedding models’ agreement by computing Pearson Correlation. The average Pearson Correlation across all models, domains and varieties is 0.77. The complete results are tabulated in Appendix G and H.

4 Results

We present the results of experiments, centered around the following questions: (a) How well do language models perform on the downstream tasks? (Section 4.1); (b) How do factors such as domain and language variety influence model performance?

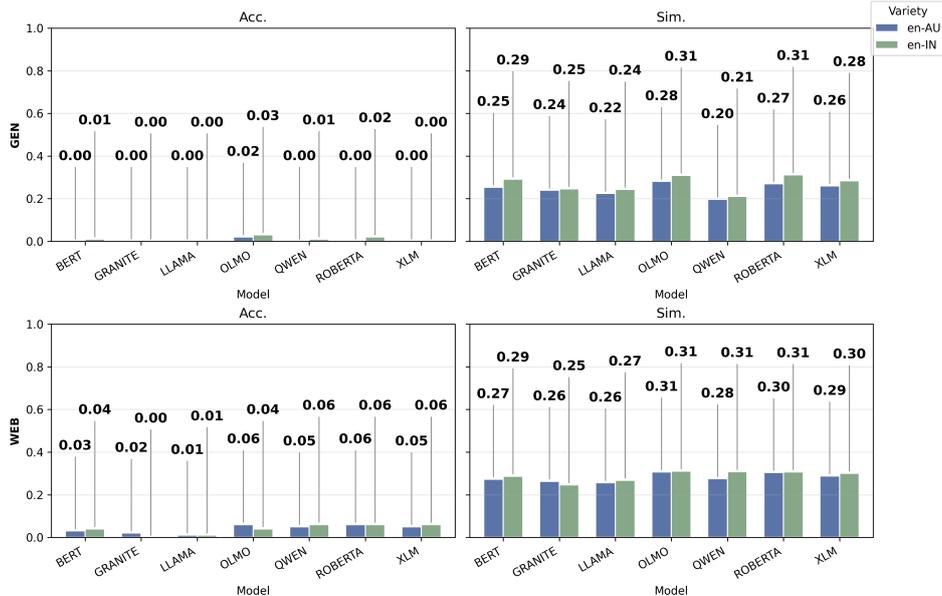


Figure 3: Performance comparison of various models on the target word prediction (TWP) task across en-AU and en-IN.

(Section 4.2); (c) What types of errors do models make, and what do these errors reveal about model behaviour? (Section 4.3).

4.1 Task Performance

We first present the overall model performance on the two downstream tasks: target word prediction (TWP) and its guided variant (TWP*) as well as target word selection (TWS). As shown in Table 3, models perform poorly on TWP, achieving an average accuracy of 0.02 and an average similarity of 0.27. TWP* performs similarly poorly with an average accuracy of 0.04 and an average similarity of 0.28. This indicates that generating the correct slang phrase in an open-ended setting remains challenging. Figure 3 and Figure 4 further show that this difficulty is consistent across model architectures and language varieties, with only marginal differences between models and a maximum accuracy of 0.13. In contrast, models report higher performance on TWS. When the task is formulated as a multiple-choice cloze test, average accuracy increases to 0.49, with a corresponding similarity score of 0.61. This suggests that while current models struggle to generate slang expressions, they are more effective at recognising the correct phrase when the search space is constrained. Figure 5 shows that the improvement is consistent across model architectures and language varieties.

4.2 Impact of Domain and Varieties

We next examine how domain and language variety influence model performance. Tables 4 and 5 summarise these effects by averaging results across all models.

Domain Effects. Table 4 shows that models consistently perform better on data drawn from the WEB domain than on the GEN domain, potentially due to data contamination. For TWP, web-based examples yield an accuracy of 0.04, compared to 0.01 on generated data. A similar drop is observed in semantic similarity. TWP* yields a larger drop, with accuracy decreasing from 0.07 to 0.02 across the sets. This gap is also present for TWS, where accuracy decreases from 0.50 on web data to 0.49 on generated data. These results indicate that domain shift has a measurable impact on the open-vocabulary tasks. Generated examples from GEN appear to be more challenging, likely because they differ in style or contextual cues from naturally occurring web data. This observation resonates with past observations regarding data contamination and evaluation (Dong et al., 2024). Therefore, we recommend that **generating new data (either using a human or an LLM) is a more robust evaluation technique for language varieties.**

Language Variety Effects. Table 5 reports performance differences between en-AU and en-IN.

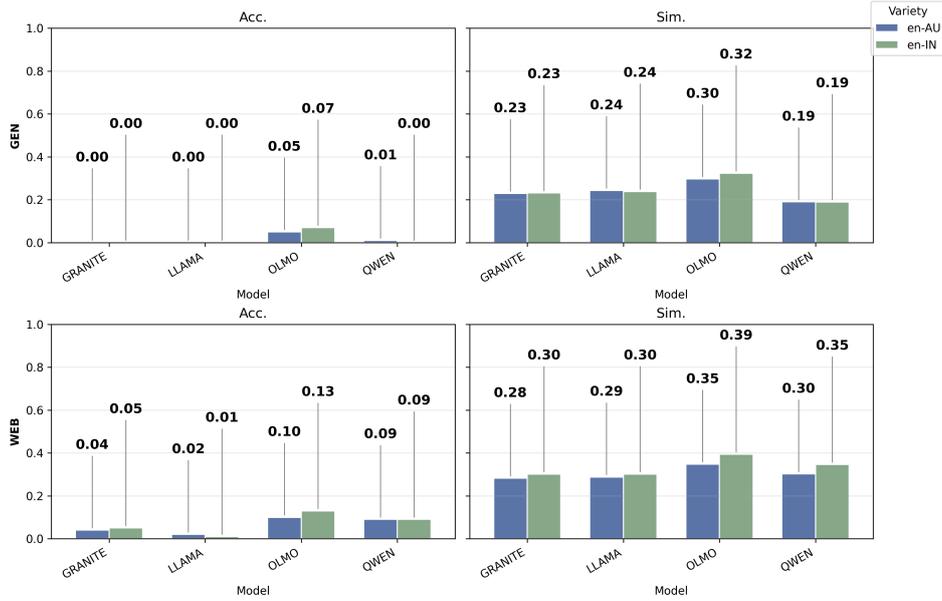


Figure 4: Performance comparison of various models on the guided target word prediction (TWP*) task across en-AU and en-IN.

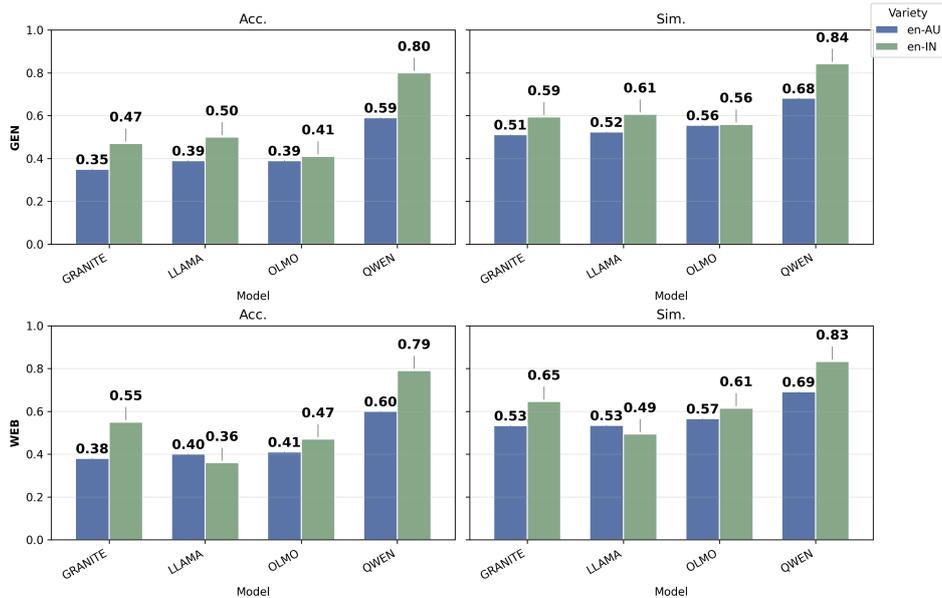


Figure 5: Performance comparison of various models on the target word selection (TWS) task across en-AU and en-IN.

Across all tasks, models perform better on en-IN than on en-AU. For TWP, en-IN shows higher accuracy (0.03 vs. 0.02) and higher similarity (0.28 vs. 0.26). Models perform similarly across both metrics for TWP*. The difference is even more pronounced for TWS, where accuracy improves by 0.10 and similarity by 0.08 when moving from en-AU to en-IN. This consistent improvement suggests

that models are better aligned with slang usage patterns found in en-IN data, possibly due to greater representation or stylistic overlap in pre-training data. Importantly, the variety effect is larger than the domain effect for TWS, indicating that language variety plays a particularly significant role when models must discriminate between competing slang candidates.

Task	Acc.	Sim.
TWP	0.02	0.27
TWP*	0.04	0.28
TWS	0.49	0.61

Table 3: Average performance of language models across tasks, averaged over all models, varieties, and datasets.

4.3 Error Analysis

To better understand the sources of model errors, we conduct a qualitative error analysis on the TWP task. We focus on this task because it exhibits the lowest average performance across all tasks, and therefore provides greater insight into model limitations. We select the best-performing model based on average similarity on the GEN dataset, which is OLMO. For each dataset (WEB, GEN) and language variety (en-AU, en-IN), we extract the 30 lowest-scoring test instances according to semantic similarity, yielding a total of 120 examples. These examples are manually analysed by native speakers of the language varieties, and categorised into five recurrent failure types.

Error Categories We briefly describe each error category below and provide illustrative examples in Table 6.

- *Literalisation* occurs when the model predicts a literal referent or standard lexical item corresponding to the slang phrase. While semantically appropriate, such predictions lose the idiomatic and regional character of the original expression.
- *Generic Substitution* refers to cases where the model preserves the general tone or evaluative meaning but replaces the target phrase with a non-regional or broadly applicable alternative. This reflects limited sensitivity to locale-specific lexical choice.
- *Semantic Drift* captures instances where the model remains within the correct topical or semantic field but selects a conceptually adjacent phrase that alters the intended meaning. These errors often arise in contexts that under-specify the precise pragmatic function of the slang term.
- *Contextual Misinterpretation* occurs when the model fails to correctly interpret situational or

discourse cues, resulting in a prediction that changes the narrative or pragmatic force of the sentence.

- *True Failure* denotes outputs that are incoherent, grammatically ill-formed, or entirely unrelated to the surrounding context.

Table 7 summarises the distribution of error categories across datasets and language varieties. Across both datasets and varieties, the most frequent failure types are *generic substitution*, *semantic drift*, and *contextual misinterpretation*. These categories reflect cases where the model captures aspects of the meaning or tone of the masked phrase but fails to recover the regionally appropriate slang expression. Notably, true failures, where the output is incoherent or unrelated, are relatively rare for en-IN but more prevalent for en-AU in the WEB dataset, suggesting uneven robustness across varieties and domains.

Overall, this analysis shows that most errors arise not from complete semantic failure, but from an inability to recover regionally appropriate slang expressions. Models frequently demonstrate partial understanding of meaning and tone, yet struggle with the cultural and pragmatic specificity required for accurate slang generation. This finding aligns with the large gap observed between similarity and accuracy in the TWP task.

5 Related Work

Recent work has increasingly recognised that LLMs exhibit systematic performance gaps when processing non-standard English varieties. Wu-raola et al. (2024) demonstrates that leading LLMs under-perform on comprehension tasks involving Nigerian English, particularly in emotion labeling and paraphrasing tasks. Similar studies on Indian English (Khanuja et al., 2020), African-American English (Deas et al., 2023) and Nigerian English (Srirag et al., 2025a) reveal language models perform poorly in comparison to Standard American and British English varieties. These findings highlight that despite training on broad web corpora, contemporary language models fail to adequately represent linguistic diversity.

Slangs are a lexical evidence of the cultural knowledge of a community. Recent work explores whether language models capture cultural knowledge and conventions. Seth et al. (2024) employ knowledge elicitation tasks to assess cultural familiarity across diverse contexts, finding that models

Domain	TWP		TWP*		TWS	
	Acc.	Sim.	Acc.	Sim.	Acc.	Sim.
WEB	0.04	0.29	0.07	0.32	0.50	0.61
GEN	0.01 (0.03)	0.26 (0.03)	0.02 (0.05)	0.24 (0.08)	0.49 (0.01)	0.61 (0.00)

Table 4: Effect of domain on task performance, averaged across models and varieties. The decrease in performance is shown in (red).

Variety	TWP		TWP*		TWS	
	Acc.	Sim.	Acc.	Sim.	Acc.	Sim.
en-AU	0.02	0.26	0.04	0.27	0.44	0.57
en-IN	0.03 (0.01)	0.28 (0.02)	0.05 (0.01)	0.29 (0.02)	0.54 (0.10)	0.65 (0.08)

Table 5: Effect of language variety on task performance, averaged across models and domains. The increase in performance is shown in (blue).

Category	Target phrase	Description	Prediction
Literalisation	maggot bag	Australian slang for a meat pie	pie
Generic Substitution	absolute unit	Someone very heavy or large	behemoth
Semantic Drift	bogan	An unsophisticated person from a working-class background	heavy metal band
Contextual Misinterpretation	a kangaroo loose in the top paddock	Intellectually inadequate	more to this picture
True Failure	amber fluid	Australian slang for beer	34

Table 6: Representative examples of error categories observed in OLMO predictions on the GEN dataset.

Category	GEN		WEB	
	en-AU	en-IN	en-AU	en-IN
Literalisation	2	3	3	5
Generic Substitution	6	8	1	7
Semantic Drift	7	6	3	4
Contextual Misinterpretation	7	7	4	9
True Failure	8	6	19	5
Total	30	30	30	30

Table 7: Distribution of error categories across the 30 lowest-scoring instances per dataset and language variety for the best-performing model (OLMO).

underperform for non-Anglo-centric cultures. Rao et al. (2025) develop NormAd, a scenario-based evaluation framework for cultural norm understand-

ing through closed-form QA, demonstrating that models struggle with non-Western cultural contexts. These works collectively demonstrate that language models’ cultural knowledge remains predominantly Western-centric.

The focus of this paper is slang. Slang represents a particularly challenging domain for LLMs due to its dynamic, community-specific nature. Informal language exemplifies the transient nature of evolving languages; thus, language models trained on temporally fixed corpora raise questions about continuous adaptability to emerging linguistic concepts. Mei et al. (2024) propose a causal inference framework for slang comprehension using an Urban Dictionary-based dataset, demonstrating meth-

ods for adapting models to novel slang terms. Sun et al. (2024) compile a slang detection dataset from movie transcripts and achieve strong performance with both open and closed-source models on identification tasks. However, these works either focus exclusively on Standard American English or do not distinguish between English varieties, limiting their applicability to understanding variety-specific slang. Ours is the first work focusing on two varieties of English: Australian and Indian English which by themselves represent Englishes spoken as the first and additional language respectively.

6 Conclusion

This work presents the first systematic evaluation of language models’ ability to understand variety-specific slang across Indian English (en-IN) and Australian English (en-AU). Through two complementary datasets; web-sourced examples (WEB) and synthetically generated scenarios (GEN), we assessed seven language models ranging from 110M to 8B parameters using a mask-filling evaluation framework encompassing both target word prediction and target word selection tasks. Our findings revealed three key insights. First, all evaluated models—regardless of architecture or scale—struggle to spontaneously generate appropriate slang terms in context, achieving at most 0.13 average accuracy in open-vocabulary prediction. This suggests that slang comprehension remains a significant blind spot even for state-of-the-art systems. Second, models show dramatic improvement when the task is reformulated as multiple-choice selection (with average accuracies reaching up to 0.8), this success highlights a fundamental asymmetry: models can discriminate between plausible slang alternatives more effectively than they can generate contextually appropriate slang. Third, larger models exhibit greater sensitivity to distributional shifts between naturalistic and synthetic contexts, with performance gaps of up to 0.1, suggesting that increased scale may lead to overfitting on specific textual patterns rather than robust understanding of lexical variation.

Our systematic failure analysis of 120 worst-performing predictions reveals the underlying nature of these limitations. We identify five distinct error categories that illuminate how models fail: literalization, generic substitution, semantic drift, contextual misinterpretation, and true failure. Critically, the prevalence of literalization and generic

substitution errors reveal that models understand the semantic content but fail to recognize or generate the culturally embedded linguistic forms that distinguish regional varieties. The prevalence of these errors demonstrates that models have learned to map slang to underlying concepts but do not appropriately deploy variety-specific expressions. These results have important implications for the development of language technologies that serve diverse global populations. As language models are increasingly deployed in applications ranging from content moderation to educational tools, their limitations in understanding non-standard varieties and informal language pose risks of bias and reduced utility for speakers of these varieties. Our datasets and evaluation framework provide a foundation for future research on variety-specific informal language understanding, enabling more comprehensive assessments of linguistic diversity in natural language processing systems. Future work should explore methods for improving slang comprehension, including targeted data collection, variety-aware training objectives, and continual learning approaches that can adapt to evolving linguistic phenomena.

Limitations

While we introduce two novel datasets, the WEB component is relatively small, comprising 377 unique entries. Furthermore, these examples are sourced exclusively from Wiktionary and Urban Dictionary. While these platforms provide valuable peer-contributed data, they may contain noise or demographic biases that do not fully reflect the breadth of spoken slang in these regions. Secondly, our analysis is restricted to two specific varieties: Indian English (en-IN) and Australian English (en-AU). While these varieties offer a comparison between two English typologies, our observations regarding model scaling behaviors and performance gaps may not generalize to other non-standard varieties, such as African American Vernacular English or Nigerian English. Finally, our dataset curation process relied on a single expert annotator per variety to validate the examples. Although this ensures native-level verification, it precludes the calculation of inter-annotator agreement metrics and leaves the dataset potentially susceptible to individual biases. Similarly, observations on the WEB dataset likely indicate data contamination.

Ethical Considerations

As the regulation corpus was sourced using publicly available repositories, there are no significant ethical considerations to report.

Acknowledgment

This work was funded by Google’s ExploreCSR grant, awarded to Aditya Joshi in 2024.

References

- Parul Awasthy, Aashka Trivedi, Yulong Li, Mihaela Bornea, David Cox, Abraham Daniels, Martin Franz, Gabe Goodhart, Bhavani Iyer, Vishwajeet Kumar, Luis Lastras, Scott McCarley, Rudra Murthy, Vignesh P, Sara Rosenthal, Salim Roukos, Jaydeep Sen, Sukriti Sharma, Avirup Sil, and 3 others. 2025. [Granite embedding models](#). *Preprint*, arXiv:2502.20204.
- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsen, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. [Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities](#). *Preprint*, arXiv:2507.06261.
- 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.
- Nicholas Deas, Jessica Grieser, Shana Kleiner, Desmond Patton, Elsbeth Turcan, and Kathleen McKeeown. 2023. [Evaluation of African American language bias in natural language generation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6805–6824, Singapore. Association for Computational Linguistics.
- 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.
- Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. 2024. [Generalization or memorization: Data contamination and trustworthy evaluation for large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12039–12050, Bangkok, Thailand. Association for Computational Linguistics.
- Bethany K. Dumas and Jonathan Lighter. 1978. [Is slang a word for linguists?](#) *American Speech*, 53(1):5–17.
- IBM Granite Team. 2025. [Granite 4.0 language models](#). Technical Report.
- 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, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2025. [Natural language processing for dialects of a language: A survey](#). *ACM Comput. Surv.*, 57(6).
- Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020. [GLUECoS: An evaluation benchmark for code-switched NLP](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3575–3585, Online. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. 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](#). *Preprint*, arXiv:1907.11692.
- Lingrui Mei, Shenghua Liu, Yiwei Wang, Baolong Bi, and Xueqi Cheng. 2024. [SLANG: New concept comprehension of large language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12558–12575, Miami, Florida, USA. Association for Computational Linguistics.
- Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, and 24 others. 2025. [2 olmo 2 furious](#). *Preprint*, arXiv:2501.00656.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. [Language models are unsupervised multitask learners](#). *OpenAI blog*, 1(8):9.

- Abhinav Sukumar Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2025. [NormAd: A framework for measuring the cultural adaptability of large language models](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2373–2403, Albuquerque, New Mexico. Association for Computational Linguistics.
- 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.
- Haviland Ferguson Reves. 1926. [What is slang?: A survey of opinion](#). *American Speech*, 1(4):216–220.
- Agrima Seth, Sanchit Ahuja, Kalika Bali, and Sunayana Sitaram. 2024. [DOSA: A dataset of social artifacts from different Indian geographical subcultures](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5323–5337, Torino, Italia. ELRA and ICCL.
- Dipankar Srirag, Aditya Joshi, and Jacob Eisenstein. 2025a. Predicting the target word of game-playing conversations using a low-rank dialect adapter for decoder models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 8–17.
- Dipankar Srirag, Nihar Ranjan Sahoo, and Aditya Joshi. 2025b. [Evaluating dialect robustness of language models via conversation understanding](#). In *Proceedings of the Second Workshop on Scaling Up Multilingual & Multi-Cultural Evaluation*, pages 24–38, Abu Dhabi. Association for Computational Linguistics.
- Zhewei Sun, Qian Hu, Rahul Gupta, Richard Zemel, and Yang Xu. 2024. [Toward informal language processing: Knowledge of slang in large language models](#). 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 1683–1701, Mexico City, Mexico. Association for Computational Linguistics.
- Ifeoluwa Wuraola, Nina Dethlefs, and Daniel Marciniak. 2024. [Understanding slang with LLMs: Modelling cross-cultural nuances through paraphrasing](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15525–15531, Miami, Florida, USA. Association for Computational Linguistics.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.

A Novel Scenario Generation Prompt

Given a slang phrase and its definition, create four unique scenarios. Each scenario must introduce named characters, describe the setting, and contain a single quotation that uses the word exactly once. Do not include any titles, scenario numbers, labels, emojis, or explanatory text; only output the four resulting paragraphs, separated by a line break.

phrase: {**phrase**}

definition: {**definition**}

B Dataset Examples

Refer to Table 8 for data set examples.

C Target Word Prediction Prompt Example

Fill in the blank with the best-fitting answer.

Sentence: i'd rather ride on the ____ as opposed to taking the stairs.

Answer (return only the answer with no extra text):

D Guided Target Word Prediction Prompt Example

Fill in the blank with the best-fitting answer.

The answer is a Australian English slang word or phrase.

Sentence: head on down to the beach this sunday, theres gonna be a ____ for the homeless.

Answer (return only the answer with no extra text):

E Target Word Selection Prompt Example

Fill in the blank with the best-fitting answer.

Sentence: "what's up jake?" "i just got bitten by a ____."

Options: [mossie, gonski, ripping,

Variety	en-AU	en-IN
phrase p	smoko	prepone
definition d	a slang term used on building sites in Australia, meaning a morning-tea break, or a smoke break.	Function: transitive verb Inflected forms: pre•poned; pre•pon•ing Etymology: Latin preponere to place before, prepone, from pre- + ponere to place – more at POSITION Date: Has been in use in urban English spoken in India since at least the 1950s To advance an event or activity to an earlier time. The closest American usage is “to advance” the timing of something. The word came into vogue in urban India as the opposite of “postponing” something.
usage example u	“we’ll knock off at 11 for smoko”	“To make sure we get to enjoy the fireworks display that starts promptly at 9 PM, let us prepone the dinner engagement to 7 rather than 8 tomorrow evening”
usage scenario u	The fluorescent lights of the London advertising agency hummed, a stark contrast to the quiet focus on Anika’s face as she stared at the mock-up on her screen. Her colleague, Ben, who had recently transferred from their Melbourne office, spun around in his chair. "You’ve been at that for hours, I’m grabbing a coffee and a biscuit, you keen for a quick smoko?"	Anjali tapped her stylus against the glass wall of the conference room in their bustling Bangalore office, catching Rohan’s attention as he walked by. The quarterly progress charts were displayed on the large monitor, but her focus was clearly elsewhere. "Rohan, I just got off the phone with the clients from Singapore; they’re flying in two days earlier than planned, so we need to prepone the final project presentation to Tuesday morning."

Table 8: Examples from WEB and GEN datasets.

mented]

Answer (return only the text of one option with no extra text)

F Experimental Language Models Statistics

Refer to Table 9 for experimental language model’s statistics.

Language Model	Parameters	Release Date
bert-base-uncased	110M	October 2018
roberta-large	355M	July 2019
xlm-roberta-large	550M	November 2019
granite-4.0-1b	1B	October 2025
llama-3.2-3b-instruct	3B	September 2024
olmo-2-1124-7b-instruct	7B	November 2024
qwen3-4b-instruct-2507	4B	July 2025

Table 9: Language Models used in Experimental Setup, their approximate Parameter Counts and Public Release Date

G Model Performance Metrics With the Novel Scenario Dataset GEN

Refer to Table 10 for model’s individual results using the GEN dataset.

H Model Performance Metrics With the Web-Based Example Dataset WEB

Refer to Table 11 for model’s individual results using the WEB dataset.

Language Model	Variety	Total	EM	Acc.	Sim (G)	Sim (M)	Corr
Target Word Prediction							
bert-base-uncased	en-AU	1238	2	0.002	0.726	0.254	0.341
bert-base-uncased	en-IN	243	3	0.012	0.729	0.291	0.670
roberta-large	en-AU	1238	5	0.004	0.730	0.270	0.464
roberta-large	en-IN	243	6	0.025	0.738	0.312	0.742
xlm-roberta-large	en-AU	1238	3	0.002	0.726	0.260	0.370
xlm-roberta-large	en-IN	243	0	0.000	0.727	0.284	0.494
granite-4.0-1b-Q8_0	en-AU	1232	1	0.001	0.729	0.239	0.307
granite-4.0-1b-Q8_0	en-IN	245	1	0.004	0.729	0.246	0.457
Llama-3.2-3B-Instruct-Q8_0	en-AU	1232	0	0.000	0.737	0.224	0.504
Llama-3.2-3B-Instruct-Q8_0	en-IN	245	1	0.004	0.733	0.243	0.716
olmo-2-1124-7B-instruct-Q8_0	en-AU	1232	23	0.019	0.741	0.281	0.713
olmo-2-1124-7B-instruct-Q8_0	en-IN	245	8	0.033	0.744	0.309	0.733
Qwen3-4B-Instruct-2507-Q8_0	en-AU	1232	3	0.002	0.729	0.197	0.482
Qwen3-4B-Instruct-2507-Q8_0	en-IN	245	3	0.012	0.732	0.210	0.617
Guided Target Word Prediction							
granite-4.0-1b-Q8_0	en-AU	1232	3	0.002	0.728	0.229	0.439
granite-4.0-1b-Q8_0	en-IN	245	1	0.004	0.727	0.231	0.436
Llama-3.2-3B-Instruct-Q8_0	en-AU	1232	6	0.005	0.740	0.243	0.701
Llama-3.2-3B-Instruct-Q8_0	en-IN	245	1	0.004	0.731	0.238	0.754
olmo-2-1124-7B-instruct-Q8_0	en-AU	1232	60	0.049	0.752	0.297	0.845
olmo-2-1124-7B-instruct-Q8_0	en-IN	245	17	0.069	0.757	0.323	0.816
Qwen3-4B-Instruct-2507-Q8_0	en-AU	1232	10	0.008	0.728	0.190	0.616
Qwen3-4B-Instruct-2507-Q8_0	en-IN	245	0	0.000	0.724	0.189	0.419
Target Word Selection							
granite-4.0-1b-Q8_0	en-AU	1232	437	0.355	0.828	0.511	0.978
granite-4.0-1b-Q8_0	en-IN	245	116	0.474	0.857	0.594	0.981
Llama-3.2-3B-Instruct-Q8_0	en-AU	1232	483	0.392	0.837	0.524	0.978
Llama-3.2-3B-Instruct-Q8_0	en-IN	245	123	0.502	0.863	0.606	0.985
olmo-2-1124-7B-instruct-Q8_0	en-AU	1232	485	0.394	0.839	0.555	0.982
olmo-2-1124-7B-instruct-Q8_0	en-IN	245	100	0.408	0.839	0.559	0.981
Qwen3-4B-Instruct-2507-Q8_0	en-AU	1232	725	0.589	0.890	0.681	0.987
Qwen3-4B-Instruct-2507-Q8_0	en-IN	245	196	0.800	0.946	0.842	0.987

Table 10: Model performance metrics with GEN across Target Word Prediction, Guided Target Word Prediction and Target Word Selection. We provide results on Exact Matches (EM), Accuracy (Acc), Sim(G) and Sim(M) refer to average cosine similarity scores using embedding models Granite-embedding-125m-english embeddings and (Awasthy et al., 2025) Sentence-BERT embeddings (Reimers and Gurevych, 2019) respectively. We compute Pearson Correlation (Corr.) between the resulting similarity scores to measure the two models agreement.

Language Model	Variety	Total	EM	Acc.	Sim (G)	Sim (M)	Corr
Target Word Prediction							
bert-base-uncased	en-AU	375	11	0.029	0.736	0.273	0.777
bert-base-uncased	en-IN	77	3	0.039	0.737	0.287	0.796
roberta-large	en-AU	375	22	0.059	0.743	0.304	0.824
roberta-large	en-IN	77	5	0.065	0.747	0.307	0.796
xlm-roberta-large	en-AU	375	17	0.045	0.737	0.288	0.831
xlm-roberta-large	en-IN	77	5	0.065	0.741	0.301	0.861
granite-4.0-1b-Q8_0	en-AU	374	6	0.016	0.737	0.263	0.700
granite-4.0-1b-Q8_0	en-IN	77	0	0.000	0.724	0.246	0.385
Llama-3.2-3B-Instruct-Q8_0	en-AU	374	3	0.008	0.738	0.256	0.673
Llama-3.2-3B-Instruct-Q8_0	en-IN	77	1	0.013	0.734	0.268	0.732
olmo-2-1124-7B-instruct-Q8_0	en-AU	374	24	0.064	0.748	0.307	0.836
olmo-2-1124-7B-instruct-Q8_0	en-IN	77	3	0.039	0.744	0.311	0.822
Qwen3-4B-Instruct-2507-Q8_0	en-AU	374	17	0.046	0.745	0.275	0.851
Qwen3-4B-Instruct-2507-Q8_0	en-IN	77	5	0.065	0.757	0.308	0.884
Guided Target Word Prediction							
granite-4.0-1b-Q8_0	en-AU	374	16	0.043	0.741	0.282	0.853
granite-4.0-1b-Q8_0	en-IN	77	4	0.052	0.744	0.301	0.849
Llama-3.2-3B-Instruct-Q8_0	en-AU	374	8	0.021	0.749	0.288	0.750
Llama-3.2-3B-Instruct-Q8_0	en-IN	77	1	0.013	0.736	0.301	0.748
olmo-2-1124-7B-instruct-Q8_0	en-AU	374	38	0.102	0.768	0.348	0.893
olmo-2-1124-7B-instruct-Q8_0	en-IN	77	10	0.130	0.774	0.394	0.889
Qwen3-4B-Instruct-2507-Q8_0	en-AU	374	34	0.091	0.760	0.302	0.888
Qwen3-4B-Instruct-2507-Q8_0	en-IN	77	7	0.091	0.759	0.346	0.899
Target Word Selection							
granite-4.0-1b-Q8_0	en-AU	374	142	0.380	0.836	0.533	0.978
granite-4.0-1b-Q8_0	en-IN	77	42	0.546	0.873	0.646	0.990
Llama-3.2-3B-Instruct-Q8_0	en-AU	374	151	0.404	0.842	0.534	0.976
Llama-3.2-3B-Instruct-Q8_0	en-IN	77	28	0.364	0.824	0.494	0.988
olmo-2-1124-7B-instruct-Q8_0	en-AU	374	155	0.414	0.846	0.566	0.980
olmo-2-1124-7B-instruct-Q8_0	en-IN	77	36	0.468	0.855	0.615	0.981
Qwen3-4B-Instruct-2507-Q8_0	en-AU	374	225	0.602	0.893	0.691	0.987
Qwen3-4B-Instruct-2507-Q8_0	en-IN	77	61	0.792	0.942	0.833	0.991

Table 11: Model performance metrics with WEB across Target Word Prediction, Guided Target Word Prediction and Target Word Selection. We provide results on Exact Matches (EM), Accuracy (Acc), Sim(G) and Sim(M) refer to average cosine similarity scores using embedding models Granite-embedding-125m-english embeddings and (Awasthy et al., 2025) Sentence-BERT embeddings (Reimers and Gurevych, 2019) respectively. We compute Pearson Correlation (Corr) between the resulting similarity scores to measure the two models agreement.

Effects of Speaker Bias in Dialect Identification and Automatic Transcription with Self-Supervised Speech Models

Olli Kuparinen

Faculty of Information Technology and Communication Sciences

Tampere University

olli.kuparinen@tuni.fi

Abstract

A major issue in audio modeling is speaker bias, in which the models learn language external traits, such as a speaker’s timbre or pitch, and use this information as a shortcut to a language task. This is especially problematic for dialectology, as it is typical in dialect corpora that only a few speakers represent a complete dialect area. In this paper, we explore the effects of speaker bias in two dialectal tasks: dialect identification and automatic dialectal transcription. We build two different data partitions of dialect interviews in Finnish and Norwegian: 1) a speaker dependent partition in which all of the speakers appear in training, development, and test sets, and 2) a speaker independent partition where each speaker only appears in exactly one set. We further experiment with modifications of the training data by augmenting the original audio with pitch shifts and noise, as well as changing the original speakers’ voices with voice conversion models. We show that the dialect identification models are highly affected by speaker bias, whereas automatic dialectal transcription models are not. The audio modifications do not offer major performance gains for either of the languages or tasks.

1 Introduction

Natural language processing (NLP) has long been focused on texts, mostly collected from the internet. The recent development of self-supervised speech models such as wav2vec2.0 (Baeovski et al., 2020) and Whisper (Radford et al., 2022) has however shifted the focus more towards audio data, offering possibilities for automatic speech recognition (ASR), speech synthesis, and spoken language identification, for instance. A similar shift can be seen in dialectologically inclined NLP, for which data has typically been text in the form of (phonetically) transcribed speech or user-generated content from social media.

One major difference between text and audio in

dialectal tasks is the nature of the medium: speech consists of many speaker-related effects (timbre, pitch, duration, etc.) in addition to the linguistic content, whereas text is formally more consistent (written mostly in standardized alphabets). This raises potential issues, as the speech models learn speaker-specific traits instead of (or at least in addition to) dialectal traits. Since dialectal datasets often include only a few speakers per dialect, this can lead trained models to neglect dialectal information and only focus on the speaker effects as a shortcut to dialect identification, for instance. This effect can be called **speaker bias** (or speaker leakage) in audio models (Abdullah et al., 2025).

In this work, we analyze the effects of speaker bias in two dialect-focused NLP tasks: **dialect identification** and **automatic dialectal transcription**. We use interview data from two unrelated languages in Finnish and Norwegian, and create two different data partitions to showcase the potential issues in data processing. Based on these partitions, we analyze how speaker bias can alter the perceived performance of models in dialectal audio tasks. We further explore typical methods used to mitigate speaker bias in speech modeling, such as audio augmentation and voice conversion, as possible solutions to the raised issues. Working on pre-trained models, the focus of the paper is not in best possible performance, but the performance differences introduced in data preprocessing. The main contributions of the paper are thus:

- show and analyze the effects of speaker bias in dialect identification and automatic dialectal transcription,
- explore possible solutions in audio augmentation and voice conversion, and
- analyze performance in the two tasks in two unrelated languages.

Style	SKN (fi)	LIA (no)
Transcript	j os vuav ver lähttöö	denn fysste kjirrkjå så va byggde # sto på enn annja plass
Standard	j os vain veri lähtee	den første kyrkja som var bygd stod på ein annan plass
English	if only the blood bleeds	the first church that was built was in another place

Table 1: Examples from the two datasets with the dialectal transcription on top and a standard language alternative below. Our Norwegian dataset is standardized to Nynorsk. The # in the Norwegian example denotes a pause in speech. An English gloss is presented at the bottom.

2 Related Work

2.1 Dialect Identification

Language and dialect identification from text is a standard task in natural language processing. The automatic distinction between distant languages has been declared solved (McNamee, 2005), but for similar languages and dialects the task is still relevant.

Dialect identification has been extensively studied in the the VarDial workshops, that have often included a shared task in discriminating between similar languages and dialects (e.g., Gaman et al., 2020; Chakravarthi et al., 2021; Aepli et al., 2023). For a long time, traditional linear classifiers such as support vector machines, naïve Bayes, and logistic regression, offered the best performance in the dialect identification tasks (e.g., Wu et al., 2019; Jauhainen et al., 2019; Camposampiero et al., 2022). Another popular option has been to fine-tune BERT for classification tasks (e.g., Zaharia et al., 2020; Bengoetxea et al., 2025). Related to the languages concerned in this paper, Hämäläinen et al. (2021) train a text only and text+audio dialect classifiers on the same Finnish dataset that we use. However, they split their data based on utterances only, corresponding to the speaker dependent set of our work (see Section 3.1), and use a more fine-grained dialect division.

Dialect identification from audio files has gained more interest after the release of the large pre-trained audio models. Systems utilizing Whisper (e.g., Elleuch et al., 2025) and wav2vec2.0 (e.g., Gutscher and Pucher, 2025) for dialect identification have become increasingly popular for instance in the Interspeech conferences. Many works comment on the problems with identifying dialects directly from the original audio and propose different workarounds, such as low-pass filtering and F0 monotonization (Parsons et al., 2025) or voice conversion (Abdullah et al., 2025; Fischbach et al., 2025), as well as model modifications (Luo and

Zhou, 2023). Kakouros and Hiovain-Asikainen (2023) present results on North Sami dialect identification, and also experiment with splitting their data into speaker dependent and speaker independent partitions, which is also a part of this study (see Section 3.1.)

2.2 Automatic Speech Recognition on Dialects

Most works on automatic speech recognition (ASR) focus on producing standard language text, even if the spoken language would be non-standard. This is a natural goal, given that many downstream applications need commands in the standard language. There is a broad field of studies on making dialectal speech automatically recognized to the standard language (e.g., Plüss et al., 2022; Miwa and Kai, 2023; Lin et al., 2024).

Another popular direction in automatic speech recognition has been automatic phoneme recognition, aiming to train (universal) systems that recognize the phonemes in speech and output corresponding IPA symbols (e.g., Li et al., 2020, 2022; Glocker et al., 2023). Automatic dialectal transcription can be characterized by being somewhere in between the standard language ASR and phoneme recognition: transcriptions aim to be phonetically precise, but are often language-specific and might not make some distinctions that are not relevant for the language (e.g., sibilants in Finnish are transcribed as /s/ irrespective of their true phonetic nature). The difference between dialectal transcriptions and standard languages in our datasets is presented in Table 1.

Works that aim to automatically transcribe speech to this domain instead of the standard language or the phoneme level are scarce. Suwanbandit et al. (2023) release a dataset of Thai dialects with transcriptions and translations to standard Thai, and further report on ASR experiments with the dataset. Kuparinen (2025) trains automatic dialectal transcription models on the same Finnish and Norwegian datasets that are described

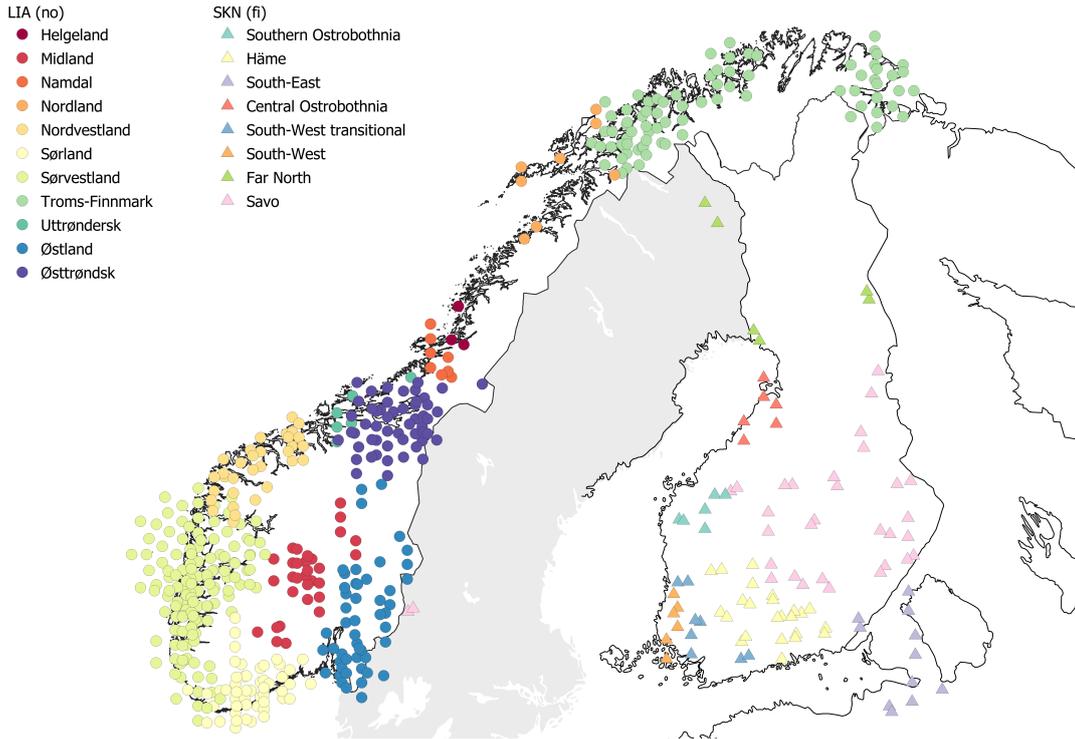


Figure 1: The speakers of both datasets on a map with dialect information as color. Norwegian speakers are presented with circles and Finnish speakers with triangles. If there are multiple speakers per location, the points are dislocated (e.g., Bergen on the West coast of Norway). In the Norwegian classification, Namdal and Utrøndersk are concatenated as well as Helgeland and Nordland. Map is made with QGIS.

in this work, but only utilizes the speaker dependent data partition. Some works also aim to produce both standard language text and dialectal text and compare the systems (Nigmatulina et al., 2020; Blaschke et al., 2025). Since dialectal transcriptions tend to have varying styles and quality, the task can be compared to low resource ASR, where data volume is likewise low and available texts might not be highly standardized.

2.3 Audio Modification

A typical way to mitigate speaker bias in speech modeling is augmenting the audio with modifications, such as changing the pitch of the voice or incorporating time or frequency masking. Speed perturbation (Ko et al., 2015) and SpecAugment (Park et al., 2019) are well known systems designed to augment the feature inputs with modifications of speech, pitch, and frequency.

Augmentations have been found to be beneficial, for example, in automatic speech recognition of dysarthric speech (Bhat et al., 2022), emotion recognition (Wu and Lee, 2023) and language assessment (Lun et al., 2024). Ullah et al. (2024) experiment with different augmentation techniques

for pre-training low resource speech models, and find that the combination of noise and pitch shifting offers best performance. We use similar augmentations in this work.

Besides augmentation of the speech features, another option for speaker bias mitigation has been the synthetic creation of new voices with systems such as HiFi-GAN (Kong et al., 2020) or using voice conversion. Voice conversion aims to transform a spoken sequence to the voice characteristics of another speaker. In essence, it makes an utterance sound as if it was spoken by someone else.

Voice conversion has been utilized, for instance, in automatic speech recognition (Casanova et al., 2023) and keyword recognition (Wubet and Lian, 2022). It has been found to be helpful in low-resource settings, where the number of natural speakers tends to be small (Baas and Kamper, 2022). This is similar to many dialectal datasets, where each dialect is often presented by only a handful of speakers. Abdullah et al. (2025) present results on Arabic dialect identification, comparing the effects of different modifications on the original audio. They show that both data augmentation

and voice conversion enhance performance both in domain and out of domain. Given their positive results, our setup follows their work.

3 Data

We use dialectal datasets of two unrelated languages, namely Finnish and Norwegian.¹ For Finnish, we use the Samples of Spoken Finnish (Institute for the Languages of Finland, 2014), which includes 99 interviews from 50 Finnish-speaking locations that present all of the Finnish dialects. The interviews were mostly recorded in the 1960s, and have since been digitized, annotated and transcribed phonetically using the Uralic phonetic alphabet. We filter the dataset to only include audio segments with a duration between 1 and 20 seconds. The filtered data includes approximately 65 hours of audio from the dialect speakers. We use the dialect division put forth by Itkonen (1989) with 8 dialect areas. The interview locations and their traditional dialect area are presented in Figure 1.

For Norwegian, we use the LIA Norwegian speech corpus (Norwegian University of Science and Technology et al., 2019), which is a joint effort of four Norwegian universities, aiming to collect the old dialectal interviews into one corpus (Hagen et al., 2021). We only use the interviews that were labeled as free talk (*fritale*) and exclude the ones that had other tasks or a special focus (such as place names). After the filtering, the dataset includes 465 speakers from 159 locations, and around 173 hours of audio from the dialect speakers. The interviews are transcribed phonetically using the Dano-Norwegian alphabet with # as a pause marker. We use the same duration limit for Norwegian as for Finnish. Furthermore, for the computationally heavy task of automatic transcription, we only use half of the available Norwegian data to make training possible with our resources.

We use the dialect division presented by Skjekke-land (1997)². The dataset is unbalanced on dialects however, with 137 speakers of the South-Western (Sørvestland) dialects and only three speakers from Helgeland. We thus combine some of the smaller dialect areas based on their top-level dialect: Namdalsk and Utrøndersk (both part of the Trøndersk

¹The segmented data are available at <https://huggingface.co/collections/okuparinen/dialectal-transcription-fi-no>.

²We utilize the mapping of municipality names and dialect areas by Phoebe Parsons, available at <https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-92/>.

dialects), and Nordlandsk and Helgelandsk (both part of the Northern dialects). The interview locations and their dialect areas are presented in Figure 1.

3.1 Data Splitting

In natural language processing tasks, (dialectal) data are often split into training, development, and test sets by splitting interviews in proportions (e.g., 80% to training, and 10% to both development and test sets). Since all speakers appear in all data splits, the tasks might be easier than actual use cases in the wild. This is especially true for audio tasks, where models learn speaker-specific traits, and can use these traits as a shortcut to the actual task (e.g., dialect identification). We will construct a data split with the described basic setup, with each speakers' utterances divided 80/10/10, which will be called a **speaker dependent** set (following Kakouros and Hiovain-Asikainen, 2023).

As a comparison, we also construct a **speaker independent** setup. In this setup, we split the data into training, development, and test sets on full interviews, assigning 80% of interviews to training, 10% to development, and 10% to testing (see also Kuparinen et al., 2023). Furthermore, as one of our tasks is dialect identification, we make sure that all dialects are always presented in all of the data splits, assigning interviews to the sets based on the amount of speakers per dialect as presented in Figure 1. We will also do further augmentation and voice conversion on this data split, described in Sections 3.2 and 3.3.

3.2 Audio Augmentation

For audio augmentation, we use both pitch shift and additional noise. Using torchaudio (Hwang et al., 2023), we take the original audio sample and randomly shift the pitch of the voice with either -4, -2, +2, or +4 steps. Furthermore, we add low level noise to the pitch shifted waveform matching the duration of the signal. Thus, the original audio will have a slightly different pitch than before, and the whole segment will have added noise on the background. To keep the setup relatively simple, we do not use any further augmentations (such as frequency or time masking). Both pitch shift and noise were found to be beneficial for Arabic dialect identification by Abdullah et al. (2025).

Setup	Training	Train size	Orig.	Aug.	VC	Test
Speaker dependent						
Original	all speakers	N	✓	✗	✗	all speakers
Voice per dialect	all speakers	N	✗	✗	✓	all speakers
Speaker independent						
Original	80% of speakers	N	✓	✗	✗	10% of sp.
Orig. + Pitch shift and noise	80% of speakers	$2 \times N$	✓	✓	✗	10% of sp.
Orig. + VC with one voice	80% of speakers	$2 \times N$	✓	✗	✓	10% of sp.
Orig. + VC with four voices	80% of speakers	$5 \times N$	✓	✗	✓	10% of sp.

Table 2: Summary of the different data splits and operations on the training set. The first two setups serve as benchmarks for speaker bias in speech models. N =the size of original training utterances. Orig.=original audio, Aug.=audio augmentation with pitch shifts and noise, VC=voice conversion of the original audio. Sp. = speakers. Development set has the same setup as the test set. Exact sizes of the different sets are presented in Appendix A.

3.3 Voice Conversion

We use nearest neighbor voice conversion (kNN-VC) presented in Baas et al. (2023).³ The approach utilizes reference segments from another speaker to convert the original voice, while still maintaining the dialectal content. As reference speakers, we use external dialectal datasets: for Finnish, we utilize the Finnish Dialect Corpus of the Syntax Archive (University of Turku and Institute for the Languages of Finland, 2021), and for Norwegian we use readings of the North Wind and the Sun in different dialects.⁴

For our voice conversion setups, all of the training samples are converted to the voice of another speaker. These voice converted training sets are then concatenated with the original training samples, effectively multiplying the size of the training data (as is done with the augmented data as well). We experiment with one voice (training data twice the size of the original) and with four voices (training five times the size of the original).

As a final experiment, we also construct a voice converted version of the speaker dependent dataset, where each dialect is represented by exactly one voice. This setup is called **voice per dialect** and is designed to highlight what happens when speaker traits and dialect traits are combined. If our assumption is correct, and the models use speaker identification as a shortcut to dialect identification, classifiers trained on the voice per dialect set should perform poorly when tested on new voices. The different data setups are summarized in Table 2. The

³The implementation is available at <https://github.com/bshall/knn-vc>.

⁴Available at <https://www.hf.ntnu.no/nos/>.

development and test sets are composed of the original speech samples throughout the experiments without any modifications.⁵

4 Task Setup

To evaluate the effect of speaker bias in dialectological speech modeling, we set up two different dialect-focused tasks: dialect identification and automatic dialectal transcription. As the point of the work is not in achieving best possible results for each task, we do not experiment with different base architectures and hyperparameters. Instead, we use wav2vec2.0 (Baevski et al., 2020) based models with basic settings for both tasks. For Finnish, we use the base model trained with 150,000 hours of Finnish speech⁶, including 2740 hours of colloquial Finnish (Getman et al., 2025). Such large base models were not available for Norwegian, which is why we use the base version of the multilingual MMS model⁷ (Pratap et al., 2023). We further report MMS results of the Finnish speaker independent set with original audio to facilitate base model comparison.

4.1 Dialect Identification

For the dialect identification task, we extract embeddings of our audio samples encoded by the base models. As it is known that the different layers of the models encode different aspects of the speech signal, we experimented with the first layer, middle layers (6 for Finnish, 6 and 12 for Norwegian), and

⁵The code for the paper is available at: <https://github.com/okuparinen/DialectSpeakerBias>.

⁶<https://huggingface.co/GetmanY1/wav2vec2-base-fi-150k>

⁷<https://huggingface.co/facebook/mms-300m>

the final layer (12 for Finnish, 24 for Norwegian) for the embedding extraction. We found that layer number 6 offers best performance in our classification task for both languages. We thus train our final classifiers on the embeddings extracted from this layer. A comparison of the different layers is presented in Appendix B.

The embeddings have dimensions of $N \times 768$ for the Finnish model, and $N \times 1024$ for the MMS model, where N is the number of 25 ms frames in the utterance. Since the utterances have differing length, we aggregate the embeddings over the utterances. We take the mean and standard deviation over the full utterance (resulting in 1536 dimensions per utterance for Finnish and 2048 dimensions for Norwegian), and use these utterance-based embeddings to train a linear support vector machine classifier.

For our classifier, we first scale the data for zero mean and unit variance before optionally applying principal component analysis on the scaled data. The data is then fed to a linear, class-balanced SVM using one-vs-rest classification strategy. We experiment with the input embedding dimensions (using mean and standard deviation, or only mean or only standard deviation), as well as with using PCA or not using it. For PCA, we also experiment with the number of components (128 or 256 for Finnish and 128, 256 or 512 for Norwegian), and select the best design for each dataset. As a text comparison, we train a similar SVM model with a tf-idf vectorizer on character n-grams of 2 to 4 characters, without applying PCA, on the manual transcriptions.

We use the classifier to predict dialect labels for the development and test sets based on their embedding representations. We evaluate the results on macro F1 and accuracy. We also report the 95% confidence intervals for the results.⁸

4.2 Automatic Dialectal Transcription

We finetune the base models with our transcribed dialectal data for a maximum of 15 epochs (early stopping of 10 epochs) with connectionist temporal classification (CTC) loss. We freeze the feature extractor before finetuning and use a learning rate of $5 \cdot 10^{-4}$ for the smaller Finnish dataset and $1 \cdot 10^{-4}$ for the larger Norwegian dataset. The finetuning is done with the Huggingface Transformers toolkit (Wolf et al., 2020). We evaluate the systems on

⁸Calculated with <https://github.com/luferrer/ConfidenceIntervals> with 1000 bootstrap sets.

character error rate (CER), which in transcription text is more informative than word error rate.⁹

5 Results

5.1 Dialect Identification

The results for the dialect identification experiments are presented for both languages in Table 3. The first rows of the tables have the same data structure (each interview split into training, development, and test sets based on utterances), as do the bottom rows (each interview appears in exactly one set).

Starting with the speaker dependent setup, very clear evidence of speaker bias can be observed. Using only the original audio, the scores are very high, indicating that the classifier learns to connect the speaker traits to a dialect in the training phase. This effect becomes evident when comparing it to the voice per dialect set, which uses the exact same data split but with converted voices for each dialect. When evaluated against the same test set, the classification performance collapses as the classification model has learned to connect the speakers to the dialects, but is then faced with unheard voices. The text baseline is solid for Finnish, but worse than the audio based model. For Norwegian, the text model is far behind the audio model.

It is also noticeable that for the speaker dependent set, both languages have very similar performance in the audio models: around 90% accuracy scores for the original audio and around 17% for the voice per dialect setup. This is not true for the speaker independent setup, where Norwegian has much higher scores throughout. This also indicates that, for the speaker dependent setup, the linguistic factors are not as important as the speaker traits for the classification. In essence, the model is more of a speaker recognition system than a dialect identification system.

For the speaker independent data split, we see a performance drop of around 60 points on original audio for Finnish, and around 40 points for Norwegian. This further indicates there is a major speaker related effect in the classification. For the text based models, there is also a considerable effect of the data split change for Finnish, but not so much for Norwegian. For Finnish, the text based model clearly outperforms audio, whereas for Norwegian the audio models are slightly better.

⁹We use the implementation provided in <https://github.com/nsmartinez/WERpp>.

SKN (fi)		
Setup	Macro F1 \uparrow	Accuracy \uparrow
Speaker dependent		
Transcription text	74.85 ^{73.10–76.53}	77.83 ^{76.48–79.12}
Original audio	88.77 ^{87.44–90.04}	89.79 ^{88.84–90.73}
Voice per dialect	16.07 ^{14.72–17.41}	17.01 ^{15.75–18.28}
Speaker independent		
Transcription text	61.88 ^{60.32–63.51}	65.23 ^{63.74–66.70}
Original audio	24.29 ^{23.06–25.52}	33.38 ^{31.83–34.90}
Orig. + Pitch shift and noise	27.55 ^{26.10–28.89}	32.85 ^{31.25–34.40}
Orig. + VC with one voice	21.05 ^{19.81–22.26}	26.85 ^{25.49–28.37}
Orig. + VC with four voices	29.63 ^{28.21–30.98}	38.09 ^{36.65–39.64}
LIA (no)		
Setup	Macro F1 \uparrow	Accuracy \uparrow
Speaker dependent		
Transcription text	53.01 ^{51.89–54.05}	58.91 ^{57.98–59.76}
Original audio	91.69 ^{91.04–92.30}	92.56 ^{92.09–93.01}
Voice per dialect	12.58 ^{11.95–13.19}	17.45 ^{16.75–18.18}
Speaker independent		
Transcription text	47.17 ^{46.03–48.25}	54.30 ^{53.34–55.24}
Original audio	50.47 ^{49.59–51.37}	64.52 ^{63.55–65.47}
Orig. + Pitch shift and noise	51.65 ^{50.89–52.44}	64.97 ^{64.12–65.90}
Orig. + VC with one voice	51.32 ^{50.58–52.15}	65.32 ^{64.39–66.21}
Orig. + VC with four voices	49.31 ^{48.55–50.07}	62.91 ^{61.93–63.88}

Table 3: Macro F1 and accuracy scores for the classification task with 95% confidence intervals in superscript. The top table shows the SKN (Finnish) results, and the bottom table shows the LIA (Norwegian) results. Per-chance accuracy is 12.5 for Finnish and 11.11 for Norwegian. For comparison, using the MMS model for the Finnish speaker independent set with original audio achieves a macro F1 score of 27.05^{25.81–28.19} (+2.79 vs. monolingual). Results on the development set and model details (embedding dimensions, PCA parameters) are provided in Appendix B.

Setup	SKN (fi)	LIA (no)
Speaker dependent		
Original audio	9.65 ^{9.43–9.87}	19.30 ^{19.02–19.55}
Voice per dialect	16.81 ^{16.47–17.19}	29.38 ^{29.04–29.72}
Speaker independent		
Original audio	15.89 ^{15.46–16.31}	20.21 ^{19.91–20.50}
Orig. + Pitch shift and noise	16.43 ^{16.00–16.83}	19.53 ^{19.31–19.87}
Orig. + VC with one voice	16.35 ^{15.91–16.76}	20.17 ^{19.88–20.46}
Orig. + VC with four voices	17.55 ^{17.16–17.95}	20.99 ^{20.71–21.26}

Table 4: Character error rate % (\downarrow) results of the automatic transcription on the different data splits with 95% confidence intervals in superscript. The MMS model finetuned for the Finnish speaker independent set with original audio achieves a character error rate of 19.45^{19.08–19.85} (+3.56 vs. monolingual).

The performance of the audio models is poor for Finnish across the board, whereas for Norwegian the models fare better. Regarding the audio modifications, there is a lack of a consistent and meaningful effect. For both languages, the augmentation with pitch shifts and noise offers a tiny performance gain in macro F1 over the original audio, whereas the voice conversions seem unstable. For Finnish, the model with four additional voices is clearly the best of the audio models, but for Norwegian it is the worst. Likewise the system with one additional voice is the best model for Norwegian in terms of accuracy but clearly the worst for Finnish.

The classification results are very different from the ones reported by e.g., [Abdullah et al. \(2025\)](#), who present a consistent performance gain from augmentation and voice conversion. Possible causes for this are that we are working with more dialect classes (8 and 9 vs. 5) and older data with possibly inconsistent quality, but the lack of positive results is still surprising.

5.2 Automatic Dialectal Transcription

The results for the automatic dialectal transcription task are presented in Table 4. For the speaker dependent models, the original audio still offers better performance than the voice per dialect system. For Finnish, the difference is not as large as for the classification task, however. In fact, the voice per dialect system is only slightly worse than the original audio based system in the speaker independent split for Finnish (although the speaker dependent split is easier in that the same topics appear in both training and testing, even if the voices are different).

For Norwegian, the difference between the original audio and voice per dialect systems in the speaker dependent set is very large. Interestingly, the difference between the dependent and independent sets is quite small (0.91 points in CER % for the original audio). For Finnish the difference is larger, but still not as big as for the classification task. This indicates that automatic transcription is not as dependent on the speakers as the classification.

In terms of character error rate, the modifications of the original audio are consistently harmful for the transcription quality in the Finnish experiments. Adding more voices and thus more training data with the same transcriptions seems to make the models worse, possibly overfitting to the train-

ing data. For Norwegian, however, the scores are similar for all versions, but the augmented data offers best results. This was also the case in the classification task.

Comparing the two languages, results for Finnish are consistently better than for Norwegian. This is most likely a result of at least two causes: 1) for Finnish, we could use a language-specific base model but for Norwegian we had to use a multilingual one¹⁰, and 2) for the Norwegian data, we noticed inconsistent quality in both the audio and the transcriptions, most likely resulting from the fact that the corpus is collected from multiple sources. This is also reflected in the transcription-based classification, where the Finnish models are performing better than the Norwegian ones, indicating possible variation in the Norwegian transcription quality. Finally, in the audio classification the results were converse: Finnish models are much worse than Norwegian. This could be a result of clearer dialectal differences in the Norwegian data, which makes identification easier, but building a unified transcription model harder.

5.3 Classifying on Automatic Transcriptions

As a final experiment, we analyze if the classification scores for Finnish can be elevated by classifying the utterances based on the automatically created transcriptions presented in Section 5.2. We use the text models trained on manual transcriptions for the original classification experiments in Section 5.1, but infer on the automatically created transcriptions of the test set. We only analyze the results on the Finnish data, as the Norwegian audio models outperformed the text based systems. The results are presented in Table 5.

The classification on the automatic transcriptions enhances performance dramatically with macro F1 scores around 25 points higher compared to the audio based classification. Even the highly speaker dependent system of voice per dialect achieves a macro F1 score of 45.82. In the speaker independent setup, the ASR model based on original audio was the best in terms of character error rate, but the worst in terms of classification F1 (albeit the differences are small for both cases). In conclusion, the text based models massively outperform audio in speaker independent setups for Finnish, even if the text is automatically created.

¹⁰Using MMS for Finnish enhanced performance in the classification task, but worsened it in the transcription task, following the Norwegian results.

Setup	Audio	ASR
Speaker dependent		
Original audio	88.77 ^{87.44–90.04}	62.86 ^{60.96–64.59}
Voice per dialect	16.07 ^{14.72–17.41}	45.82 ^{44.06–47.49}
Speaker independent		
Original audio	24.29 ^{23.06–25.52}	48.10 ^{46.53–49.72}
Orig. + Pitch shift and noise	27.55 ^{26.10–28.89}	50.53 ^{48.99–52.12}
Orig. + VC with one voice	21.05 ^{19.81–22.26}	49.94 ^{48.30–51.59}
Orig. + VC with four voices	29.63 ^{28.21–30.98}	49.66 ^{48.12–51.22}
Manual transcription		
Speaker dependent		74.85 ^{73.10–76.53}
Speaker independent		61.88 ^{60.32–63.51}

Table 5: Macro F1 scores \uparrow for Finnish SVM classifiers with 95% confidence intervals in superscript. On the left, we present the results for classifiers trained and evaluated on audio embeddings, and on the right, the models trained on manual transcriptions and evaluated on automatic transcriptions. The corresponding classification results for the manual transcriptions are presented on the bottom. The Norwegian results are omitted, as the audio models already outperformed the manual transcription based systems.

6 Conclusion

In this paper, we have shown how speaker bias severely affects audio modeling in dialect identification and to a lesser extent in automatic dialectal transcription, with data from two unrelated languages. The speaker dependent setup highlighted how audio models shortcut to speaker recognition instead of dialect identification: using original audio in training achieved excellent scores for both languages, whereas using the same data split with a single voice per dialect ended in collapse.

We further aimed to mitigate the effects of bias with traditionally used techniques in audio augmentation and voice conversion. While there were some positive effects, the overall usefulness of these methods on the tasks remained negligible. This is in contrast to previous findings on, for instance, Arabic dialect identification (Abdullah et al., 2025). Especially for Finnish, dialect identification proved to be difficult for all models, and classification on transcriptions provided far better results. Conversely for Norwegian, identification from audio outperformed text, but automatic dialectal transcription performed worse than for Finnish.

This work has focused on speaker bias mitigation solutions that are applied on the input (i.e., the waveform itself). More elaborate systems could also be applied post-hoc, by targeting the embedding dimensions that hold the most information on speaker traits, and filtering or down-weighting such dimensions. Thebaud et al. (2024) use Inte-

grated Gradients to trace which phonemes affect speaker recognition the most, but a similar system could be applied also to trace speaker effects from model embeddings. Zhu et al. (2025) train ECAPA-TDNN (Desplanques et al., 2020) based speaker embeddings and use SHAP values to trace the speaker-affected dimensions from the content embeddings of different self-supervised models. They further experiment how filtering the speaker information affects ASR accuracy. Systems based on explainability methods could thus provide interesting possibilities for dialectal audio modeling as well, but they are beyond the scope of this paper and thus left for future work.

Limitations

The building of the speaker independent set is a possible source of variation. As the data splits are constructed from full interviews, the scores are highly affected by the interviews chosen to the development and test sets. A possible way to undermine this effect would be to build several folds of the data split. We restricted our experiments to one fold due to resource limitations, but included the 95% confidence intervals to show possible fluctuation.

This work focuses on two languages spoken in the Nordic countries. Even though the languages represent different families, the datasets themselves are largely collected following similar dialectological and cultural ideologies.

Acknowledgments

This work is supported by the Research Council of Finland through project No. 360356 “Speech as Speech – Acoustic Modeling in Variational Linguistics”. The author also wishes to acknowledge CSC – IT Center for Science, Finland, for computational resources.

References

- Badr M. Abdullah, Matthew Baas, Bernd Möbius, and Dietrich Klakow. 2025. [Voice Conversion Improves Cross-Domain Robustness for Spoken Arabic Dialect Identification](#). In *Interspeech 2025*, pages 2790–2794.
- Noëmi Aepli, Çağrı Çöltekin, Rob Van Der Goot, Tommi Jauhiainen, Mourhaf Kazzaz, Nikola Ljubešić, Kai North, Barbara Plank, Yves Scherrer, and Marcos Zampieri. 2023. [Findings of the VarDial Evaluation Campaign 2023](#). In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 251–261, Dubrovnik, Croatia. Association for Computational Linguistics.
- Matthew Baas and Herman Kamper. 2022. [Voice Conversion Can Improve ASR in Very Low-Resource Settings](#). In *Interspeech 2022*, pages 3513–3517.
- Matthew Baas, Benjamin van Niekerk, and Herman Kamper. 2023. [Voice Conversion With Just Nearest Neighbors](#). In *Interspeech 2023*, pages 2053–2057.
- Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.
- Jaione Bengoetxea, Mikel Zubillaga, Ekhi Azurmendi, Maite Heredia, Julen Etxaniz, Markel Ferro, and Jeremy Barnes. 2025. [HiTZ at VarDial 2025 NorSID: Overcoming Data Scarcity with Language Transfer and Automatic Data Annotation](#). In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 209–219, Abu Dhabi, UAE. Association for Computational Linguistics.
- Chitralekha Bhat, Ashish Panda, and Helmer Strik. 2022. [Improved ASR Performance for Dysarthric Speech Using Two-stage DataAugmentation](#). In *Interspeech 2022*, pages 46–50.
- Verena Blaschke, Miriam Winkler, Constantin Förster, Gabriele Wenger-Glemser, and Barbara Plank. 2025. [A Multi-Dialectal Dataset for German Dialect ASR and Dialect-to-Standard Speech Translation](#). In *Interspeech 2025*, pages 913–917.
- Giacomo Camposampiero, Quynh Anh Nguyen, and Francesco Di Stefano. 2022. [The Curious Case of Logistic Regression for Italian Languages and Dialects Identification](#). In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 86–98, Gyeongju, Republic of Korea. Association for Computational Linguistics.
- Edresson Casanova, Christopher Shulby, Alexander Korolev, Arnaldo Candido Junior, Anderson da Silva Soares, Sandra Aluísio, and Moacir Antonelli Ponti. 2023. [ASR Data Augmentation in Low-Resource Settings Using Cross-Lingual Multi-Speaker TTS and Cross-Lingual Voice Conversion](#). In *Interspeech 2023*, pages 1244–1248.
- Bharathi Raja Chakravarthi, Gaman Mihaela, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Ruba Priyadharshini, Christoph Purschke, Eswari Rajagopal, Yves Scherrer, and Marcos Zampieri. 2021. [Findings of the VarDial Evaluation Campaign 2021](#). In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–11, Kiyv, Ukraine. Association for Computational Linguistics.
- Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. [ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification](#). In *Interspeech 2020*, pages 3830–3834.
- Haroun Elleuch, Salima Mdhaffar, Yannick Estève, and Fethi Bougares. 2025. [ADI-20: Arabic Dialect Identification Dataset and Models](#). In *Interspeech 2025*, pages 2775–2779.
- Lea Fischbach, Akbar Karimi, Caroline Kleen, Alfred Lameli, and Lucie Flek. 2025. [Improving Low-Resource Dialect Classification Using Retrieval-based Voice Conversion](#). In *Interspeech 2025*, pages 2780–2784.
- Mihaela Gaman, Dirk Hovy, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Christoph Purschke, Yves Scherrer, and Marcos Zampieri. 2020. [A Report on the VarDial Evaluation Campaign 2020](#). In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–14, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).
- Yaroslav Getman, Tamás Grósz, Tommi Lehtonen, and Mikko Kurimo. 2025. [Is Your Model Big Enough? Training and Interpreting Large-Scale Monolingual Speech Foundation Models](#). In *Interspeech 2025*, pages 231–235.
- Kevin Glocker, Aaricia Herygers, and Munir Georges. 2023. [Allophant: Cross-Lingual Phoneme Recognition with Articulatory Attributes](#). In *Proc. Interspeech 2023*.
- Lorenz Gutscher and Michael Pucher. 2025. [Audio-Based Classification and Geographic Regression of Austrian Dialects](#). In *Interspeech 2025*, pages 2765–2769.

- Kristin Hagen, Gjert Kristoffersen, Øystein A. Vangsnes, and Tor A. Åfarli, editors. 2021. *Språk i arkiva: Ny forskning om eldre talemål frå LIA-prosjektet*. Novus forlag.
- Mika Hämmäläinen, Khalid Alnajjar, Niko Partanen, and Jack Rueter. 2021. *Finnish Dialect Identification: The Effect of Audio and Text*. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8777–8783, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Jeff Hwang, Moto Hira, Caroline Chen, Xiaohui Zhang, Zhaoheng Ni, Guangzhi Sun, Pingchuan Ma, Ruizhe Huang, Vineel Pratap, Yuekai Zhang, Anurag Kumar, Chin-Yun Yu, Chuang Zhu, Chunxi Liu, Jacob Kahn, Mirco Ravanelli, Peng Sun, Shinji Watanabe, Yangyang Shi, and 5 others. 2023. *TorchAudio 2.1: Advancing Speech Recognition, Self-Supervised Learning, and Audio Processing Components for PyTorch*. *Preprint*, arXiv:2310.17864.
- Institute for the Languages of Finland. 2014. *Samples of Spoken Finnish, Downloadable Version*.
- Terho Itkonen. 1989. *Nurmijärven murrekirja*. Suomalaisen Kirjallisuuden Seuran toimituksia ; 498. Suomalaisen kirjallisuuden seura, Helsinki.
- Tommi Jauhiainen, Krister Lindén, and Heidi Jauhiainen. 2019. *Discriminating Between Mandarin Chinese and Swiss-German Varieties Using Adaptive Language Models*. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 178–187, Ann Arbor, Michigan. Association for Computational Linguistics.
- Sofoklis Kakouros and Katri Hiovain-Asikainen. 2023. *North Sámi Dialect Identification with Self-supervised Speech Models*. In *Interspeech 2023*, pages 5306–5310.
- Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. *Audio Augmentation for Speech Recognition*. In *Interspeech 2015*, pages 3586–3589.
- Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. *HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis*. *Preprint*, arXiv:2010.05646.
- Olli Kuparinen. 2025. *Automatic Dialectal Transcription: An Evaluation on Finnish and Norwegian*. In *Interspeech 2025*, pages 2390–2394.
- Olli Kuparinen, Aleksandra Miletic, and Yves Scherrer. 2023. *Dialect-to-Standard Normalization: A Large-Scale Multilingual Evaluation*. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13814–13828, Singapore. Association for Computational Linguistics.
- Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R Mortensen, Graham Neubig, Alan W Black, and Metze Florian. 2020. *Universal Phone Recognition with a Multilingual Allophone System*. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8249–8253. IEEE.
- Xinjian Li, Florian Metze, David R. Mortensen, Alan W Black, and Shinji Watanabe. 2022. *Phone Inventories and Recognition for Every Language*. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1061–1067, Marseille, France. European Language Resources Association.
- Jiayan Lin, Shenghui Lu, Hukai Huang, Wenhao Guan, Binbin Xu, Hui Bu, Qingyang Hong, and Lin Li. 2024. *MinSpeech: A Corpus of Southern Min Dialect for Automatic Speech Recognition*. In *Interspeech 2024*, pages 2330–2334.
- Tin Mei Lun, Ekaterina Voskoboynik, Ragheb Al-Ghezi, Tamas Grosz, and Mikko Kurimo. 2024. *Oversampling, Augmentation and Curriculum Learning for Speaking Assessment with Limited Training Data*. In *Interspeech 2024*, pages 4019–4023.
- Qibao Luo and Ruohua Zhou. 2023. *Exploring the Impact of Back-End Network on Wav2vec 2.0 for Dialect Identification*. In *Interspeech 2023*, pages 5356–5360.
- Paul McNamee. 2005. *Language Identification: A Solved Problem Suitable for Undergraduate Instruction*. *Journal of Computing Sciences in Colleges*, 20(3):94–101.
- Shogo Miwa and Atsuhiko Kai. 2023. *Dialect Speech Recognition Modeling using Corpus of Japanese Dialects and Self-Supervised Learning-based Model XLSR*. In *Interspeech 2023*, pages 4928–4932.
- Iuliia Nigmatulina, Tannon Kew, and Tanja Samardzic. 2020. *ASR for Non-standardised Languages with Dialectal Variation: the case of Swiss German*. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 15–24, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).
- Norwegian University of Science and Technology, University of Bergen, University of Oslo, and The Arctic University of Norway. 2019. *Lia norsk - korpus av eldre dialektoptak*.
- Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. *SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition*. In *Interspeech 2019*, pages 2613–2617.
- Phoebe Parsons, Heming Strømholth Bremnes, Knut Kvale, Torbjørn Svendsen, and Giampiero Salvi. 2025. *Effects of Prosodic Information on Dialect Classification Using Whisper Features*. In *Interspeech 2025*, pages 2785–2789.

- Michel Plüss, Manuela Hürlimann, Marc Cuny, Alla Stöckli, Nikolaos Kaptotis, Julia Hartmann, Malgorzata Anna Ulasik, Christian Scheller, Yanick Schraner, Amit Jain, Jan Deriu, Mark Cieliebak, and Manfred Vogel. 2022. [SDS-200: A Swiss German Speech to Standard German Text Corpus](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3250–3256, Marseille, France. European Language Resources Association.
- Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. [Scaling Speech Technology to 1,000+ Languages](#). *arXiv*.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. [Robust Speech Recognition via Large-Scale Weak Supervision](#). *Preprint*, arXiv:2212.04356.
- Martin Skjækkeland. 1997. *Dei norske dialektane : tradisjonelle særdrag i jamføring med skriftmåla*. Høyskoleforlaget, Kristiansand.
- Artit Suwanbandit, Burin Naowarat, Orathai Sangpetch, and Ekapol Chuangsuwanich. 2023. [Thai Dialect Corpus and Transfer-based Curriculum Learning Investigation for Dialect Automatic Speech Recognition](#). In *Interspeech 2023*, pages 4069–4073.
- Thomas Thebaud, Gabriel Hernández, Sarah Flora Samson Juan, and Marie Tahon. 2024. [A Phonetic Analysis of Speaker Verification Systems through Phoneme selection and Integrated Gradients](#). In *The Speaker and Language Recognition Workshop (Odyssey 2024)*, pages 59–66.
- Asad Ullah, Alessandro Ragano, and Andrew Hines. 2024. [Reduce, Reuse, Recycle: Is Perturbed Data Better than Other Language Augmentation for Low Resource Self-Supervised Speech Models](#). In *Interspeech 2024*, pages 77–81.
- University of Turku and Institute for the Languages of Finland. 2021. [The Finnish Dialect Corpus of the Syntax Archive, Downloadable Version](#).
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. [Transformers: State-of-the-Art Natural Language Processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Nianheng Wu, Eric DeMattos, Kwok Him So, Pin-zhen Chen, and Çağrı Çöltekin. 2019. [Language Discrimination and Transfer Learning for Similar Languages: Experiments with Feature Combinations and Adaptation](#). In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 54–63, Ann Arbor, Michigan. Association for Computational Linguistics.
- Ya-Tse Wu and Chi-Chun Lee. 2023. [MetricAug: A Distortion Metric-Lead Augmentation Strategy for Training Noise-Robust Speech Emotion Recognizer](#). In *Interspeech 2023*, pages 3587–3591.
- Yeshanew Ale Wubet and Kuang-Yow Lian. 2022. [Voice Conversion Based Augmentation and a Hybrid CNN-LSTM Model for Improving Speaker-Independent Keyword Recognition on Limited Datasets](#). *IEEE Access*, 10:89170–89180.
- George-Eduard Zaharia, Andrei-Marius Avram, Dumitru-Clementin Cercel, and Traian Rebedea. 2020. [Exploring the Power of Romanian BERT for Dialect Identification](#). In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 232–241, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).
- Xiaoxu Zhu, Junhua Li, Aaron J. Li, Yiming Ren, and Baoxiang Li. 2025. [Speaker Disentanglement of Speech Pre-trained Model Based on Interpretability](#). *Preprint*, arXiv:2507.17851.

A Data Size

Table 6 presents the data sizes in whitespace separated tokens and audio duration in hours and minutes. Because the speaker independent sets consist of full interviews, they show more variation in size than the speaker dependent sets that are split on utterances.

B Development Set Results

Table 7 presents the results of the classification task on the different layers of the development set. The results clearly show that layer 6 offers best performance for all setups, and it was thus used for the final classification.

Table 8 presents the development set results for the classification task, as well as the inputs for the SVM model (using the mean and standard deviation (std), or just mean or just standard deviation of the utterance embeddings). If PCA was used, the number of components is also presented. The development set results are broadly 5–10 points higher than the test set results for Finnish, and around 5 points higher for Norwegian. For the development set, the setup with original audio and 4 voices achieves the best score of the speaker independent set for both languages.

Table 9 presents the development set results for the automatic dialectal transcription task. For

Dataset	Split	Train		Dev.		Test	
		Tokens	Audio	Tokens	Audio	Tokens	Audio
SKN (fi)	Dependent	498,587	52:02	62,907	06:35	62,356	06:33
	Independent	489,168	51:12	64,031	06:36	70,651	07:23
LIA (no)	Dependent	1,562,453	138:16	194,773	17:19	200,969	17:53
	Independent	1,575,923	140:04	206,627	17:59	175,645	15:24

Table 6: Statistics of the used data. Audio duration in hh:mm. Dev. = development set.

Setup	SKN (fi)			LIA (no)		
	L1	L6	L12	L1	L6	L12
Speaker independent						
Original audio	20.34	32.40	23.16	42.33	53.01	47.44
Orig. + Pitch shift and noise	21.06	32.43	21.97	45.85	54.08	47.72
Orig. + VC with one voice	24.31	33.82	24.89	45.56	54.68	47.83
Orig. + VC with four voices	25.87	34.88	25.70	43.07	54.87	44.97

Table 7: Layer-wise macro F1 scores in the classification task for the development set. L = layer.

Setup	SKN (fi)		LIA (no)	
	SVM	Macro F1 \uparrow	SVM	Macro F1 \uparrow
Speaker dependent				
Transcription	–	74.15	–	52.77
Orig. audio	mean+std	90.04	mean+std	92.26
Voice per dialect	std, PCA 256	15.55	std, PCA 528	13.04
Speaker independent				
Transcription	–	66.85	–	49.39
Orig. audio	mean, PCA 256	32.40	mean+std	53.01
Orig. + Aug.	std	32.43	mean+std	54.08
Orig. + VC1	mean, PCA 128	33.82	mean+std	54.68
Orig. + VC4	mean+std	34.88	mean+std	54.87

Table 8: Macro F1 scores on the development set in the classification task, as well as the SVM model inputs for the best model. The number of PCA components are presented if PCA was used.

Finnish, the results are very slightly better than for the test set, whereas for Norwegian the test set results are better by around 2 points across the speaker independent set.

Setup	SKN (fi)	LIA (no)
Speaker dependent		
Original audio	9.74	19.37
Voice per dialect	16.75	29.75
Speaker independent		
Original audio	15.00	22.00
Orig. + Aug.	15.26	21.51
Orig. + VC1	15.42	22.12
Orig. + VC4	16.99	23.01

Table 9: Character error rate % (\downarrow) results of the automatic transcription on the development set.

OcWikiDialects: A Wikipedia Dataset With Rich Metadata for Occitan Dialect Identification

Oriane Nédey Rachel Bawden Thibault Clérice Benoît Sagot
Inria, Paris, France
{firstname.lastname}@inria.fr

Abstract

Occitan is a Romance language spoken mostly in the South of France and characterised by rich dialectal variation, which can pose problems for certain NLP tools. This shortfall is largely attributable to the scarcity of dialect-annotated corpora, in a context where linguistic classification within the Occitan dialect continuum is still debated and major nomenclatures, such as ISO 639, fail to provide granular codes for varieties below the generic “Occitan” label. In this paper, we introduce OcWikiDialects, a new dataset comprising articles from the Occitan Wikipedia. The corpus features rich metadata, including dialect labels, and is segmented at both paragraph and sentence levels. Combined with previously released datasets, we explore approaches for Occitan dialect identification by training three types of model on up to 8 labels: linear SVM classifiers based on word and character n -grams, FastText classifiers based on pretrained vectors, and BERT-based neural classifiers adapted through fine-tuning. Evaluations across in- and out-of-domain test sets demonstrate the substantial impact of our new dataset for the task. However, a peak macro-averaged F1 score of 58.15 underscores persistent challenges for underrepresented Occitan varieties, supported by our per-dialect analysis. Code, dataset and models are available: <https://github.com/DEFI-COLaF/OcWikiDialects>.

1 Introduction

Current NLP technologies offer very good performance in particular with large language models (LLMs), including in some low-resource settings (Pomeranke et al., 2025). However, most available tools consider the supported languages as standardised monolithic entities, thus hiding many aspects of variation that occur in natural languages (Bird, 2022). Ignoring these aspects in the NLP development process has an impact on speakers of under-represented and less standardised varieties,

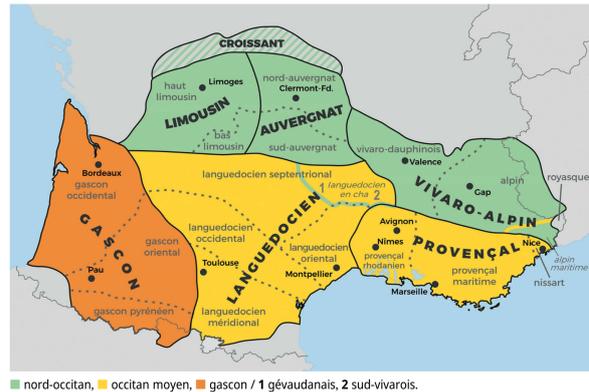


Figure 1: The dialect classification for Occitan proposed in (Sibille, 2024).

as revealed in studies comparing performance between a standard variety (e.g. Standard American English or Modern Standard Arabic) and related non-standardised variants (e.g. African-American English or Moroccan Arabic) (Khondaker et al., 2023; Okpala and Cheng, 2025; Gupta et al., 2025).

In this paper, we focus on Occitan, a dialect continuum spoken mainly in the south of France and supported as a monolithic entity in several NLP models such as pretrained multilingual BERT (Devlin et al., 2019) and the machine translation model NLLB (Costa-jussà et al., 2024). Existing Occitan NLP datasets increasingly cover diverse aspects of internal variation, including dialectal (Miletic et al., 2020), orthographical (Poujade et al., 2024) and other diastratic (Nédey et al., 2025) factors. However, to the best of our knowledge, there are no models available for the identification of the Occitan variety of a text, although such tool could be crucial for building and evaluating dialect-aware NLP systems.

Our contribution is twofold: (1) we introduce a new dialect-labelled dataset OcWikiDialects, and (2) we use it in combination with other datasets to train baseline models for the Occitan dialect identification (OCDI) task. Our dataset is made up

of articles from the Occitan Wikipedia and comprises metadata about dialects and users, aggregated across revisions. Our OCDI experiments involve three types of model: linear SVM based on words and character n -grams, FastText (Joulin et al., 2017) with pretrained vectors, and BERT (Devlin et al., 2019) with pretrained models. By analysing results across three model types, five datasets and up to eight dialects, we demonstrate the impact of the OcWikiDialects dataset and emphasise the need to develop more resources for underrepresented Occitan varieties.

2 Related Work

2.1 Occitan Dialects and Datasets

Occitan is a dialect continuum that spans across the south of France and some border regions of Spain and Italy (see Figure 1). Linguistic studies based on isoglosses and dialectometry experiments led to a classification into dialects within the continuum (Esher and Sibille, 2024) that are sometimes used by speakers when referring to their language. Dialectal variation of Occitan occurs at multiple linguistic levels including phonetic, morphologic, syntactic and lexical. While the so-called and most popular *classical* spelling convention tends to reduce dialect variation in writing, some features remain visible (see some examples from our dataset OcWikiDialects in Table 1), such as the vocalisation of [l] into [w] in final positions (e.g. *aquel* vs. *aqueu* for ‘this’), the presence of the enunciative particle *que*, or the use of definite plural articles *li*, *lu*, or *lei* (vs. *los* or *las*).

Next to large textual datasets using the monolithic label “Occitan” (Schwenk et al., 2021; Miletic and Scherrer, 2022; Costa-jussà et al., 2024; Penedo et al., 2025), smaller Occitan corpora include dialect labels: Tolosa Treebank (Miletic et al., 2020) is a corpus of literary texts from four Occitan varieties, manually annotated with morphosyntactic information. Similarly, CorpusArièja (Poujade et al., 2024) focuses on morphosyntactic annotation of literary texts from a transitional area between the Languedocian and Gascon varieties. The organisation *Lo Congrès permanent de la lenga occitana* compiled bilingual sentences (Occitan and French or other standardised languages) from their websites (Séguier and Lo Congrès, 2023a, 2024) and from various translated tools (Séguier and Lo Congrès, 2023b) into corpora where the document-level dialect label is available for each sentence. The

organisation also released the ReVoc dataset (Lo Congrès, 2024), which contains not only sentences and corresponding dialect labels, but also information about the speakers of the related speech data collection campaign. The inclusion of sociological metadata was also emphasised in Nédey et al. (2025) for the dataset ForumOccitania, which is made of posts from an online forum, accompanied by user-declared information such as dialect, geographical location and age.

Despite the existence of these datasets, the ratio of dialect annotated data remains very low, and the distribution of dialects is very uneven, with a large over-representation of the Languedocian variety while some others are barely represented (e.g. Auvergnat, Vivaroalpine).

2.2 Language Identification for Similar Varieties and Dialects

While the most popular tools for language identification cover a large number of languages (Grave et al., 2018; Kargaran et al., 2023; Costa-jussà et al., 2024), many dialectal variants remain unsupported. This is partly due to their reliance on major language nomenclatures such as ISO 639, which typically lack granular codes for varieties below the level of macro languages, for instance Brazilian and European Portuguese. Works published during VarDial workshops reveal that the identification of similar languages and dialects remains challenging (Aeppli et al., 2023), especially in the context of dialect continuum where established linguistic categories have overlapping and occasionally contested boundaries (Aeppli et al., 2022).

For such varieties, a text can often be valid for several labels, motivating recent work in similar language identification (Keleg and Magdy, 2023; Bernier-colborne et al., 2023; Chifu et al., 2024; Fedorova et al., 2025) to emphasize a shift toward multi-label classification. Although FastText is the most popular architecture to discriminate between many high-resource languages, its usage seems less popular in the context of discriminating between similar languages (Fedorova et al., 2025), where the most common approaches are based on statistical models such as Naïve Bayes and SVM, using n -gram-based features, and on pretrained language models derived from BERT (Devlin et al., 2019). The limited amount of data and lack of diversity (Cahyawijaya et al., 2023) that is common in annotated corpora for these languages make the models (especially statistical ones) more likely to over-rely

on named entities or other semantic aspects of the corpora (e.g. names of cities or languages, topics) instead of linguistic features. Sousa et al. (2025) address this bias by creating a multi-domain corpus for two national varieties of Portuguese and by randomly masking named entities and replacing words with their part of speech tag.

There are no publicly available systems to discriminate between varieties of Occitan, although some works on classification have been carried out: Seguíer (2015) explores statistical approaches based on word and character n -gram frequencies, as well as manually defined grapheme-based and grammatical-based features, resulting in an almost perfect accuracy on 45 samples in 8 Occitan varieties when using only the character n -gram features. More recently, Nédey et al. (2025) carry out unsupervised topic modelling experiments that result in a classifier able to distinguish between 4 Occitan varieties with a macro F1 score of 85.50, when evaluated on 9.5k in-domain test samples.

3 Dataset Creation

We fetch the clean Markdown text of the Occitan Wikipedia¹ articles from the FineWiki dataset² (Penedo, 2025), and we use the XML dump version with complete history³ and texts with wiki markup (Wikitext) to extract more detailed metadata about the dialects used, the article revisions, and the contributors. We retrieve the dialect label of each article from a tag in the Wikitext when it exists, and drop the article from our dataset otherwise. From the revision history, we extract user IDs and timestamps. We also parse user pages to extract declared Occitan language proficiency levels⁴ and dialects, and to mark bot users.

We derive additional metadata from the revision history, such as timestamps at creation and latest update, Occitan level of the first contribution, and highest Occitan level. We use the Wikitext version of each revision to rank users by number and size⁵ of contributions on the article, and to aggregate the number and size of contributions by Occitan level.

¹<https://oc.wikipedia.org>

²<https://huggingface.co/datasets/HuggingFaceFW/finewiki>

³<https://dumps.wikimedia.org/ocwiki/20250901/>. Since FineWiki was built from the HTML dump of 20250820, we parse the history only up to that date.

⁴See scale in Appendix A.

⁵Absolute difference in bytes with the previous contribution, where negative differences are divided by two, with a minimum value of one.

The clean articles are split into paragraphs (based on empty lines) and into sentences (using NLTK (Bird and Loper, 2004)).

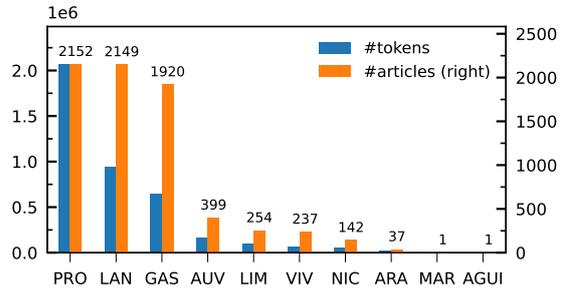


Figure 2: Dialect distribution in OcWikiDialects.

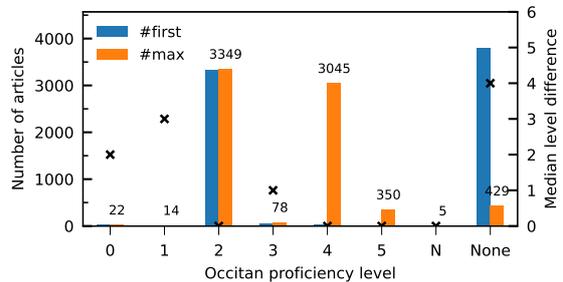


Figure 3: Distribution of self-declared Occitan proficiency levels in OcWikiDialects articles. As #first, we report the number of articles per user-declared Occitan level when considering the first non-empty contribution, and as #max, the number when considering the maximum level over the article’s history. Bar labels correspond to #max.

The resulting OcWikiDialects dataset contains 7,292 articles with a dialect label, comprising approximately 57k paragraphs, 290k sentences, and 4M tokens.⁶ Dialect labels have ten possible values, six of them corresponding to the usual high-level Occitan varieties (Auvergnat (AUV), Gascon (GAS), Limousine (LIM), Languedocian (LAN), Provençal (PRO) and Vivaroalpine (VIV)), two corresponding to local, very distinct varieties (Aranese (ARA), linguistically closer to Gascon, and Niçard (NIC), linguistically closer to Provençal), and the last two corresponding to transitional varieties between Occitan and *langues d’Oil*⁷ (Aguianese (AGUI) and Marchese (MAR)). The dialect distribution presented in Figure 2 reveals that PRO and LAN are the

⁶Counts after splitting on whitespace and removing punctuation (more frequent with the Markdown format). Paragraphs and sentences containing only punctuation are ignored.

⁷A continuum of northern Romance varieties, encompassing standard French and multiple regional varieties.

most frequently assigned labels (ca. 2k articles), closely followed by GAS. Notably, PRO articles are substantially longer on average than those in the other classes, resulting in PRO being the most represented variety in terms of token count, with over 2M tokens. Moreover, even if dialects AUV and VIV are under-represented in OcWikiDialects, Table 2 shows that the number of samples for these dialects surpasses by far that of other datasets, enabling more reliable evaluations on these varieties.

Out of 1249 active users, only 35 declared a dialect in their user page (see Figure 5 in Appendix B.1), with values only for the five most represented dialects in the dataset. Yet, we observe that over 90% of articles labelled as PRO were edited at least once by one of the 8 active users declaring this dialect.

Our analysis of user-declared Occitan levels (see Figure 3) shows that almost all articles were created by users with a level either unknown or intermediate. However, when considering the maximum level across contributions for each article, a clear shift towards high levels is visible, especially as almost 47% of articles were edited by a user of level 4 (near-native ability) or above, whereas the ratio of articles with an unknown level drops from 52% (#first) to 6% (#max).

Additional analyses in Appendix B.2 show that bots are frequent authors of small contributions, suggesting a limited impact on the textual contents of the dataset. Nevertheless, a seemingly important proportion of articles in OcWikiDialects concern municipalities with very similar patterns and templates (see excerpts in Table 1). This lack of topic, style and syntactic diversity could have a negative impact on downstream NLP applications, as observed by Lambrecht et al. (2022) in Machine Translation experiments on dialect-labelled articles of the Alemannic Wikipedia.

4 Occitan Dialect Identification (OCDI) Experiments

4.1 Methodology

We train OCDI model baselines using three complementary modelling approaches, each providing benefits relevant to specific application scenarios.

SVM We combine n -grams of characters (2-5) and full words (unigrams), vectorised with TF-IDF, to train a linear SVM classifier using

scikit-learn (Pedregosa et al., 2018).⁸ The learned features provide the most straightforward basis for interpretability.

FastText We choose this type of model as it usually provides very good results as well as the fastest inference (Fedorova et al., 2025; Suarez et al., 2026). We train a classifier based on FastText embeddings pretrained on Occitan data. For the embeddings, we compare results obtained when using existing vectors trained on Occitan data from Common Crawl and Wikipedia (CC)⁹ (Grave et al., 2018), or when training a new embedding model.

BERT We fine-tune mBERT-cased¹⁰ and mBERT-uncased¹¹ models (Devlin et al., 2019), multilingual Transformer (Vaswani et al., 2017) encoders that were pretrained partly on Occitan data, and oc-mBERT,¹² the result of continued pretraining mBERT-cased specifically on Occitan data (Hopton and Aepli, 2024).

Hyperparameters for each type of model are described in Appendix D.

We evaluate models using accuracy, recall, precision and F1 metrics, for each dialect and macro-averaged per dialect class. As we train and evaluate separately on datasets that contain differing sets of labels (see Table 2), at test time we group predicted labels that are absent from the test set under a single “other” label, to avoid over-penalising these predictions when computing macro-average scores.

To enable comparison in settings where source datasets differ in both class sets and sample distributions, we report the theoretically expected performance of a random-label baseline. This baseline assigns labels uniformly at random to samples, resulting in a fixed per-class recall equal to the inverse of the number of classes, while the precision for each class corresponds to its empirical prevalence in the test set (i.e., the number of samples of that class divided by the test set size).

4.2 Data

We use the following dialect-labelled datasets to train and evaluate the models on the OCDI task:

⁸<https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html>

⁹<https://fasttext.cc/docs/en/crawl-vectors.html>

¹⁰<https://huggingface.co/google-bert/bert-base-multilingual-cased>

¹¹<https://huggingface.co/google-bert/bert-base-multilingual-uncased>

¹²https://huggingface.co/zhopto3/oc_mbert

LAN	Menèrba es una comuna lengadociana situada dins lo departament d’Erau.
PRO	Barcilona es una comuna provençala situada dins lo departament deis Aups d’ Auta Provença.
NIC	Lo Puget Tenier es una comuna d’Occitània, dins lo departament dei Aups Maritims.
VIV	Chastelnòu de Bordeta es una comuna occitana de Daufinat situaa dins lo departament de Droma.
AUV	Mauriac z-es ’na comuna d’Auvèrnhe ; z-es administrada pel departament delh Chantal .
LIM	Lo dolmen situat entre lo vilatge de la Valada e la D 979 fuguet fortament ’bimat en 1862 e completament destruch en 1902.
GAS	Biriato qu’ei ua comuna de la província tradicionau de Labord, administrada peu departament deus Pirenèus Atlantics dens la region de Navèra Aquitània.
ARA	Les ei ua vila e municipi dera Val d’Aran, en eth terçon de Quate Lòcs, ath marge deth riu Garona.

Table 1: Excerpts from OcWikiDialects with some elements of dialectal variation in bold, as compared to LAN. Translations are available in Appendix C.

Dialect	WIKI			TTB			CONGRES			SOFT			FORUM			CONCAT		
	Tr	D	Te	Tr	D	Te	Tr	D	Te	Tr	D	Te	Tr	D	Te	Tr	D	Te
LAN	14575	507	501	615	61	437	9592	500	500	72119	500	500	1410	381	250	98311	1949	2188
GAS	13383	467	500	134	18	103	6821	500	500	103	102	204	1594	242	250	22035	1329	1557
PRO	17985	509	508	45	0	32	59	60	119	666	500	500	704	86	250	19459	1155	1409
LIM	1723	500	503	36	0	41	67	68	135	61	61	122	3619	291	250	5506	920	1051
AUV	2787	512	501	0	0	0	11	12	24	39	39	78	0	0	0	2837	563	603
NIC	127	145	274	0	0	0	0	0	0	861	500	500	0	0	0	988	645	774
VIV	223	223	414	0	0	0	9	10	18	39	40	78	0	0	0	271	273	510
ARA	57	58	122	0	0	0	15	16	32	0	0	0	0	0	0	72	74	154
CONCAT	50860	2921	3323	830	79	613	16574	1166	1328	73888	1742	1982	7327	1000	1000	149479	6908	8246

Table 2: Number of samples in each dataset and split (Tr = train, D = development, Te = test), with distributions per dialect label.

- *OcWikiDialects* (WIKI),¹³ 8 dialect labels. Article-level labels were projected onto paragraph-level samples.
- *Tolosa Treebank*¹⁴ (TTB) (Miletic et al., 2020), 4 dialect labels. Document-level labels were projected onto sentence-level samples.
- *ForumOccitania* (FORUM) (Nédey et al., 2025), 4 dialect labels. User dialects were projected onto post-level anonymised samples.
- CONGRES: the deduplicated concatenation of corpora *Lo Congrès Websites*¹⁵ and *Lo Congrès News*¹⁶, 7 dialect labels.¹⁷ Document-level labels were projected onto sentence-level samples.
- *SoftwaresOccitanTranslations*¹⁸ (SOFT), 7 dialect labels. Document-level labels were projected onto sentence-level samples.

We use the existing train/dev/test splits for TTB and FORUM and create splits for the other corpora.

¹³AGUI and MAR were excluded due to insufficient data. The version of the dataset used for the experiments does not include articles created after 2025-08-01.

¹⁴From UD 2.17. https://universaldependencies.org/treebanks/oc_ttb/index.html

¹⁵<https://zenodo.org/records/12192029>

¹⁶<https://zenodo.org/records/8411197>

¹⁷Cisalpine was excluded due to insufficient data.

¹⁸sic. <https://zenodo.org/records/8411351>

The dev and test sets are built by iterating over dialect subsets without exceeding 500 or 50% of samples for each dialect, in order to produce fairly balanced test sets, in terms of dialects and domains. We use paragraph-level samples for OcWikiDialects, and do not mix articles between splits. We concatenate all sources (WIKI, TTB, FORUM, CONGRES and SOFT) with their splits into CONCAT.

Our FastText embedding model is trained on the concatenation of CONCAT-train and the following datasets: NLLB¹⁹ (Costa-jussà et al., 2024), OcWikiDisc²⁰ (Miletic and Scherrer, 2022), FineWiki²¹ (Penedo, 2025), FineWeb2²² (Penedo et al., 2025), and Tatoeba.²³

In order to assess the importance of preprocessing on the OCDI task, we train our models separately on raw text and on preprocessed texts. The chosen preprocessing steps take into account the presence of User-Generated Content in ForumOccitania. First, we remove URLs and emails, then we convert accents to their ASCII equivalent. Sequences of three or more of the same character

¹⁹<https://huggingface.co/datasets/allenai/nllb>. We use the Occitan samples with a LID score ≥ 0.8 .

²⁰<https://zenodo.org/records/7079580>

²¹<https://huggingface.co/datasets/HuggingFaceFW/finewiki>

²²<https://huggingface.co/datasets/HuggingFaceFW/fineweb-2>

²³<https://opus.nlpl.eu/Tatoeba/oc&fr/v2023-04-12/Tatoeba> (Tiedemann, 2012)

are normalised into a single one. Finally, we remove punctuation and numbers, and we turn the remaining text to lowercase.

4.3 Results and Discussion

Results in Table 3 show that our pretrained FastText vectors increased the overall macro-F1 by 4.48 points. Similarly, our BERT classifier fine-tuned from oc-mBERT performed better than the original mBERT-cased pretrained model, and also slightly better than the mBERT-uncased model (see Table 4).

While preprocessing is usually recommended to prevent overfitting on irrelevant features when training language identification models (Fedorova et al., 2025; Sousa et al., 2025), our experiments on domain-specific and multi-domain data resulted in lower or similar performance when using it, as shown in Table 5. The rest of our analysis will therefore focus only on models trained and evaluated on non-preprocessed datasets, and adapted from the pretrained FastText vectors and BERT model that resulted in the best performance scores.

FastText vectors	F1	Recall	Precision	Accuracy
CC	47.11	45.30	69.25	62.70
Ours	51.59	49.30	70.39	65.95

Table 3: Results of FastText OCDI models depending on the pretrained vectors used. Training set: CONCAT-train. Test set: CONCAT-test. Except for accuracy, scores are macro-averaged over dialect classes.

Pretrained model	F1	Recall	Precision	Accuracy
mBERT-cased	54.51	50.78	75.99	65.69
mBERT-uncased	57.89	53.66	77.64	67.89
oc-mBERT	58.15	54.27	78.49	66.61

Table 4: Performance depending on the pretrained BERT model, fine-tuned on CONCAT-train and tested on CONCAT-test. Except for accuracy, scores are macro-averaged over dialect classes.

Model type	w/ prep	w/o prep
SVM	53.47	55.45
FastText	51.54	51.59
BERT	54.84	58.15

Table 5: F1 scores macro-averaged, comparing approaches with and without text preprocessing, fine-tuned on CONCAT-train and tested on CONCAT-test.

Model type	F1	Recall	Precision	Accuracy	Time
Random baseline	11.32	12.50	12.50	12.50	-
SVM	55.45	52.12	72.32	67.05	3
FastText + our vectors	51.59	49.30	70.39	65.95	0.01
oc-mBERT	58.15	54.27	78.49	66.61	192

Table 6: Performance of models trained on CONCAT-train and evaluated on CONCAT-test, without preprocessing. Recall, precision and F1 scores are macro-averaged. Best score for each column is indicated in **bold**. Time corresponds to the average runtime per sample, expressed in milliseconds, measured on an 13th Gen Intel Core i7-1370P CPU. For oc-mBERT, time is measured on 1000 random samples from CONCAT-test instead of the full test set.

Performance scores in Table 6, obtained from the evaluation on CONCAT-test, report a macro-average F1 score of 11.32 for the random baseline, which is surpassed by all models trained on CONCAT-train. The best F1 score of 58.15 is achieved with the BERT model, which corresponds to an accuracy of 66.61. Performance using the SVM classifier is lower with 55.45 F1, and the lowest using FastText, with 51.59.

While these scores are low, we expect our evaluation methodology to underestimate the systems’ performance, due to the projection of single labels from original documents onto the samples, which hides situations where a same text could be valid in more than one variety, especially for shorter texts.

A more in-depth analysis of results indicates performance disparities between dialects and models (see Table 7). Dialects that are under-represented in the training data (ARA, VIV, NIC) tend to have higher a precision (up to 91.76 for ARA) and a lower recall (only 19.51 for NIC) than bigger classes. As shown in Figure 4, this is due to the over-prediction of the larger classes, especially LAN, which gets a 91.86 recall, and only a precision of 59.10 with oc-mBERT. In usage cases where recall is preferred over precision (e.g. data mining), we recommend using the best-performing model in terms of recall per Occitan variety, rather than a single model across dialects: SVM for AUV, GAS, NIC, and VIV, FastText for GAS and LIM, and oc-mBERT for ARA, LAN and PRO.

The analysis of dialect-specific features ranked by the SVM weights (see Table 8) reveals that some numbers and punctuation are incorrectly over-weighted, and that the model is biased with features related to names of languages (like ‘aran’ for class ARA). However, some important features of each

Model	ARA		AUV		GAS		LAN		LIM		NIC		PRO		VIV	
	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R
Random baseline	1.87	12.50	7.31	12.50	18.88	12.50	26.53	12.50	12.75	12.50	9.39	12.50	17.09	12.50	6.18	12.50
SVM	85.29	18.83	62.30	37.81	76.49	83.82	60.22	89.03	81.83	70.69	88.16	27.91	60.11	67.49	64.12	21.37
FastText	86.36	12.34	66.21	32.17	75.61	83.82	60.49	90.40	79.06	71.84	86.38	23.77	57.18	66.15	51.82	13.92
oc-mBERT	91.76	50.65	79.69	33.83	77.37	82.15	59.10	91.86	76.50	66.60	88.82	19.51	58.22	68.35	96.43	21.18

Table 7: Precision (P) and recall (R) per dialect for each model trained on CONCAT-train and evaluated on CONCAT-test. Best score for each column is indicated in bold.

dialect stand out in the learned features, such as the lexical variant ‘dab’ (*with*) specific to GAS, the non-vocalised ‘l’ at word ending in LAN (as in ‘sul’ *on the*), the aphaeresis of [a] in LIM (e.g. ‘quò’ *this*), the muting of intervocalic [d] in VIV (as in ‘poètz’ *you can*), or the variant ‘lu’ for the masculine plural definite article in NIC. However, we notice that the character n -gram features ‘ou’ and ‘i’ correspond to the so-called *mistralian* writing norm. While this norm seems to be used only for quotes and explanations in the Occitan Wikipedia, their over-weighted association to NIC is problematic as this writing norm is not restricted to only this Occitan variety; in fact it is also used frequently by PRO speakers.

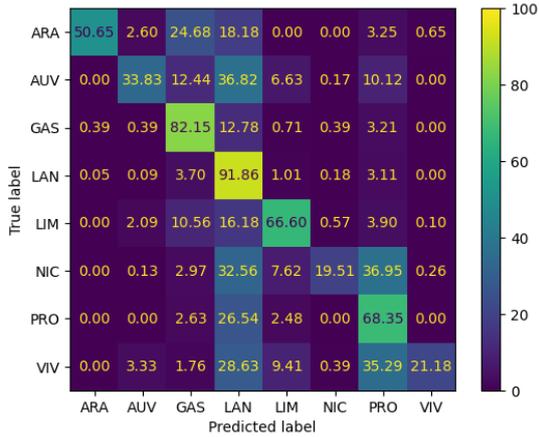
Dialect	Features
ARA	aran determinants referéncias der comandar mès lesan galin telediario substantius
AUV	delh cercar pueis 3,45 z-es 982 dreita 26,43 favier 20,17
GAS	dens dab ei shens mei ua au deu dreita 2019
LAN	sense del sul ∅,∅ l∅ ambe ∅, aude ièr iè
LIM	maïres daus quò queu 'na ente quo quel 833 quela
NIC	ou à∅ lu caracter adressa ligna statut recerca da i∅
PRO	lei deis ò leis dau dei direccions #∅ premiera ∅1
VIV	internes recèrcha poètz aa navigaor pechon predefinia segua mas mesmas

Table 8: Top 10 learned features per dialect for the SVM model without preprocessing. Word features are highlighted in blue, and character n -gram features in green. Word boundaries are indicated with ∅.

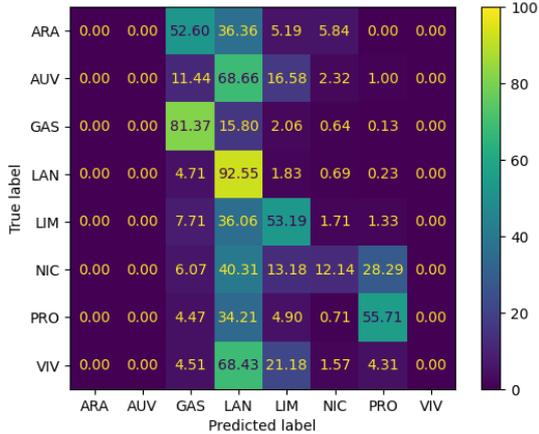
From the results of models trained on a single data source (cf. Table 9), the best macro-F1 scores are obtained by models trained and evaluated on only four classes, i.e. TTB and FORUM. Out of the two, all three models trained on FORUM exhibit larger performance gaps on TTB-test compared to the TTB models evaluated on FORUM-test, suggesting potential overfitting of the FORUM models, even

if scores on TTB-test are within the range of scores with the same-source model. Models trained on CONGRES and SOFT (with similar sets of seven labels) display a low performance on the OCDI task in general, and a clear drop of performance out-of-domain on the test sets with seven or eight labels, with some scores that are close to the random baselines, or even lower for two out of the three models trained on SOFT and evaluated on WIKI. Possible reasons for these drops are the very specific and technical domain of SOFT, a large class imbalance in the training datasets, and for CONGRES, a very low number of samples apart from the LAN and GAS classes. Models trained on WIKI obtain the best results when evaluated in-domain, and often surpass performance of other single-source models in the out-of-domain evaluation settings. Scores on SOFT-test are similar to those obtained with the same-source models, and for CONGRES-test, WIKI models reached a better performance than the same-source ones. Training on the concatenated datasets results in higher scores across test sets, as the CONCAT models obtained the best scores on three test sets (CONCAT, CONGRES and SOFT), and second or third rank otherwise. Also, the macro F1 scores on CONCAT-test are higher for the CONCAT models than for the WIKI ones, showing that augmenting both the quantity of data and the diversity of domains has an impact on the OCDI task.

In order to assess the impact of our dataset OcWikiDialects on the OCDI task, we performed an ablation study comparing models trained on the full concatenation of available training data (CONCAT) with models trained on the same data excluding WIKI-train (CONCATNOWIKI). Evaluation was performed on the CONCAT-test set in both settings. As shown in Table 10, the inclusion of OcWikiDialects results in substantial performance gains, with improvements of up to 23.48 macro F1 points for the oc-mBERT model. This impact is particularly pronounced for the rarer classes, as shown when comparing the confusion matrices in Figure 4: while the oc-mBERT model trained with-



(a) CONCAT-train



(b) CONCATNOWIKI-train

Figure 4: Confusion matrices for the oc-mBERT models fine-tuned from CONCAT-train (top) or CONCATNOWIKI-train (down), without preprocessing, and evaluated on CONCAT-test.

out OcWikiDialects never predicts the rarer classes ARA, AUV and VIV, incorporating the WIKI samples to the training leads to more frequent predictions of these labels, and results in higher recall across all classes except LAN, while also reducing over-prediction of the LAN label (i.e. improving precision).

5 Conclusion

In this paper, we introduce a new corpus OcWikiDialects, comprising over 7k articles (4M tokens) from the Occitan Wikipedia, segmented into paragraphs and sentences. In doing so, we contribute to increasing the amount of dialect-labelled Occitan data, necessary to increase the visibility of Occitan variation as a non-standard language in datasets for NLP. Metadata from revision histories and user pages included in the dataset enabled us to label

Trainset and model	Testset					
	CONCAT	WIKI	CONGRES	SOFT	TTB	FORUM
Random baseline	11.32	12.07	10.59	12.52	19.08	25.00
CONCAT-train						
SVM	55.45	56.15	42.74	42.33	60.57	68.98
FastText	51.59	51.27	39.27	39.45	62.42	67.85
oc-mBERT	58.15	57.72	59.15	43.91	56.79	70.22
WIKI-train						
SVM	53.61	60.21	44.81	31.99	44.16	57.61
FastText	49.23	49.86	43.01	34.71	57.67	66.06
oc-mBERT	48.12	53.18	50.45	31.67	45.99	52.37
CONGRES-train						
SVM	20.45	15.98	38.33	14.21	33.29	36.77
FastText	24.01	16.63	37.46	15.60	37.51	43.14
oc-mBERT	20.71	18.48	33.24	16.66	37.66	48.32
SOFT-train						
SVM	20.57	11.52	15.00	39.35	25.60	20.02
FastText	19.42	9.06	19.86	30.18	38.52	38.41
oc-mBERT	24.44	14.21	26.16	31.17	42.22	45.70
TTB-train						
SVM	20.92	17.56	28.27	17.48	66.66	47.36
FastText	26.00	19.71	32.86	25.23	73.19	68.85
oc-mBERT	15.48	11.23	22.73	14.20	46.38	36.36
FORUM-train						
SVM	29.11	26.90	36.36	22.21	53.82	88.93
FastText	32.66	28.91	41.63	31.35	71.38	89.01
oc-mBERT	29.22	26.11	37.29	22.14	63.95	86.62

Table 9: F1 scores on individual test sets for each model based on its training dataset and type. Best score for each test set is marked in bold. We warn the reader about the non-comparability of scores between most columns, as the number of labels differs between the test sets.

	\bar{W}	$C - \bar{W}$			
	F1	F1	Recall	Precision	Acc.
SVM	36.94	+18.51	+15.52	+16.46	+11.70
FastText	35.70	+15.89	+13.20	+16.63	+10.38
oc-mBERT	34.67	+23.48	+17.39	+17.92	+9.25

Table 10: Performance gaps between models trained on CONCAT-train (C) or on CONCATNOWIKI-train (\bar{W}), and evaluated on CONCAT-test. Except for accuracy (Acc.), scores are macro-averaged over dialect classes.

articles with ten Occitan varieties and allowed us to carry out in-depth analyses of user-declared proficiency levels and bot contributions.

We used this dataset in combination with previously published dialect-labelled corpora to train Occitan dialect identification (OCDI) models in single-domain and multi-domain settings to discriminate between up to 8 dialects. Our approaches include n -gram based SVM classifiers, FastText classifiers based on pretrained vectors, and fine-tuned BERT models. Results favour the fine-tuned oc-mBERT model without text preprocessing, with a macro-average F1 score of 58.15 and fairly balanced scores across domains. However, performance differs between dialect classes, with low recall scores for less represented dialects, and low precision scores for the over-represented classes

which tend to be over-predicted, especially Languedocian. Future work should therefore prioritise the development of additional resources for underrepresented Occitan varieties.

We release models from all three main approaches (SVM, FastText, BERT), as they have different advantages, in particular interpretation of predictions, inference speed and overall accuracy, and we hope that our new dataset and OCDI models will foster more dialect-aware NLP research and applications for the Occitan language.

Limitations

The models presented in this paper are meant as baselines for future research on the topic of Occitan Dialect Identification, with several limitations that may restrict their suitability in certain settings, especially when a high confidence is required.

Beside a limited performance described in this paper, in particular for underrepresented Occitan varieties (Section 4.3), we did not train our models to predict a class ‘other’ when prompted with texts from other varieties, closely related to Occitan (e.g. Croissant transitional dialects) or not (e.g. French or Italian).

Furthermore, the datasets used in this study only contain texts in the classical spelling, except for the Niçard subset of SoftwaresOccitanTranslations which contains both samples in classical and mistralian norms. Therefore, using our models on texts written with the mistralian norm might lead to biased predictions towards the Niçard label.

Assessing performance in a multi-domain setting enhances the reliability of results, however the test samples have been selected automatically from the available datasets without removing irrelevant samples such as bibliography entries with titles in languages other than Occitan. Manual annotation of the test samples by Occitan specialists could be particularly relevant to improve the quality of test sets, and also to move towards a multi-label classification task where the samples might be valid in multiple varieties.

Acknowledgments

The authors would like to thank Zachary Hopton for releasing the oc-mBERT model upon our request, the reviewers for their valuable feedback, and the CLEPS infrastructure from Inria Paris for providing computational resources. This work was partly funded by Rachel Bawden and Benoît

Sagot’s chairs in the PRAIRIE institute, funded by the French national agency ANR, as part of the “Investissements d’avenir” programme under the reference ANR-19-P3IA-0001 and by Benoît Sagot’s chair in its follow-up, PRAIRIE-PSAI, also funded by the ANR as part of the “France 2030” strategy under the reference ANR23-IACL-0008.

References

- Noëmi Aepli, Antonios Anastasopoulos, Adrian-Gabriel Chifu, William Domingues, Fahim Faisal, Mihaela Gaman, Radu Tudor Ionescu, and Yves Scherrer. 2022. [Findings of the VarDial evaluation campaign 2022](#). In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–13, Gyeongju, Republic of Korea. Association for Computational Linguistics.
- Noëmi Aepli, Çağrı Çöltekin, Rob Van Der Goot, Tommi Jauhiainen, Mourhaf Kazzaz, Nikola Ljubešić, Kai North, Barbara Plank, Yves Scherrer, and Marcos Zampieri. 2023. [Findings of the VarDial evaluation campaign 2023](#). In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 251–261, Dubrovnik, Croatia. Association for Computational Linguistics.
- Gabriel Bernier-colborne, Cyril Goutte, and Serge Leger. 2023. [Dialect and variant identification as a multi-label classification task: A proposal based on near-duplicate analysis](#). In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 142–151, Dubrovnik, Croatia. Association for Computational Linguistics.
- Steven Bird. 2022. [Local languages, third spaces, and other high-resource scenarios](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7817–7829, Dublin, Ireland. Association for Computational Linguistics.
- Steven Bird and Edward Loper. 2004. [NLTK: The natural language toolkit](#). In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.
- Samuel Cahyawijaya, Holy Lovenia, Fajri Koto, Dea Adhista, Emmanuel Dave, Sarah Oktavianti, Salsabil Akbar, Jhonson Lee, Nuur Shadieq, Tjeng Wawan Cenggoro, Hanung Linuwih, Bryan Wilie, Galih Muridan, Genta Winata, David Moeljadi, Alham Fikri Aji, Ayu Purwarianti, and Pascale Fung. 2023. [NusaWrites: Constructing high-quality corpora for underrepresented and extremely low-resource languages](#). In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 921–945,

- Nusa Dua, Bali. Association for Computational Linguistics.
- Adrian-Gabriel Chifu, Goran Glavaš, Radu Tudor Ionescu, Nikola Ljubešić, Aleksandra Miletic, Filip Miletic, Yves Scherrer, and Ivan Vulić. 2024. [VarDial evaluation campaign 2024: Commonsense reasoning in dialects and multi-label similar language identification](#). In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 1–15, Mexico City, Mexico. Association for Computational Linguistics.
- Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Mailard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, and 20 others. 2024. [Scaling neural machine translation to 200 languages](#). *Nature*, 630(8018):841–846. Publisher: Nature Publishing Group.
- 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.
- Louise Esher and Jean Sibille, editors. 2024. *Manuel de linguistique occitane*. De Gruyter.
- Mariia Fedorova, Jonas Sebulon Frydenberg, Victoria Handford, Victoria Ovedie Chruickshank Langø, Solveig Helene Willoch, Marthe Løken Midtgaard, Yves Scherrer, Petter Mæhlum, and David Samuel. 2025. [Multi-label Scandinavian language identification \(SLIDE\)](#). In *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025)*, pages 179–189, Tallinn, Estonia. University of Tartu Library, Estonia.
- Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. [Learning word vectors for 157 languages](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Abhay Gupta, Jacob Cheung, Philip Meng, Shayan Sayyed, Kevin Zhu, Austen Liao, and Sean O’Brien. 2025. [EnDive: A cross-dialect benchmark for fairness and performance in large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 16830–16855, Suzhou, China. Association for Computational Linguistics.
- Zachary Hopton and Noëmi Aepli. 2024. [Modeling Orthographic Variation in Occitan’s Dialects](#). In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 78–88, Mexico City, Mexico. Association for Computational Linguistics.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. [Bag of Tricks for Efficient Text Classification](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.
- Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. [GlotLID: Language identification for low-resource languages](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218, Singapore. Association for Computational Linguistics.
- Amr Keleg and Walid Magdy. 2023. [Arabic dialect identification under scrutiny: Limitations of single-label classification](#). In *Proceedings of ArabicNLP 2023*, pages 385–398, Singapore (Hybrid). Association for Computational Linguistics.
- Md Tawkat Islam Khondaker, Abdul Waheed, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. [GPTAraEval: A comprehensive evaluation of ChatGPT on Arabic NLP](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 220–247, Singapore. Association for Computational Linguistics.
- Louisa Lambrecht, Felix Schneider, and Alexander Waibel. 2022. [Machine translation from Standard German to alemannic dialects](#). In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 129–136, Marseille, France. European Language Resources Association.
- Lo Congrès. 2024. [ReVoc Corpus](#).
- Aleksandra Miletic, Myriam Bras, Marianne Vergez-Couret, Louise Esher, Clamença Poujade, and Jean Sibille. 2020. [A four-dialect treebank for Occitan: Building process and parsing experiments](#). In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 140–149, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).
- Aleksandra Miletic and Yves Scherrer. 2022. [OcWikiDisc: a corpus of Wikipedia talk pages in Occitan](#). In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 70–79, Gyeongju, Republic of Korea. Association for Computational Linguistics.
- Oriane Nédey, Juliette Janès, Rachel Bawden, Thibault Clérice, and Benoît Sagot. 2025. [ForumOccitania: a Corpus of User-Generated Content for Multiple Occitan Varieties](#).

- Ebuka Okpala and Long Cheng. 2025. [Large Language Model Annotation Bias in Hate Speech Detection](#). *Proceedings of the International AAAI Conference on Web and Social Media*, 19:1389–1418.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2018. [Scikit-learn: Machine Learning in Python](#). *arXiv preprint*. ArXiv:1201.0490 [cs].
- Guilherme Penedo. 2025. [FineWiki](#).
- Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. [FineWeb2: One Pipeline to Scale Them All — Adapting Pre-Training Data Processing to Every Language](#).
- David Pomeranke, Jonas Nothnagel, and Simon Ostermann. 2025. [The AI Language Proficiency Monitor – Tracking the Progress of LLMs on Multilingual Benchmarks](#). *arXiv preprint*. ArXiv:2507.08538 [cs] version: 1.
- Clamenca Poujade, Myriam Bras, and Assaf Urieli. 2024. [CorpusArièja: Building an annotated corpus with variation in Occitan](#). In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 66–71, Torino, Italia. ELRA and ICCL.
- Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. [WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.
- Eve Segulier. 2015. [Reconnaissance automatique des dialectes occitans à l’écrit](#). Master’s thesis, Université Toulouse Jean Jaurès, Toulouse, France.
- Jean Sibille. 2024. [16 Les dialectes occitans](#). In Louise Esher and Jean Sibille, editors, *Manuel de linguistique occitane*, pages 423–471. De Gruyter.
- Hugo Sousa, Rúben Almeida, Purificação Silvano, Inês Cantante, Ricardo Campos, and Alípio Jorge. 2025. [Enhancing portuguese variety identification with cross-domain approaches](#). In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, volume 39 of AAAI’25/IAAI’25/EAAI’25, pages 25192–25200. AAAI Press.
- Pedro Ortiz Suarez, Laurie Burchell, Catherine Arnett, Rafael Mosquera-Gómez, Sara Hincapie-Monsalve, Thom Vaughan, Damian Stewart, Malte Ostendorff, Idris Abdulmumin, Vukosi Marivate, Shamsuddeen Hassan Muhammad, Atnafu Lambebo Tonja, Hend Al-Khalifa, Nadia Ghezaiel Hamouda, Verrah Otiende, Tack Hwa Wong, Jakhongir Saydaliev, Melika Nobakhtian, Muhammad Ravi Shulthan Habibi, and 78 others. 2026. [CommonLID: Re-evaluating State-of-the-Art Language Identification Performance on Web Data](#). *arXiv preprint*. ArXiv:2601.18026 [cs].
- Aure Séguier and Lo Congrès. 2023a. [Occitan Corpus from Lo Congrès news](#).
- Aure Séguier and Lo Congrès. 2023b. [SoftwaresOccitanTranslations corpus](#).
- Aure Séguier and Lo Congrès. 2024. [Lo Congrès websites Corpus](#).
- Jörg Tiedemann. 2012. [Parallel data, tools and interfaces in OPUS](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is All you Need](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

A Scale of Occitan Proficiency Levels

The Occitan language proficiency levels used in OcWikiDialects correspond to the Babel scale defined by Wikimedia Commons.²⁴, with the following definitions to which we add level ‘None’:

- ‘None’ means that no Occitan level was found in the user page
- 0 indicates someone who does not understand the language.
- 1 stands for basic knowledge: the ability to understand and answer simple questions in the language.
- 2 stands for intermediate knowledge.
- 3 stands for advanced or fluent knowledge: the ability to correct spelling and grammar errors in the language.
- 4 stands for near-native ability.
- 5 stands for professional proficiency.
- N stands for native language.

²⁴<https://commons.wikimedia.org/wiki/Commons:Babel> Consulted on 2025-12-30.

B Additional Statistics from Dataset OcWikiDialects

B.1 Dialects Distribution

Figure 5 shows the distribution of dialect tags declared in articles and in user pages.

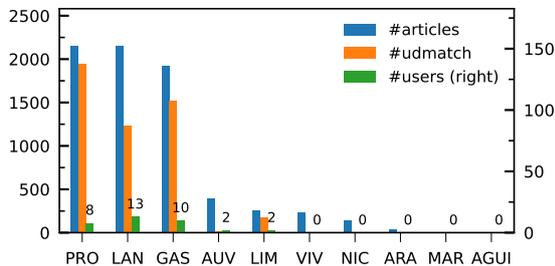


Figure 5: Distribution of article dialect tags (#articles) and user-declared dialects (#users, only active users) in OcWikiDialects, and number of articles per dialect for which at least one contribution was made by a user of the same dialect (#udmatch).

B.2 Bot Contributions

Regarding the role of the 135 active users detected as bots, they have been used to create 13% of the articles in the dataset and 41% of the total contributions. However, this ratio drops to 14% when considering only contributions with at least +100 bytes (based on the difference before and after), and to 8% for the first non-empty contributions²⁵ authored by a bot.

C Excerpts from OcWikiDialects

Table 11 shows a few excerpts from the dataset OcWikiDialects with their translations into English.

D Hyperparameters

For SVM models, the impact of class imbalance is reduced by setting the `class_weight` option to “balanced”. Features occurring in only one sample (`min_df`) or in more than 90% of the training data (`max_df`) are ignored.

For FastText, our embedding skipgram model has the same dimension of 300 than the CC vectors, and is trained for 5 epochs with a learning rate of 0.05. The classification layer is trained for 20 epochs with a hierarchical softmax loss and a learning rate of 0.1. In preliminary experiments we

²⁵For each article, we consider the earliest revision with at least 100 bytes.

found that using word unigrams for the OCDI task results in performance that is better than or similar to using word bigrams.

For BERT models, we use a learning rate of $1e-5$ with linear scheduling and AdamW optimiser, an effective batch size of 64. Each model is trained for 50 epochs with 0.2% of the total steps (i.e. 1% of the training set) used for warmup and early stopping with a patience of 10.

LAN	Menèrba (Minerve en francés) es una comuna lengadociana situada dins lo departament d'Erau e la region d'Occitània, ancianament de Lengadòc-Rosselhon. <i>Menèrba (Minerve in French) is a Languedocian municipality located in the Hérault departement and the Occitania region, formerly Languedoc-Roussillon.</i>
PRO	Barcilona o benlèu Barciloneta (Barcelonnette en francés) es una comuna provençala situada dins lo departament deis Aups d'Auta Provença e la region de Provença-Aups-Còsta d'Azur. <i>Barcilona or maybe Barciloneta (Barcelonnette in French) is a Provence municipality located in the Alps of Upper Provence departement and Provence-Alpes-Côte d'Azur region.</i>
NIC	Lo Puget Tenier (Lou Puget-Teniés en nòrma mistralenca; Puget-Théniers en francés) es una comuna d'Occitània, dins lo País Niçard e lo departament dei Aups Maritims. <i>Lo Puget Tenier (Lou Puget-Teniés in the mistralian norm; Puget-Théniers in French) is a municipality of Occitania, in the County of Nice and the departement of Maritime Alps.</i>
VIV	Chastelnòu de Bordeta (Châteauneuf-de-Bordette en francés) es una comuna occitana de Daufinat situaa dins lo departament de Droma e la region d'Auvèrnhe-Ròse-Aups, ancianament de Ròse-Aups. <i>Chastelnòu de Bordeta (Châteauneuf-de-Bordette in French) is an Occitan municipality of Dauphiny located in the Drôme departement and the Auvergne-Rhône-Alpes region, formerly Rhône-Alpes.</i>
AUV	Mauriac (Mauriac en francés) z-es 'na comuna d'Auvèrnhe ; z-es administrada pel departament delh Chantal de la region d'Auvèrnhe-Ròse-Aups, ancianament d'Auvèrnhe. <i>Mauriac (Mauriac in French) is an Auvergnat municipality ; it is administrated by the Cantal departement of Auvergne-Rhône-Alpes region, formerly Auvergne.</i>
LIM	Lo territòri de la comuna de Sent Deunis fuguet 'bitat desjà la Preistòria. Ne'n tesmonha lo dolmen situat entre lo vilatge de la Valada e la D 979, qui fuguet fortament 'bimat en 1862 e completament destruch en 1902. <i>The territory of Sent Deunis municipality was already inhabited during Prehistory. This is testified by the dolmen located between the village of La Valada and the D 979 [road], which was strongly damaged in 1862 and completely destroyed in 1902.</i>
GAS	Biriato (Biriātu en basco, Biriātu en francés) qu'ei ua comuna de la província tradicionau de Labord, administrada peu departament deus Pirenèus Atlantics dens la region de Navèra Aquitània, ancianament d'Aquitània. <i>Biriato (Biriātu in Basque, Biriātu in French) is a municipality of the traditional province of Labord, administrated by the Pyrénées Atlantiques departement in New Aquitaine region, formerly Aquitaine.</i>
ARA	Les (en catalan: Lés) ei ua vila e municipi dera Val d'Aran, en eth terçon de Quate Lòcs, ath marge deth riu Garona. <i>Les (in Catalan: Lés) is a city and municipality of the Val d'Aran, in the district of Quate Lòcs, along the Garonne river.</i>

Table 11: Excerpts from OcWikiDialects with their translation into English.

Language Mixture to Develop Accurate Galician Dependency Parsers: An Exploration of Its Effects

Xabier Irastortza-Urbieta¹, José M. García-Miguel², Marcos Garcia³

¹ HiTZ Center - Ixa, University of the Basque Country

² iLingua, Universidade de Vigo

³ CiTIUS, Universidade de Santiago de Compostela

xabier.irastorza@ehu.eus, gallego@uvigo.gal, marcos.garcia.gonzalez@usc.gal

Abstract

The development of accurate syntactic parsers remains a challenge for low-resource languages. To overcome it, the literature has proposed leveraging syntactic annotations from typologically related languages. This work investigates the viability and adequacy of this approach for Galician, evaluating the use of annotations from major Romance languages as source data. Our methodology extends beyond standard automatic evaluation to incorporate a detailed error analysis, which precisely quantifies the effects of multilingual training and assesses the practical scalability of the method. The results establish the necessity of embedding models for effective cross-lingual transfer and demonstrate that even languages not particularly close can yield adequate parsers. This work confirms the benefits of cross-lingual data augmentation while delineating its scalability limits. Furthermore, the error analysis identifies specific, typologically conditioned grammatical dependencies that remain persistent challenges for accurate dependency parsing.

1 Introduction

Galician is a Romance language spoken in the southwestern part of the Iberian Peninsula, with approximately 2.4 million speakers. It is the official language of the region of Galicia, alongside Spanish. Like many other minoritized languages, it faces significant challenges in developing language technologies, primarily due to a scarcity of data and resources. This scarcity is a particularly critical challenge in the field of syntax parsing, the focus of this work, as it requires costly, expert human annotation to develop high-quality systems.

To address this data scarcity, we focus on Galician’s membership in the Romance language family. Other major languages in this family, such as Portuguese, French, and Spanish, benefit from abundant human-annotated data for parser development (Nivre et al., 2020). Building on this, our

work investigates methods for combining data from these languages with Galician data to leverage their syntactic similarities and, consequently, develop more accurate syntactic parsers.

1.1 Goal

This work investigates the effects of combining syntactic annotations from different languages, focusing on Galician. We aim to identify the most effective source languages for Galician and analyze the specific parsing errors that are introduced or resolved by incorporating data from other languages. To guide this investigation, we base our work on two research questions:

1. RQ1: Typological Proximity. What is the minimum degree of typological proximity required between Galician and a source language to improve syntactic parsing performance?
2. RQ2: Error Analysis. What are the most prominent syntactic dependency errors in the resulting parsers? To what extent can these errors be attributed to (a) inherent treebank structure, (b) specific language combinations, or (c) the use of cross-lingual embedding models? Furthermore, do these factors primarily exacerbate or mitigate specific error types?

Ultimately, this work aims to propose guidelines for constructing more accurate Galician syntactic parsers. Although focused on a single language, the core approach—leveraging typological proximity within a language family—provides a reproducible framework for improving parsing systems for other low-resource languages (Dione, 2021).

This paper proceeds as follows. We first survey related work in Section 2. Section 3 describes our methodology, the results of which are discussed in Section 4. We present the final conclusions in Section 5 and the limitations in Section 6.

Language	Nb. Tr.	Nb. Sent.	UAS/LAS
English	13	97,896	96.8 / 94.4
Portuguese	7	82,183	96.5 / 95.2
Italian	11	42,297	95.8 / 93.8
Basque	1	8,993	88.2 / 85
Galician	3	5,993	86.8 / 82.3
Welsh	1	2,717	88.4 / 82.6

Table 1: Syntax parsing resources of languages. Respectively, the number of treebanks in UD, the sum of the total number of sentences of them and the highest UAS and LAS obtained by each language with the default UDPipe 2 models (version of 2025).

2 Related Work

Linguistically, the majority of modern syntactic parsers are based on dependency parsing. The Universal Dependencies (UD) project, a large-scale collection of treebanks, has been instrumental in the widespread adoption of this framework. UD provides treebanks for over 178 languages and has significantly contributed to the standardization of annotation guidelines, ensuring cross-linguistic consistency (de Marneffe et al., 2014).

Grounded in Universal Dependencies, several dependency parsers have been successfully developed using machine learning techniques, including UDPipe (Straka and Straková, 2017) and Stanza (Qi et al., 2020). Subsequently, the rise of multilingual parsers like UDify (Kondratyuk and Straka, 2019) and the creation of algorithms such as TOWER (Glavaš and Vulić, 2021) to combine annotations from different languages were important milestones in the field’s development. Furthermore, some authors have proposed using embeddings as input to parsers instead of raw text, demonstrating the positive effects of this approach (Adelmann et al., 2021).

The Large Language Model revolution in recent years has also influenced the field; some authors have employed them to create synthetic training data (Zhang et al., 2024), while others have used them directly as parsers (Ezquerro et al., 2025). Some research has suggested that LLMs are capable of achieving state-of-the-art performance in dependency parsing, at least in some languages (Hromei et al., 2024). However, it is important to note that some authors have found significant performance gaps of LLMs between high- and low-resource languages in common NLP tasks such as translation (Court and Elsner, 2024) or annotation

Parser	UAS	LAS	Innovations
UDPipe2	78.66	74.25	Graph-based neural architecture
UDify	84.08	76.77	Multilingual parser
Stanza	77.27	71.86	Bi-LSTM parser
Tower	77.57	66.87	Hierarchical clustering
Sarym.	89.03	85.30	Mixture of languages, embeddings

Table 2: Performance of various models in Galician. The rows follow a chronological order. Scores were obtained by parsing the TreeGal treebank.

(Jadhav et al., 2025), which may also be the case for dependency parsing in low-resource languages.

With regard to low-resource languages, the lack of manual syntactic annotations has traditionally hindered the training of accurate parsers (Taghizadeh and Faili, 2022). To address this problem, previous work has proposed using annotations from other languages by means of different types of transfer learning (Ghiffari et al., 2023). Table 1 compares the available resources for each language in Universal Dependencies. Note that the major languages have a considerably larger number of annotated sentences; thus, their parsers achieve considerably higher scores on standard metrics, indicating better performance.

Universal Dependencies has released three Galician treebanks, two of which consist of manually revised annotations: TreeGal (Garcia, 2016) and PUD-gl (Sánchez-Rodríguez et al., 2024). It is worth mentioning the treebank PUD-gl, which is the Galician version of a parallel corpus available in several languages; it was first published for the CoNLL Shared Task of 2017 (Zeman et al., 2017).

Regarding dependency parsers for Galician, the earliest systems were rule-based (Gamallo and González, 2012). Subsequently, Galician was included among the languages supported by UDPipe 2 (Straka, 2018), and further research has been conducted to improve these parsers—most notably the study by Sarymsakova et al. (2024). In that study, the authors explored the use of annotations from Portuguese and Spanish for data augmentation, as well as the use of embeddings from several language models, to develop more accurate parsers for Galician. The present work takes that study as its point of departure.

3 Methods

This work uses UDPipe 2 (Straka, 2018) as the basis for creating syntactic parsers, following the approach of several previous studies (Lopes and Pardo, 2024; Sarymsakova et al., 2024). In all experiments, we provided the models with gold-standard tokenization. All treebanks used for training were sourced from Universal Dependencies. For the Galician treebanks PUD-gl and TreeGal, we split the sentences into three sets: from each treebank, 800 sentences were assigned to the training set, 50 to the development set, and 150 to the test set. All other treebanks used in this work were employed exclusively as training data. We excluded the Galician treebank *Corpus Técnico do Galego*, CTG (Guinovart, 2017), because it contains semi-automatically annotated data, unlike the manually annotated treebanks used in this work.

In parallel, across all experiments we investigate the impact of using embeddings as input to the parsers. To this end, we generated embeddings using three language models: Bertinho (Vilares Calvo et al., 2021), BertGL (Garcia, 2021), and XML-Roberta (Conneau et al., 2020). In all cases we have used the *Base* version of those models. The first two are monolingual models trained on Galician, while the latter is a multilingual model. All models were sourced from HuggingFace, and embeddings were calculated using the Wembedding Service tool provided by UDPipe 2.

Evaluation was conducted consistently across all experiments. First, we assessed the general performance of the parsers using two standard automatic metrics: Unlabeled Attachment Score or UAS (Eisner, 1996) and Labeled Attachment Score or LAS (Nivre et al., 2004). We placed greater emphasis on the LAS metric due to its closer alignment with human judgments, as established in prior work (Plank et al., 2015). Second, we performed a fine-grained error analysis, disaggregating results by grammatical dependency types to measure parser precision for each specific relation. This approach allowed us to describe in detail the impact of cross-lingual data mixtures on parser performance and to pave the way for future qualitative analyses. The following subsections describe each experiment.

3.1 Impact of Typological Proximity

In this experiment, we trained parsers using separate, monolingual datasets from Galician and other languages, ensuring an equal number of sentences

per language without mixing data across languages within a single training run. Our goal was to confirm that using data from languages typologically close to Galician—even in the absence of any Galician training sentences—can produce effective parsers, which should perform significantly better than parsers trained on data from typologically distant languages. To this end, we selected five high-resource languages representing a gradient of typological proximity to Galician: Portuguese and Spanish (both Ibero-Romance languages, the subgroup to which Galician belongs)¹; French and Italian (Romance languages outside the Ibero-Romance subgroup); and German (a Germanic, non-Romance language). Thus, the first pair is phylogenetically closest to Galician, while German is the most distant.

To conduct this experiment, we used the Parallel Universal Dependencies (PUD) treebanks for the five languages mentioned above. We selected 800 sentences from each language’s PUD treebank and trained all models exclusively on this data. For each language, we trained four models: three models used the embeddings generated by each of the three language models (Bertinho, BertGL, and XML-Roberta) as input, and one baseline model used no embeddings. All models were trained on monolingual data (i.e., each model used data from only one source language).

In line with the above, we created six distinct training data partitions, each corresponding to a single source language: G (Galician), P (Portuguese), S (Spanish), F (French), I (Italian), and D (German). These codes will be used in the following sections to reference the corresponding partitions.

3.2 Comparison Between Treebanks

In this experiment, we compared parsers trained on the two manually annotated Galician treebanks, PUD-gl and TreeGal. The aim was to quantify the impact of differences in annotation guidelines and minor formatting conventions between treebanks, even when they represent the same language. To this end, we constructed two separate training sets of 800 sentences each, one from each treebank. For each set, we trained four parser variants: three using embeddings from the respective language models (Bertinho, BertGL, XML-Roberta) and one

¹Galician and Portuguese are traditionally considered varieties of the same language with different orthographies, while Galician and Spanish, despite intense contact and a shared writing system, are distinct languages (Carvalho Calero, 1985).

baseline model without embeddings. Finally, we evaluated all trained parsers using the corresponding 150-sentence test sets defined for each treebank in the previous section.

3.3 Mixture of Languages

For this experiment, we created training partitions by combining manual annotations from different treebanks, thereby mixing languages. The purpose was to study the parsing errors that are either introduced or resolved as a direct result of this cross-lingual data combination. Specifically, we created four training partitions based on the PUD-gl Galician treebank. All partitions contained the core set of 800 training sentences from PUD-gl. To this core, we added sentences from other languages to form the following four mixed-data conditions:

1. G+P: PUD-gl + Portuguese (800 Galician + 800 Portuguese sentences).
2. G+S: PUD-gl + Spanish (800 Galician + 800 Spanish sentences).
3. G+P+S: PUD-gl + Portuguese + Spanish (800 Galician + 400 Portuguese + 400 Spanish sentences).
4. G+I+F: PUD-gl + Italian + French (800 Galician + 400 Italian + 400 French sentences).

All partitions consisted of 1,600 sentences each. All trained parsers were evaluated on the 150-sentence PUD-gl test set defined earlier. As in previous experiments, we tested the impact of different embedding types (Bertinho, BertGL, XML-Roberta, and no embeddings).

3.4 Scalability

As a final scalability experiment, we investigated whether large, high-resource treebanks from closely related languages could be used directly—without any pre- or post-processing—to develop more precise parsers for Galician, thereby addressing its data scarcity. For this purpose, we incorporated two additional large treebanks: CINTIL (Portuguese) and AnCora (Spanish).

We constructed three training partitions. Each partition contained the combined training sentences from both Galician treebanks (800 from PUD-gl + 800 from TreeGal = 1,600 sentences). To this Galician base, we added sentences from the large external treebanks to create the following conditions:

1. G+PP: Galician + CINTIL: 1,600 Galician + 2,400 Portuguese (CINTIL) sentences.
2. G+SS: Galician + AnCora: 1,600 Galician + 2,400 Spanish (AnCora) sentences.
3. G+PPSS: Galician + CINTIL + AnCora: 1,600 Galician + 1,200 Portuguese (CINTIL) + 1,200 Spanish (AnCora) sentences.

For each of the three data combinations, we trained four parser variants (using the three embedding types—Bertinho, BertGL, XML-Roberta—and one baseline without embeddings). All models were evaluated on the standard PUD-gl test set of 150 sentences.

4 Results

In this section we will present the results of the experiments described in Section 3, followed by a discussion of them.

4.1 Typological Proximity

UAS-LAS analysis. Figure 1 shows the UAS and LAS scores achieved by parsers trained on different language versions of the PUD treebank, all evaluated on the Galician PUD-gl test set. The parsers trained on Portuguese and Spanish data obtain UAS scores very close to the parser trained on Galician (within 2 and 4 points, respectively), though their LAS scores are somewhat lower (5 and 12 points difference). This proximity in performance is only observed when using embeddings; without embeddings, the performance of parsers trained on Portuguese (P) and Spanish (S) is significantly more limited. For these two closely related languages, the choice between monolingual and multilingual embedding models has a minimal effect.

In contrast, for more typologically distant languages—Italian (I), French (F), and German (D)—the type of embedding model matters substantially. When using the multilingual embedding model, parsers trained on Italian and French achieve UAS scores only 7 points below the Galician-trained parser, compared to a 30–40 point drop for parsers trained without embeddings. However, performance on the stricter LAS metric remains more limited for all parsers trained on languages other than Galician.

Dependency-level analysis. Figure 2 visualizes the error analysis for the trained parsers. Each row corresponds to one of the 14 most frequent dependency relations in the PUD treebank, and

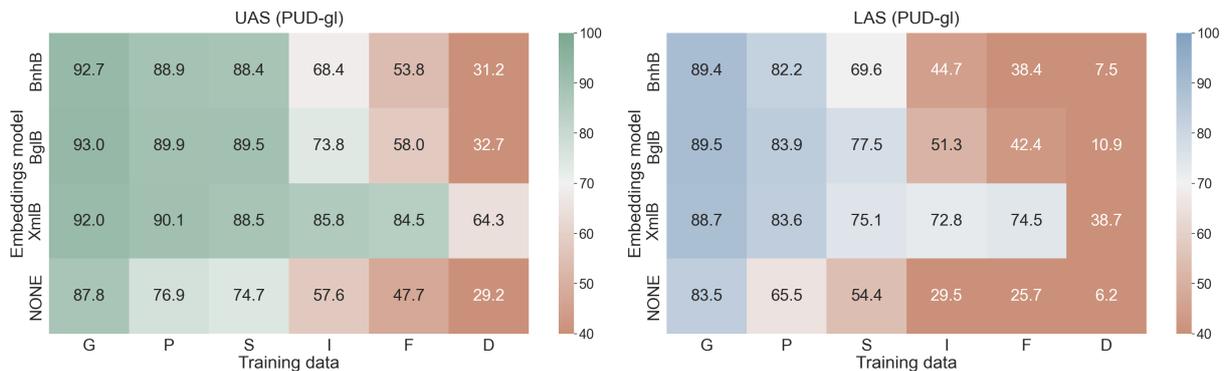


Figure 1: Performance of models trained in a sole language. Both images columns represent a different language on which the parsers was trained and each row a distinct embedding model used by the parser.

each cell shows the precision of correctly inferred dependencies for a given parser. In the row *Others* we calculated the average of that proportion for all other dependencies that were not represented in the Figure, since the total amount of dependencies is of 32.

The analysis reveals that certain dependencies are consistently more difficult to parse across all models, regardless of the source language or embedding type. The most problematic relations are the clausal modifier of noun (*acl*) and numeric modifier (*nummod*), while oblique nominals (*obl*), nominal modifiers (*nmod*), and nominal subjects (*nsubj*) also exhibit high error rates despite their high frequency. As expected, parsers trained on other source languages struggle most with infrequent syntactic dependencies. For instance, even when using embeddings, the precision gap between the Galician-trained parser (G) and the Portuguese-trained parser (P) is of 0.25 points. This gap widens dramatically for more typologically distant languages.

Although embeddings generally improve performance for all languages (noting that distant languages require multilingual embeddings), they enable competitive accuracy on some otherwise problematic dependencies for specific languages (e.g., determinants in Italian and Spanish, or oblique nominals in Portuguese). In that sense, the Figure 2 shows that each source language tends to accentuate specific error patterns. For instance, parsers trained on Portuguese and French more frequently misidentify oblique nominal (*obl*) and nominal modifier (*nmod*) relations, whereas those trained on Spanish and Italian show higher error rates on direct objects (*obj*).

In summary, typological proximity has a

stronger positive effect on UAS than on the stricter LAS. Portuguese and Spanish are sufficiently close to Galician to train parsers of decent quality, with results not far from those trained on Galician data itself. Moreover, our experiments reveal a clear performance gradient when using different source languages to train parsers for a target language. For Galician, Ibero-Romance languages (Portuguese, Spanish) yield results closest to the Galician-trained baseline. Other Romance languages (French, Italian) can achieve relevant performance, but only when multilingual embeddings are employed. Parsers trained on German data, however, fail to achieve decent results even with multilingual embeddings.

Notably, the performance gap between using other Romance languages and using German is substantially larger than the gap between Ibero-Romance and other Romance languages. This holds despite the fact that, for example, both French and German are largely intelligible to Galician speakers, suggesting that mutual intelligibility is a poor proxy for parser transferability compared to formal syntactic similarity.

4.2 Mixture of Languages

In this section we present the results of combining diverse languages with Galician in the training of parsers. The Table 3 illustrates that results.

UAS and LAS improvements. As can be seen in the Table 3, the improvements gained from mixing languages are limited. The results obtained on the PUD-gl test partition by models trained solely on Galician (G) are similar to those obtained by models trained with partition composed by multiple languages (e.g., G+P or G+S). It is true that, when no embedding models are used, language mixing

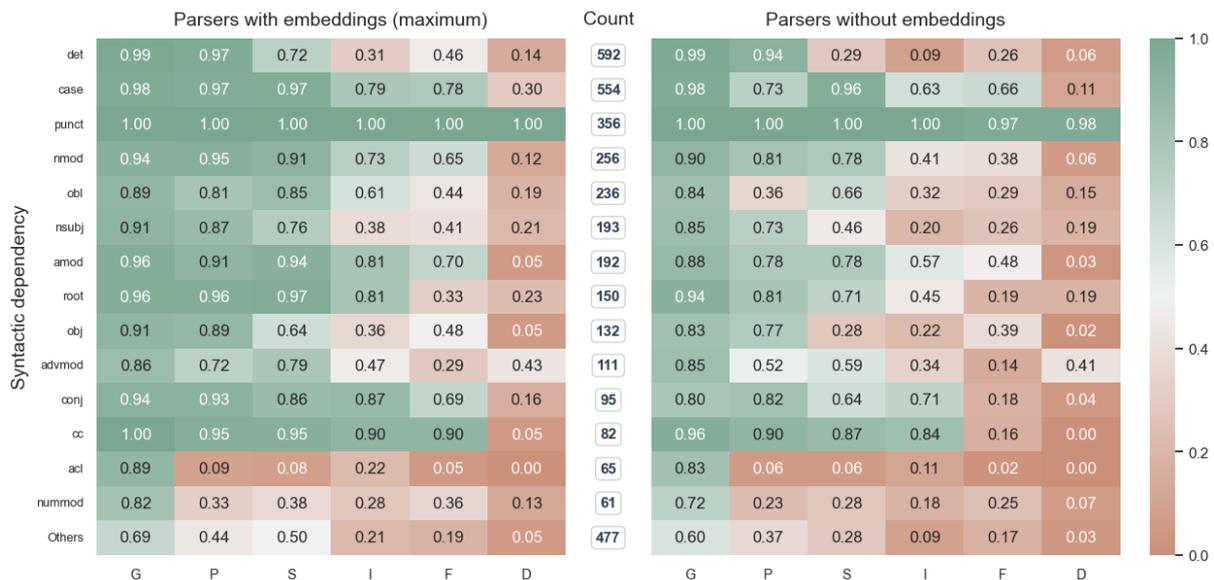


Figure 2: Error analysis disaggregated by dependencies. The left image shows the precision of the parsers that use embeddings (the maximum score obtained between the three models previously explained) and the right image the scores obtained by the parsers without embeddings. *Count* indicates the number of occurrences of each dependency.

Parsers	BglB	XmlB	NONE
G	93.0 89.5	92.0 88.7	87.8 83.5
G+P	93.1 89.5	93.0 89.8	88.0 84.9
G+S	93.3 90.1	92.6 89.3	89.3 84.8
G+P+S	93.1 89.9	92.4 89.0	87.8 84.7
G+I+F	92.4 89.2	92.4 89.2	87.9 83.6

Table 3: Results of language mixture in training. Each row represent a different training partition -thus, a specific language mixture- and each column a different embedding model. Each cell contains: UAS | LAS.

brings greater improvements: the gain between G and G+S models is 0.3 in UAS when using Bert-GL embeddings, and 1.5 when using no embeddings.

Mixing closely related languages does not introduce noise into the training; results always improve, even if only slightly. This is the case for Spanish (G+S) and Portuguese (G+P) with Galician: mixing either with Galician yields small but consistent gains. Increasing the number of mixed languages also does not seem to add more noise to the training, provided the added languages are themselves typologically close (Spanish and Portuguese, G+P+S partition), although the improvements are not greater either. However, as the typological proximity of the mixed languages decreases, they become noisier: combining Italian or French with Galician yields poorer results than training with Galician alone.

Dependency-level error analysis. The analy-

sis follows the trend observed with the automatic metrics. When using embedding models, parsers augmented with data from other languages rarely surpass the performance of the parser trained exclusively on Galician (G), and the observed gains are limited to specific dependencies. It can be affirmed that language mixing is generally unable to yield substantial improvements for the problematic dependencies identified in Section 4.1, as they remain persistent challenges.

However, when embedding models are not used, small but consistent improvements become apparent in several grammatical dependencies: oblique nominals, nominal subjects, adjectival modifiers, etc. In all cases, an improvement of at least 0.02 is observed. When evaluation is performed on the TreeGal treebank, these improvements persist, whether weaker or stronger (see Figure 4). It is noteworthy, furthermore, that for low-frequency dependencies (grouped as *Others*), language mixing has a positive effect, particularly on the TreeGal treebank. Concerning the mixing of specific languages, the most widespread improvements across dependencies are observed when combining Galician and Portuguese (the G+P parsers). There are very few dependencies for which the parsers trained with mixed languages make more errors than those trained only on Galician, even when the added languages are Italian or French.

Depen.	PUD-E	PUD-N	Tre-E	Tre-N
obl	0.02	0.03	0.06	0.01
nsubj	0.01	0.04	0.02	0.02
amod	0	0.04	0.01	0.03
obj	-0.01	0.03	0.01	0.01
conj	-0.01	0.04	-0.01	0.01
acl	0.03	0.02	0.07	0.01
Others	0.02	0.02	0.03	0.06

Table 4: Improvements in problematic dependencies. Values indicate the gain in precision when using language mixture training (G+P, G+S, G+P+S, G+I+F) over Galician-only training (G). Columns are organized by test corpus (PUD-gl, TreeGal), with each pair showing results without embeddings (first column) and with embeddings (second column).

4.3 Comparison Between Treebanks

Table 5 compares the performance scores of parsers trained and evaluated on the PUD-gl and TreeGal treebanks, using the four embedding models described previously.

Different performances between treebanks.

Despite both treebanks being in Galician and having the same size, models evaluated on the treebank they were not trained on performed significantly worse than those trained and tested on the same treebank. For example, parsers trained on PUD-gl suffered a performance drop of 8–13 UAS points and 11–16 LAS points when evaluated on TreeGal, compared to their evaluation on PUD-gl. Furthermore, parsers trained and evaluated exclusively on TreeGal outperformed those trained on PUD-gl and evaluated on TreeGal, achieving roughly 5 more UAS points and 8 more LAS points. Similarly, parsers trained on TreeGal performed worse when evaluated on PUD-gl compared to parsers trained and evaluated on PUD-gl.

TreeGal more challenging than PUD-gl. The results indicate that the TreeGal treebank appears to present a greater parsing challenge. This is observable in Table 5, where, in some configurations, parsers trained on TreeGal achieve higher scores when evaluated on PUD-gl than when evaluated on their own training treebank, TreeGal (compare columns Tr-Tr and Tr-Pu for UAS).

Embeddings as flexibility providers. This counterintuitive effect occurs only for parsers that use embeddings. The use of embeddings not only improves overall performance but also reduces the performance gap between evaluation on the in-domain treebank (the one used for training) and the

out-of-domain treebank (the other treebank). For instance, for parsers trained on PUD-gl with embeddings, the performance drop when evaluating on TreeGal (versus PUD-gl) is approximately 8 UAS points and 12 LAS points. In contrast, for parsers trained on PUD-gl without embeddings, this cross-treebank performance drop increases to 13 UAS and 16.3 LAS points. Therefore, the inclusion of embeddings enhances the parser’s generalization capacity, yielding models that are more robust and flexible across different annotation schemes.

Greater performance gap on LAS. Finally, the results demonstrate that the performance gap between in-domain evaluation (same treebank used for training and testing) and cross-treebank evaluation is wider for the LAS than for the UAS—by an average of three points. This is consistent with expectations, as LAS is a more restrictive metric that requires correct grammatical labels in addition to syntactic structure.

Based on the findings presented in this subsection, we conclude that automatic evaluation scores vary considerably across treebanks, even when they represent the same language and are of comparable size. These variations likely stem from differences in annotation quality, guideline compatibility, and the generalization capacity of the parsers themselves.

4.4 Scalability

The last objective of this work was to measure the scalability of language mixing as a step toward building more precise syntactic parsers for Galician. Table 6 presents the results of these experiments. Following what was shown in the previous section, this experiment also demonstrates that language mixing can only yield minor improvements when robust embedding models are used, even if the size of the added partition is five times larger. The UAS difference between the G parser and the scaled parsers (G+PP, G+SS, and G+PPSS) is, at best, 1.1 points in favor of the latter; the LAS difference, however, is 1.7 points. In contrast, when embedding models are not used, the scalability of language mixing is clearly observed: the GCP parser achieves a 3.2-point higher UAS and a 3.5-point higher LAS.

Considering these results, we conclude that scaling syntactic parsers for Galician by directly using large treebanks from closely related languages—without any pre- or post-processing—is not an especially promising approach, given the

Model	UAS					LAS				
	Pu-Pu	Pu-Tr	Tr-Tr	Tr-Pu	Dif.	Pu-Pu	Pu-Tr	Tr-Tr	Tr-Pu	Dif.
BnhB	91.8	82.9	87	88.4	8.9	88.4	76.2	83	83	12.2
BglB	91.4	83.1	87.4	87.9	8.3	88.3	76.8	83.8	82.5	11.5
XmlB	91	82.2	86.7	87.1	8.8	87.7	76	82.6	82	11.7
NONE	88	75	80.5	76.6	13	83.8	67.5	75.1	70.1	16.3

Table 5: Performance comparison across treebanks. The subcolumns represent the following configurations: parser trained and evaluated on PUD-gl (Pu–Pu), trained on PUD-gl and evaluated on TreeGal (Pu–Tr), trained and evaluated on TreeGal (Tr–Tr), and trained on TreeGal and evaluated on PUD-gl (Tr–Pu).

Parsers	BglB	XmlB	NONE
G	93.0 89.5	92.0 88.7	87.8 83.5
G+PP	93.8 91.0	93.2 90.0	90.5 86.7
G+SS	94.1 90.0	93.1 90.0	91.2 87.3
G+PPSS	93.6 90.7	93.3 90.1	91.0 87.0

Table 6: Scalability of language mixture in Galician. See Section 3 for information about parsers and experiment configuration. Each cell contains: UAS | LAS.

limited improvements. Two main reasons may explain this:

- The baseline performance of Galician parsers is already high, achieving UAS and LAS scores between 80 and 95 points. Therefore, the margin for improvement is not substantial, and, as shown in previous sections, only certain specific dependencies remain problematic.
- Galician already benefits from robust embedding models, which are capable of capturing on their own the types of gains that might otherwise be sought through language mixing. Thus, these models present a strong alternative to cross-lingual data augmentation.

Accordingly, we conclude that scaling parsers for Galician requires high-quality data. To effectively leverage language mixing, it would likely be necessary to develop specific pre- and post-processing techniques aimed at improving performance on particular problematic dependencies. However, the data also show that for languages and linguistic varieties with fewer resources than Galician—perhaps with greater interest in the latter—language mixing can still yield significant improvements as an alternative to embedding models.

To conclude the results section, we reproduced the experiments described above, this time evaluating them on the test partition of the TreeGal tree-

bank, and we arrived at similar results. This suggests that our findings are treebank-independent, both for the scalability experiment and for the previous ones, as shown in the data provided in Appendix C.

5 Conclusion

A promising approach for training syntactic parsers in a low-resource language is to leverage syntactic annotations from major languages with significant typological proximity. In the case of Galician, parsers trained exclusively on Portuguese or Spanish can achieve accuracy close to those trained on Galician itself. Furthermore, the use of embeddings significantly enhances parser performance, to the point where models trained on more distantly related languages become viable. However, in the pursuit of more precise Galician parsers, we have identified three key limitations: (i) the gains from mixing Galician annotations with those of larger languages are modest; (ii) parsers exhibit a weak generalization capacity across different treebanks; and (iii) certain syntactic dependencies—such as oblique nominals and adverbial modifiers for Galician—, which could be language specific, remain problematic for parsers trained on any data source. Overall, this work demonstrates that strategically exploiting data from other languages can open pathways for many languages to obtain better parsers.

6 Limitations

An important limitation of this study is one commonly noted in the dependency parsing literature: the lack of diversity in text genres used for training and evaluation. Most available treebanks consist of journalistic and encyclopedic texts, as these are the easiest to crawl from the web. However, natural language is used in a much wider variety of genres—including literary works, transcribed speech, legal documents, and technical writing—which are

not well represented in current resources. Future work should address this gap by creating treebanks with broader genre coverage and subsequently developing parsers adapted to these diverse text types.

A more technical limitation of our study is that we did not perform hyperparameter tuning during parser training, and we used only the UDPipe 2 architecture. Future research could apply our experimental framework to other systems—such as Stanza or LLM-based parsers—to determine whether our findings generalize across different architectures. In that sense, exploring the instruction of the Galician-trained LLM Carballo (Gamallo et al., 2024) for dependency parsing is a promising path. Additionally, future work should undertake a more realistic evaluation by parsing raw text directly, rather than relying on gold tokenization.

Acknowledgments

This paper was funded by MCIU/AEI/10.13039/501100011033 (grants with references PID 2021-128811OA-I00, PID2024-161928OB-I00, CNS2024-154902, and AIA2025-163322-C62), and by the Galician Government (ED431G 2023/04 and ED431B 2025/16). Xabier Irastortza-Urbieta is supported by a doctoral grant from the Basque Government (PRE_2025_1_0026).

References

- Benedikt Adelmann, Wolfgang Menzel, and Heike Zinsmeister. 2021. [The impact of word embeddings on neural dependency parsing](#). In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, pages 1–13, Düsseldorf, Germany. KONVENS 2021 Organizers.
- Ricardo Carvalho Calero. 1985. O problema ortográfico. *Agália: Publicaçom internacional da Associaçom Galega da Lingua*, 2:127–134.
- 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.
- Sara Court and Micha Elsner. 2024. [Shortcomings of LLMs for low-resource translation: Retrieval and understanding are both the problem](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 1332–1354, Miami, Florida, USA. Association for Computational Linguistics.
- 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).
- Cheikh M. Bamba Dione. 2021. [Multilingual dependency parsing for low-resource African languages: Case studies on Bambara, Wolof, and Yoruba](#). In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, pages 84–92, Online. Association for Computational Linguistics.
- Jason M. Eisner. 1996. [Three new probabilistic models for dependency parsing: An exploration](#). In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.
- Ana Ezquerro, Carlos Gómez-Rodríguez, and David Vilares. 2025. [Better benchmarking LLMs for zero-shot dependency parsing](#). In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 121–135, Tallinn, Estonia. University of Tartu Library.
- Pablo Gamallo and Isaac González. 2012. [Deppattern: A multilingual dependency parser](#). In *Demo Session of the International Conference on Computational Processing of the Portuguese Language (PROPOR 2012)*, Coimbra, Portugal.
- Pablo Gamallo, Pablo Rodríguez, Iria de Dios-Flores, Susana Sotelo, Silvia Paniagua, Daniel Bardanca, José Ramon Pichel, and Marcos Garcia. 2024. [Open Generative Large Language Models for Galician](#). *Procesamiento del Lenguaje Natural*, 73(0):259–270.
- Marcos Garcia. 2016. [Universal dependencies guidelines for the galician-treegal treebank](#). In *Technical Report*. LyS Group, Universidade da Coruña.
- Marcos Garcia. 2021. [Exploring the representation of word meanings in context: A case study on homonymy and synonymy](#). 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 3625–3640. Association for Computational Linguistics.
- Fadli Aulawi Al Ghiffari, Ika Alfina, and Kurniawati Azizah. 2023. [Cross-lingual transfer learning for Javanese dependency parsing](#). In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational*

- Linguistics: Student Research Workshop*, pages 1–9, Nusa Dua, Bali. Association for Computational Linguistics.
- Goran Glavaš and Ivan Vulić. 2021. [Climbing the tower of treebanks: Improving low-resource dependency parsing via hierarchical source selection](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4878–4888, Online. Association for Computational Linguistics.
- Xavier Gómez Guinovart. 2017. Recursos integrados da lingua galega para a investigación lingüística. In *Galacia: Estudos de lingüística portuguesa e galega*, pages 1037–1048. Universidade de Santiago de Compostela.
- Claudiu Daniel Hromei, Danilo Croce, and Roberto Basili. 2024. U-deppllama: Universal dependency parsing via auto-regressive large language models. *Italian Journal of Computational Linguistics*, 10(1):21–38.
- Suramya Jadhav, Abhay Shanbhag, Amogh Thakurdesai, Ridhima Sinare, and Raviraj Joshi. 2025. [On limitations of LLM as annotator for low resource languages](#). In *Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025)*, pages 277–282, Southern Denmark University, Odense, Denmark. Association for Computational Linguistics.
- Dan Kondratyuk and Milan Straka. 2019. [75 languages, 1 model: Parsing Universal Dependencies universally](#). 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 2779–2795, Hong Kong, China. Association for Computational Linguistics.
- Lucelene Lopes and Thiago Pardo. 2024. [Towards portparser - a highly accurate parsing system for Brazilian Portuguese following the Universal Dependencies framework](#). In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 401–410, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.
- 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, Johan Hall, and Jens Nilsson. 2004. [Memory-based dependency parsing](#). In *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004*, pages 49–56, Boston, Massachusetts, USA. Association for Computational Linguistics.
- Barbara Plank, Héctor Martínez Alonso, Željko Agić, Danijela Merkle, and Anders Søgaard. 2015. [Do dependency parsing metrics correlate with human judgments?](#) In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 315–320, Beijing, China. Association for Computational Linguistics.
- 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, Online. Association for Computational Linguistics.
- Xulia Sánchez-Rodríguez, Albina Sarymsakova, Laura Castro, and Marcos Garcia. 2024. [Increasing manually annotated resources for Galician: the parallel Universal Dependencies treebank](#). In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 587–592, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.
- Albina Sarymsakova, Xulia Sánchez Rodríguez, and Marcos García González. 2024. Towards accurate dependency parsing for galician with limited resources. *Procesamiento del Lenguaje Natural*, 73:247–257.
- 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.
- Nasrin Taghizadeh and Hesham Faily. 2022. [Cross-lingual transfer learning for relation extraction using universal dependencies](#). *Computer Speech & Language*, 71:101265.
- David Vilares Calvo, Marcos García González, and Carlos Gómez Rodríguez. 2021. Bertinho: Galician bert representations. *Procesamiento del Lenguaje Natural*, 66:13–26.
- Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Uřešová, and 43 others. 2017. [CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies](#). In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.

Meishan Zhang, Gongyao Jiang, Shuang Liu, Jing Chen, and Min Zhang. 2024. LLM-assisted data augmentation for Chinese dialogue-level dependency parsing. *Computational Linguistics*, 50(3):867–891.

A Available Resources

The code used to produce the results presented in this paper is publicly available at: <https://zenodo.org/records/18007073>.

B Glossary of Syntactic Dependencies

Through this paper we used abbreviations to reference syntactic dependencies. In this appendix we provide a glossary to explain each abbreviation. In general, the abbreviations we have used are the same employed by Universal Dependencies in all their treebanks.

- *Det*: Determinant.
- *Case*: Case marking.
- *Punct*: Punctuation.
- *Nmod*: Nominal modifier.
- *Obl*: Oblique nominal.
- *Nsubj*: Nominal subject.
- *Amod*: Adjectival modifier.
- *Root*: Root.
- *Obj*: Direct object.
- *Advmod*: Adverbial modifier.
- *Conj*: Conjunct.
- *Cc*: Coordinating conjunction.
- *Acl*: Clausal modifier of noun.
- *Nummod*: Numeric modifier.

C Supplementary Results

In this appendix, we provide a complementary set of results for the experiments described in Section 3, this time evaluating them on the TreeGal treebank. This contrasts with the primary results presented in the main paper, which were based on evaluation using the PUD-gl treebank. The strong similarity between the two sets of results reinforces the conclusions presented in Section 5.

With regard to **typological proximity**, Table 7 reiterates the results presented in Section 4.1. The

Parsers	BglB	XmlB	NONE
G	83.1 77.5	81.7 76.3	77.1 70.3
P	83.5 76.2	83.3 75.9	69.2 56.8
S	82.1 70.8	80.7 67.8	67.4 48.2
I	68.1 48.8	80.5 69.3	53.0 29.2
F	57.0 42.0	78.7 69.5	46.6 26.2
A	30.4 10.0	57.9 33.3	28.4 5.8

Table 7: Results from the typological proximity experiment (Section 4.1), evaluated on the TreeGal test partition. Each cell contains: UAS | LAS.

Parsers	BglB	XmlB	NONE
G	83.1 77.5	81.7 76.3	77.1 70.3
G+P	84.6 79.4	83.2 77.5	78.1 71.4
G+S	84.5 79.1	83.3 77.6	77.7 71.0
G+P+S	85.0 79.7	83.1 77.1	78.2 71.6
G+I+F	83.9 78.2	83.1 77.8	78.1 71.2

Table 8: Results from the language mixture experiment (Section 4.2), evaluated on the TreeGal test partition. Each cell contains: UAS | LAS.

influence of typological proximity when training parsers on different source languages remains a key factor for achieving high performance. Embedding models continue to provide a considerable improvement. It should be noted that the absolute results on TreeGal are lower than those on PUD-gl, consistent with the analysis in Section 4.3.

Language mixture. Table 8 exposes the results for the parsers from Section 4.2—trained via language mixture—when evaluated on the TreeGal treebank. Note that all parsers were trained on the PUD-gl treebank. We reach the same conclusion as in the main analysis: while language mixture can improve results, its benefits are subject to significant limitations.

Also, the results of the **scalability** experiment (Section 4.4) can be reproduced when evaluating on the TreeGal treebank, as shown in Table 9. Note that in this table, the parser in the row labeled *TRE* was trained exclusively on the 800-sentence training partition of TreeGal (in order to do a fair comparison). This same Galician data was also included in the training sets of the scaled parsers (G+PP, G+SS, G+PPSS), as described in Section 3.

On the other hand, the conclusions drawn from **error analyses** are also reproducible when evaluating on the TreeGal dataset. Figure 3 shows that the set of problematic dependencies remains similar to that identified using the PUD-gl treebank. For

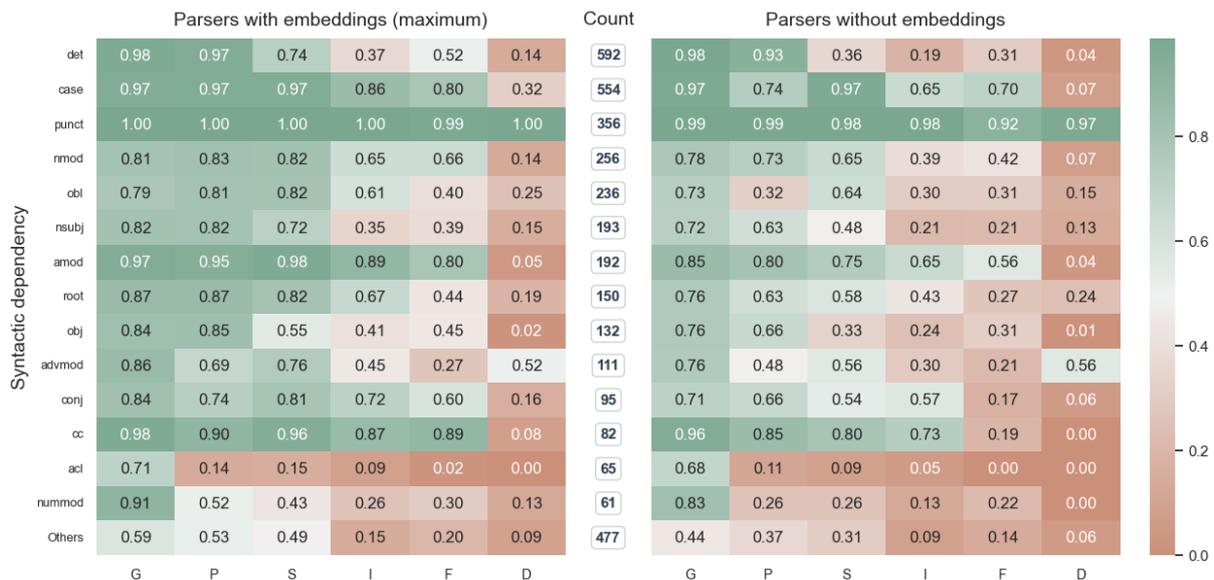


Figure 3: Error analysis disaggregated by dependencies, in the same format as Figure 2. The evaluation was conducted on the test partition of the TreeGal treebank.

Parsers	BglB	XmlB	NONE
TRE	88.8 85.7	86.8 83.7	81.5 76.5
G+PP	89.4 86.3	88.1 84.5	83.5 79.0
G+SS	90.0 87.0	89.5 86.6	86.0 82.0
G+PPSS	89.5 86.4	88.6 85.4	84.6 80.1

Table 9: Results from the scalability experiment (Section 4.4), evaluated on the TreeGal test partition. Each cell contains: UAS | LAS.

further analysis, the complete code used to generate the results in this paper is available via the link provided in Appendix A.

C.1 Ambiguity in Parsing

Finally, it should be noted that certain errors identified during the evaluation reflect **annotation decisions that are inherently debatable or ambiguous** even for human annotators. To illustrate this point, we provide the following examples:

- In the sentence "*Curio, [...], informou personalmente a César das accións de Pompeio*" ("Curio, [...], personally informed Caesar of Pompey's actions"), the parser employing BglB embeddings (parser G) predicts *César* as a direct object, whereas the gold standard classifies it as an oblique nominal. This case is ambiguous due to the fuzzy boundaries in some cases between direct objects (obj), indirect objects (iobj), and obliques (obl) in Galician.

- The same parser frequently exhibits confusion between discourse elements and adverbial modifiers (advmod), as in the examples: *Porén, cando o Senado lle respondeu categoricamente, prohibíndolle...* ("However, when the Senate answered him categorically, prohibiting him...") and *Agora, este departamento enfróntase a novos desafíos* ("Now, this department faces new challenges"). In such instances, the distinction between the adverbial function as a circumstantial modifier (advmod) and its role as a discourse marker is often subtle and not clearly defined.

Consequently, future qualitative analyses should also consider the annotation conventions of Universal Dependencies, along with the potential confusions or ambiguities they may introduce.

Crowdsourcing Piedmontese to Test LLMs on Non-Standard Orthography

Gianluca Vico and Jindřich Libovický

Charles University, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00 Praha, Czech Republic
{vico, libovicky}@ufal.mff.cuni.cz

Abstract

We present a crowdsourced dataset for Piedmontese, an endangered Romance language of northwestern Italy. The dataset comprises 145 Italian–Piedmontese parallel sentences derived from Flores+, with translations produced by speakers writing in their natural orthographic style rather than adhering to standardized conventions, along with manual word alignment. We use this resource to benchmark several large language models on tokenization parity, topic classification, and machine translation. Our analysis reveals that Piedmontese incurs a tokenization penalty relative to higher-resource Romance languages, yet LLMs achieve classification performance approaching that of Italian, French, and English. Machine translation results are asymmetric: models translate adequately from Piedmontese into high-resource languages, but generation into Piedmontese remains challenging. The dataset and code are publicly released.

1 Introduction

Piedmontese (ISO 639-3: pms) is a Romance language spoken in the Piedmont region of northwestern Italy. According to Ethnologue (Eberhard et al., 2025), it has fewer than one million speakers and is classified as endangered, with intergenerational transmission in decline.

Existing NLP resources for Piedmontese are limited and predominantly derived from Piedmontese Wikipedia. While useful, these sources largely adhere to standardized orthographic conventions and thus fail to capture the orthographic variations that are common in written Piedmontese. This discrepancy raises the question of how well current language models handle Piedmontese as it is actually written by speakers.

To address this gap, we present a crowdsourced dataset of Italian-to-Piedmontese translations, where annotators were explicitly instructed

Flores+ <i>dev</i> 114	
Ita	Si tratta della maggiore acquisizione nella storia di eBay.
Pms	A l'è la pi granda aquisission ënt la stòria d'ebay
Eng	It is the biggest acquisition in eBay's history.
Fra	C'est la plus grande acquisition de l'histoire d'eBay.

Table 1: Sample from parallel sentences for evaluating machine translation. Annotators translated the Italian sample into Piedmontese. The Italian, French and English samples are originally from Flores+.

to write in whichever orthographic style feels natural to them. The source sentences are drawn from the Flores+ dataset (NLLB Team et al., 2024), a multiparallel corpus spanning over 200 languages. We additionally provide manual word alignments between Piedmontese and Italian sentence pairs.

Using this data, we evaluate several large language models (LLMs) both intrinsically, through tokenization parity (Petrov et al., 2023) analysis, and extrinsically, on topic classification (using labels from SIB-200; Adelani et al., 2024) and machine translation (MT). Our results indicate that current LLMs exhibit reasonable comprehension of Piedmontese, achieving decent performance in classification and translation from Piedmontese into high-resource languages. However generation into Piedmontese remains substantially more challenging.

We illustrate the data collection procedure in Section 2. In Section 3, we describe the dataset, and in Section 4, we assess LLM performance on Piedmontese. Section 5 presents related datasets and in Section 6 we summarise our findings.

The dataset¹ and evaluation code² are released under an open-source license.

2 Data Collection

We collected translations via an online questionnaire administered in Italian (see Appendix H), the dominant language in the region, understood by all Piedmontese speakers. Annotators were recruited voluntarily through social media and word of mouth, with no restrictions on repeated participation. To preserve anonymity, we did not track annotator identity across sessions.

The questionnaire comprises three components. First, we elicit demographic and sociolinguistic information, including the annotator’s primary language, self-assessed proficiency in Piedmontese, age group, and method of language acquisition. We also ask whether they believe Piedmontese has a standard orthography and, if so, whether it is commonly used. These questions serve both to characterize our annotator population and to contextualize the orthographic variation in the resulting translations.

Second, we present annotators with a randomly selected Italian sentence from Flores+ (NLLB Team et al., 2024) and ask them to translate it into Piedmontese. Crucially, annotators are instructed to write in whatever manner feels natural to them, rather than adhering to any prescribed standard. Translation is optional, so annotators can still complete the other parts of the questionnaire. In this way, we can accommodate annotators who comprehend Piedmontese but do not actively write it. To address the absence of certain diacritics on standard physical keyboards, we provide a substitution scheme (e.g., /:a for ä). Mobile keyboards do not have this issue, and we observed that people either directly use diacritics or use diacritics that can be found on Italian keyboards (àèèìòù).

Third, annotators evaluate a translation submitted by a previous participant, viewing both the Italian source and the Piedmontese rendering. This peer review mechanism enables filtering of erroneous or inappropriate submissions and provides an estimate of inter-annotator agreement on translation quality. While the task is subjective, we ask annotators to take into consideration possible variations of the language and of the orthography.

¹<http://hdl.handle.net/11372/LRT-6086>

²<https://github.com/GianlucaVico/CrowdsourcedPiedmontese>

3 Dataset Description

We have collected 200 annotations, and 145 of them have valid translations: 68 are from the Flores+ *dev* set, while 77 are from the *devtest* set. 102 samples have been reviewed by at least one annotator, but due to their limited number. We use the reviews only to filter missing or offensive translations.

We organise the collected data in three datasets: 1) the raw list of annotations that can be used for further analysis, 2) a list of parallel sentences for evaluating MT systems, and 3) a list of word-aligned sentences.

3.1 Annotation

Figure 1 shows that most annotators use primarily Italian, and a few use Piedmontese. Other languages include English and Icelandic. The proportion of annotators who submitted a translation is higher among Piedmontese speakers than among Italian speakers. Additionally, most annotators declared themselves to be perfectly or fully proficient in Piedmontese. Most of the annotators are confident in their language knowledge; however, only a small portion considers Piedmontese their native language.

Our questionnaire reached mainly younger people, as shown in Figure 2, but, since this is an endangered language, older people are more likely to speak it.

On average, completing the questionnaire took approximately 7 minutes. People who did not provide a translation took approximately 3 minutes. According to 11% of the annotators, people use standard grammar when writing Piedmontese, while 42% of them disagree. 54% of them think that Piedmontese has a standard grammar, whether it is used or not, and for 25% of the annotators, there is no standard.

3.2 Parallel Sentences

Flores+ contains 2009 samples divided into the *dev* and *devtest* splits. The sentences provided to the annotators are randomly selected, so some of them have multiple translations: three samples from the *devtest* set and one sample from the *dev* set have two translations. The paired sentences have the same overall meaning, but translation quality varies; for example, annotators may use more general terms, summarise a list or remove details. 102 samples have at least one human review, which we

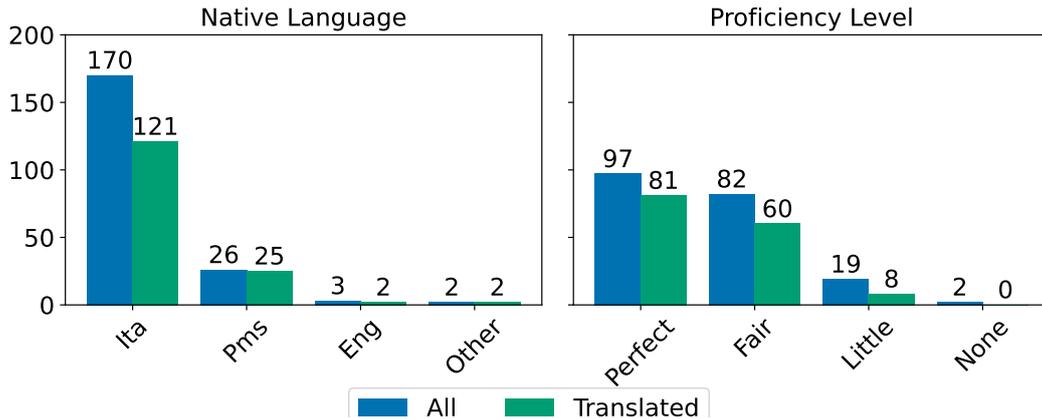


Figure 1: On the left, the main language used by the annotators; Icelandic is included in *Other*. On the right, the self-reported proficiency in Piedmontese. The majority of people uses Italian and self-reports perfect or fair proficiency in Piedmontese.

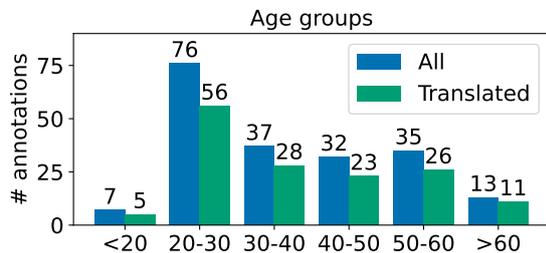


Figure 2: Age distribution of the annotators. Most annotators are 20-30 years old, while older people are more likely to know Piedmontese, but we did not reach them.

used to remove incorrect translations. We present a sample in Table 1: as can be seen, the Piedmontese text may contain incorrect capitalization or missing punctuation. Also, the use of diacritics is inconsistent among annotators.

3.3 Word Aligned Sentences

Due to the limited number of samples, the authors are able to manually word-align the Piedmontese and Italian sentences. We select pairs of corresponding spans in the paired sentences, ensuring that each span is non-overlapping (i.e., each word belongs to at most one span) In this sense, there are cases where, for example, a verb is aligned with a noun because they convey the same meaning, but the sentence structure differs. We use the white space and apostrophe to split words. As an example, *e sull'albero* (*and on the tree*) consists of three words: *[e][sull'][albero]*. One word can be aligned to multiple words, e.g., *è* (*is*) is aligned to *a l'è*, and there are unaligned words. However, we

do not consider subword alignment. The dataset comprises 3003 spans, with a median of 20 spans per sentence pair. 2902 spans are a single word aligned to another single word. The median number of characters for each span is 5 for both Italian and Piedmontese.

4 Model Evaluation

To assess LLM performance on Piedmontese, we first evaluate tokenizer parity (Petrov et al., 2023): this provides an estimate of the costs in tokens of processing Piedmontese compared to other languages. Then, we use the aligned dataset to investigate whether models can find corresponding words between Piedmontese and Italian. Finally, we use topic classification and machine translation as downstream tasks for evaluation. In topic classification, models need to be able to understand Piedmontese, while in machine translation, they also have to generate Piedmontese. The downstream tasks are evaluated in a zero-shot setup.

We consider the following open-weight models from HuggingFace: Llama 3.3 70B³ (Grattafiori et al., 2024), Gemma 3 27B⁴ (Gemma Team, 2025), Qwen 3 30B⁵ (Qwen Team, 2025), EuroLLM 9B⁶ (Martins et al., 2025), Tower Plus 9B⁷ (Rei et al., 2025); and the closed-source models: Gemini⁸ and GPT⁹. Besides Piedmontese and Italian, we also

³meta-llama/Llama-3.3-70B-Instruct

⁴google/gemma-3-27b-it

⁵Qwen/Qwen3-30B-A3B-Instruct-2507

⁶utter-project/EuroLLM-9B-Instruct

⁷Unbabel/Tower-Plus-9B

⁸gemini-2.5-flash-preview-09-2025

⁹gpt-4o-mini

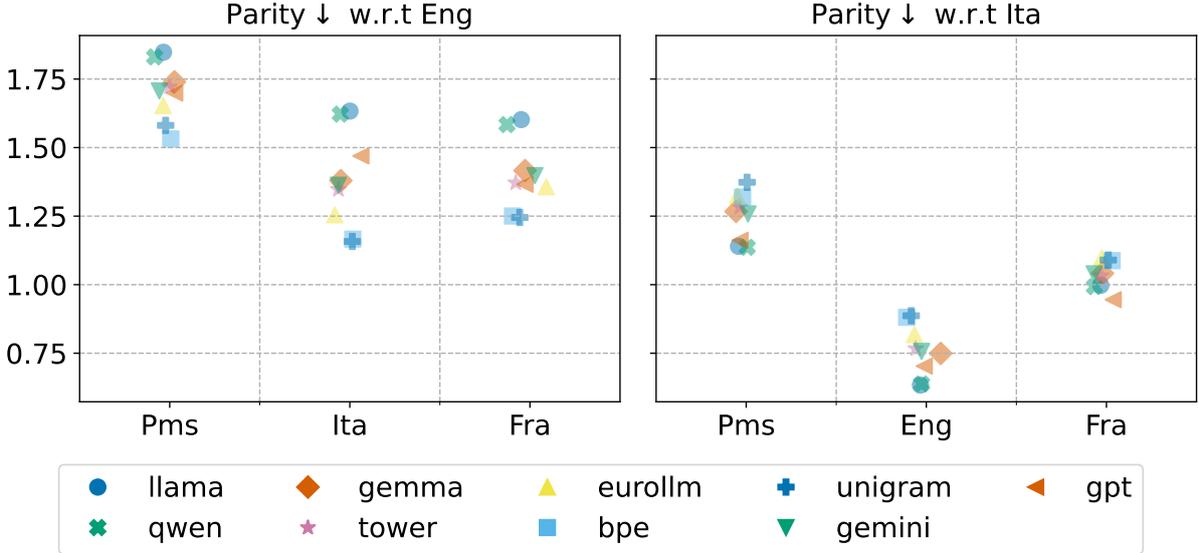


Figure 3: Parity scores with respect to English and Italian. Piedmontese has worse parity compared to the other languages; however, it is closer to one when compared to Italian.

include French, as it is the high-resource language closest to Piedmontese, other than Italian, and English, due to its widespread availability. Because the data is limited and we do not perform any parameter search, we evaluate the model on the combined *dev* and *devtest* sets. The hyper-parameters for the experiments are listed in Appendix F.

4.1 Tokenizer Parity

As shown by Ahia et al. (2023), low-resource languages are often overtokenized, resulting in higher costs and worse performance compared to high-resource languages. We evaluate the tokenizer parity (Petrov et al., 2023) for the LLMs, UnigramLM, and BPE from SentencePiece to estimate the number of extra tokens required to process the same sentence in Piedmontese.

We train the SentencePiece tokenizers using English, Italian, French, and Piedmontese data from the Glot500 Corpus (Imani et al., 2023), with 100k samples for each language and a vocabulary size of 32,000.

We average the parity of each sample, computed as:

$$p_t(s_{\text{tgt}}, s_{\text{ref}}) = \frac{|t(s_{\text{tgt}})|}{|t(s_{\text{ref}})|}$$

where t is a tokenization function that produces a list of tokens, s_{tgt} is a sentence in the target language, and s_{ref} is the corresponding sentence in the reference language. As reference languages, we use English and Italian. A parity close to one

Model	F ₁ ↑	Precision ↑	Recall ↑
Eflomal	.774	.817	.735
SimAlign	.589	.726	.496

Table 2: Alignment scores of eflomal and SimAlign.

indicates that the tokenizer produces a similar number of tokens for the source and target languages, whereas values greater than one indicate that the target language is over-tokenized.

In Figure 3, we report the parity scores of the models. Piedmontese has worse (i.e., higher) parity than the other languages, which means that using LLMs with Piedmontese is more computationally expensive. Training the tokenizer on Piedmontese can help, as BPE and UnigramLM have lower parity compared to English. However, overall, the models yield comparable results, and closed models do not have an advantage over the open-weight models. In Appendix D, we report the exact parity scores for the different setups.

4.2 Word Alignment

We use eflomal (Östling and Tiedemann, 2016) trained on our dataset as a baseline and compare with the unsupervised SimAlign (Jalili Sabet et al., 2020) with XLM-RoBERTa (Conneau et al., 2020) with subwords. XLM-RoBERTa is a multilingual model, but Piedmontese was not explicitly included in the training data. We use the same evaluation script from SimAlign, which reports precision, re-

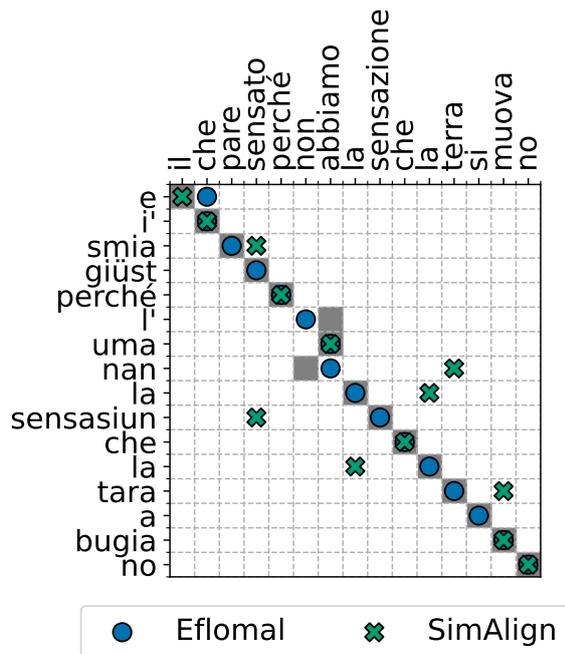


Figure 4: This is a sample alignment. The gray background is the reference alignment, while eflomal alignment is represented by the blue circles and SimAlign one by the green crosses. The English translation of the sentence is *This seems sensible, because the Earth doesn't feel as if it's moving, does it?*

call, F_1 , and alignment error rate (AER) defined as:

$$\text{prec} = \frac{|A \cap P|}{|A|}, \text{rec} = \frac{|A \cap S|}{|S|},$$

$$F_1 = \frac{2\text{prec} \cdot \text{rec}}{\text{prec} + \text{rec}}, \text{AER} = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|}$$

where A are the system alignment, S the sure reference alignment and P the possible reference alignments. However, our annotations do not include possible alignment and so AER is simply $1 - F_1$ (see Appendix E).

The results in Table 2 show that eflomal achieves better scores than SimAlign, which relies on the language model representations. The scores of SimAlign are comparable to those that its authors observed for English-Hindi alignment, indicating that the model produces reasonable alignments despite not being trained on Piedmontese. This indicates that while the XLM-RoBERTa representations are sufficient for generating zero-shot alignment, statistical methods still yield better results.

Additionally, the effect of (sub)words that are identical between Italian and Piedmontese and are therefore easier to align is unknown. We show an

example of alignment in Figure 4, where the reference alignment is mostly monotonic. SimAlign seems to align words in the wrong position (e.g., *la* aligned to the wrong *la*), while eflomal might align words that occur together (e.g., the negation *nan* with the verb *abbiamo*).

4.3 Topic Classification

We use SIB-200 (Adelani et al., 2024) to evaluate the models on topic classification with our data. SIB-200 uses sentences from Flores+, so it is possible to obtain labels for the Piedmontese sentences. The dataset contains 7 classes: *science/technology*, *travel*, *politics*, *sports*, *health*, *entertainment*, and *geography*, but some sentences are labelled as *uncategorized* and are excluded from our experiments. In total, 37 sentences from the *dev* set and 38 from the *devtest* set have a label. We use the same set of sentences for all four languages.

In Figure 5, we report the F_1 scores of the models on the different languages. We note that, while scores for Piedmontese are generally lower than for the other languages, they are still comparable, meaning that models are able to understand the language. Smaller models, such as EuroLLM, exhibit worse performance: in particular, EuroLLM struggles to follow instructions and, in French, often generates all labels or overly long explanations. Tower has a larger drop in performance in Piedmontese, but it is still able to solve the task despite its focus on machine translation. Closed models behave similarly to the larger open-weight models. See Appendix B for the exact values and additional metrics and Appendix A for the prompts used.

4.4 Machine Translation

We test zero-shot machine translation, including both $Pms \rightarrow X$ and $X \rightarrow Pms$. From Figure 6, models have similar chrF++ (\uparrow) scores when translating from the different languages to Piedmontese. Moreover, all languages achieve similar chrF++ scores when translating to Italian, including Piedmontese. However, translating from Piedmontese to French or English is worse than in other languages. While we cannot directly compare the target languages, $X \rightarrow Pms$ has noticeably lower scores.

Given that $X \rightarrow ita$ has comparable results for all languages, we use Italian as a pivot by first translating from the source language to Italian, and then to the target language. From Table 3, the pivot strategy improves translations up to +2.15 chrF++ in the $Pms \rightarrow X$ direction and +1.22 chrF++ in the

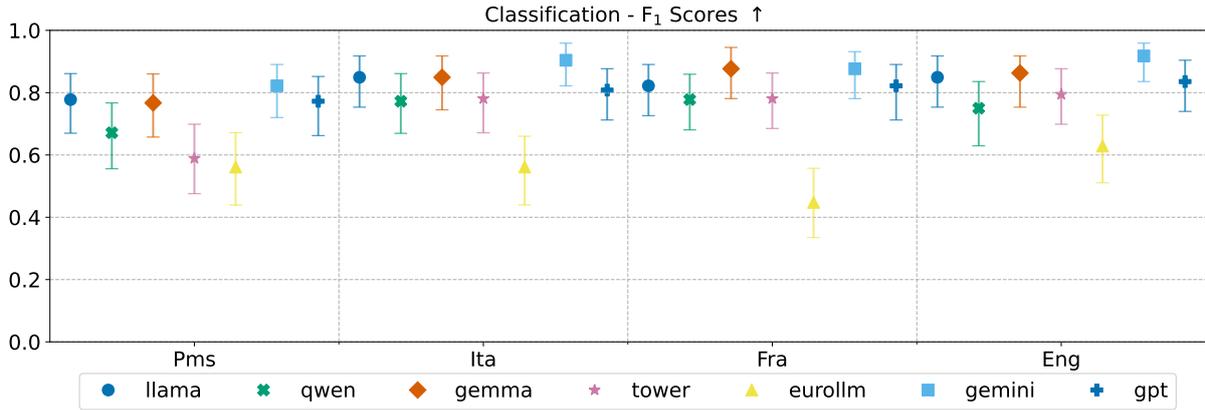


Figure 5: Comparison on the F1 scores of the models in the topic classification task. We perform bootstrapping to compute the confidence interval of the scores.

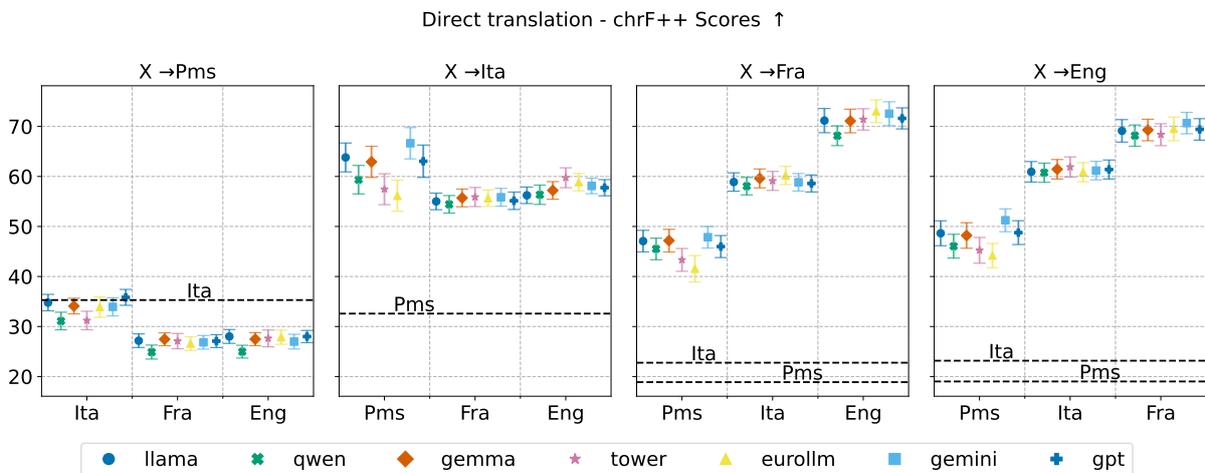


Figure 6: chrF++ scores of the different models. Each subplot shows the target language, while the source languages are on the x-axis. The dotted horizontal lines indicate the scores obtained when using the reference text in a given language as if it were the translation.

$X \rightarrow Pms$ direction.

However, evaluating $X \rightarrow Pms$ is particularly challenging, because models might produce Piedmontese that is correct but different from the reference, which does not use standard orthography, and surface-level metrics such as chrF++ penalize this. Machine-learned metrics like COMET (Rei et al., 2020) can improve this, but they need training data, which is not available. In Appendix C, we report additional metrics, including COMET (without fine-tuning on Piedmontese). Moreover, we observe that the Italian sentences are closer to the Piedmontese references than what the models are generating, as shown by the horizontal lines in Figure 6. This can also explain why LLMs are able to understand Piedmontese. In Table 4, we show some translation examples between Italian and Piedmontese from different models. See Ap-

pendix C for the exact values and additional metrics and Appendix A for the prompts used.

5 Related Work

The data for this work is derived from Flores+ (NLLB Team et al., 2024), which is an evaluation benchmark for machine translation. It contains 2009 sentences, each translated into more than 200 languages. It does not include Piedmontese, but it includes geographically close Italian regional languages, such as Ligurian and Lombard. There are other datasets that contain Piedmontese data, such as Wikipedia (68k samples) and Wikisource (4k samples) (Wikimedia Foundation), and Glot500 (Imani et al., 2023) (226k samples), which derive from the Piedmontese portion of Wikipedia, and Tatoeba (Tiedemann, 2020) (800 samples), which contains sentences annotated by volunteers. How-

Dev		Target				Devtest		Target			
		Pms	Ita	Fra	Eng			Pms	Ita	Fra	Eng
Source	Pms	-	58.80	42.48	45.95	Source	Pms	-	62.15	47.42	47.77
	Ita	32.62	-	57.83	60.98		Ita	33.23	-	60.26	61.37
	Fra	26.45	54.68	-	70.89		Fra	26.60	56.11	-	67.90
	Eng	27.02	57.60	72.25	-		Eng	26.97	57.92	70.61	-
	Pms _{Pivot}	-	-	44.62	46.91		Pms _{Pivot}	-	-	49.57	49.91
	Fra _{Pivot}	27.67	-	-	67.71		Fra _{Pivot}	27.11	-	-	65.11
	Eng _{Pivot}	28.07	-	67.96	-		Eng _{Pivot}	27.85	-	67.63	-

Table 3: Average chrF++ scores of the models on the two sets and all directions. Note that only columns are comparable. *Pivot* refers to the experiments that use Italian as a pivot for the translation.

Flores+ dev 114	
<i>It is the biggest acquisition in eBay's history.</i>	
Pms	A l'è la pì granda aquisission ënt la stòria d'ebay
Ita	Si tratta della maggiore acqui- sizione nella storia di eBay.
<i>Pms→Ita</i>	
EuroLLM	È la più grande acquisizione nella storia di eBay
Gemini	È la più grande acquisizione nella storia di eBay
<i>Ita→Pms</i>	
Gemma	A l'è la pì gròssa aquisission an la stòria d'eBay.
GPT	A l'è la piò gròssa aquisission an sla storia ëd eBay.

Table 4: Translation examples. The *Ita→Pms* translations are understandable, but have different spellings than the reference and across models. The *Pms→Ita* translations are correct, although, phrased differently than the reference translation.

ever, these datasets contain a more standardised version of Piedmontese, which differs from what people might use in real life.

Datasets derived from CommonCrawl¹⁰, like C4 (Raffel et al., 2020), FineWeb2 (Penedo et al., 2025), CulturaX (Nguyen et al., 2024), and Oscar (Ortiz Su'arez et al., 2020), also contain some Piedmontese, but correctly identifying a low-resource language is challenging, and false positives can affect the results.

Another project with the objective of collecting

¹⁰<https://commoncrawl.org/>

language data, including Piedmontese, is AlpiLinK (Rabanus et al., 2023–). AlpiLinK collects crowd-sourced spoken data of various regional languages in the Alpine regions of Italy and contains 5442 Piedmontese sentences.

Ramponi and Casula (2023) propose DIATOPIT, a corpus of social media posts written in different local languages of Italy or using regional Italian. The corpus includes 288 Piedmontese samples and, similarly to our work, does not assume a standard orthography but focuses on the languages as actually written by people.

6 Conclusion

In this paper, we presented a crowdsourced dataset for Piedmontese, whose main characteristic is the non-standard orthography. The dataset can be used for further research on the annotators' demographics, machine translation, and word alignment. Furthermore, we highlight how Piedmontese is at a disadvantage in many popular NLP models, showing that it has higher parity compared to related languages. We then test several LLMs to investigate their performance on topic classification and machine translation tasks. We note that models are able to understand Piedmontese, although they perform worse than in other languages; the scores are still comparable. However, generation still remains challenging.

7 Limitations

This work presents several limitations. Firstly, the selection of annotators is biased because it relies on social media, and people who speak the language may not be accessible. This also influences how the translations are written, because some characters are easier to type on a smartphone keyboard than

on a physical one or with pen and paper. Additionally, we do not track which variant of Piedmontese the annotators use, but we consider Piedmontese to be what the annotators themselves refer to as Piedmontese. Then, the annotators are mostly Italian native speakers, since Italian is the national language, and the number of samples is extremely small. We focus on Piedmontese in Italy and do not consider, for example, Piedmontese spoken in Argentina. The task involves translating from Italian, which can result in translationese. Also, in the questionnaire, we use the terms *orthography* and *grammar* interchangeably to make it easier to understand.

Acknowledgments

This research was supported by the Czech Science Foundation project 25-16242S. The work described herein has also been supported by the Ministry of Education, Youth and Sports of the Czech Republic, Project No. LM2023062 LINDAT/CLARIAH-CZ. We thank the annotators who contributed to this work. GV thanks friends and relatives and the Instagram pages piemontays, Spurgatocn and Abitare il Piemontese for sharing the questionnaire to a larger public.

References

- David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2024. [SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian’s, Malta. Association for Computational Linguistics.
- Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah Smith, and Yulia Tsvetkov. 2023. [Do all languages cost the same? tokenization in the era of commercial language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9904–9923, Singapore. Association for Computational Linguistics.
- 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.
- David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2025. *Ethnologue: Languages of the World*, 28th edition. SIL International, Dallas, TX.
- Gemma Team. 2025. [Gemma 3](#).
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. [Glot500: Scaling multilingual corpora and language models to 500 languages](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.
- 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.
- Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M Alves, José Pombal, Nicolas Boizard, and 1 others. 2025. [Eurollm-9b: Technical report](#). *arXiv preprint arXiv:2506.04079*.
- Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2024. [CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4226–4237, Torino, Italia. ELRA and ICCL.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2024. [Scaling neural machine translation to 200 languages](#). *Nature*, 630(8018):841–846.
- Pedro Javier Ortiz Su’arez, Laurent Romary, and Benoit Sagot. 2020. [A monolingual approach to contextualized word embeddings for mid-resource languages](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Robert Östling and Jörg Tiedemann. 2016. [Efficient word alignment with Markov Chain Monte Carlo](#). *Prague Bulletin of Mathematical Linguistics*, 106:125–146.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. [Fineweb2: One pipeline to scale them all – adapting pre-training data processing to every language](#). *Preprint*, arXiv:2506.20920.

Aleksandar Petrov, Emanuele La Malfa, Philip H.S. Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23*, Red Hook, NY, USA. Curran Associates Inc.

Qwen Team. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.

Stefan Rabanus, Anne Kruijt, Birgit Alber, Ermenegildo Bidese, Livio Gaeta, Gianmario Raimondi, Paolo Benedetto Mas, Sabrina Bertollo, Serena Bissolo, Angelica Bonelli, Dario Capelli, Jan Casalicchio, Raffaele Cioffi, Patrizia Cordin, Michele Cosentino, Silvia Dal Negro, Ilaria Driussi, Alexander Glück, Joachim Kokkelmans, and 10 others. 2023–. [AlpiLinK](#). German-Romance language contact in the Italian Alps: documentation, explanation, participation. <https://alpilink.it>. Ongoing project.

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. *J. Mach. Learn. Res.*, 21(1).

Alan Ramponi and Camilla Casula. 2023. [Diatopit: A corpus of social media posts for the study of diatopic language variation in Italy](#). In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, page 187–199, Dubrovnik, Croatia. Association for Computational Linguistics.

Ricardo Rei, Nuno M Guerreiro, José Pombal, João Alves, Pedro Teixeira, Amin Farajian, and André FT Martins. 2025. Tower+: Bridging generality and translation specialization in multilingual llms. *arXiv preprint arXiv:2506.17080*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. [COMET: A neural framework for MT evaluation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Jörg Tiedemann. 2020. [The tatoeba translation challenge – realistic data sets for low resource and multilingual MT](#). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Wikimedia Foundation. [Wikimedia downloads](#).

A Prompts

Topic classification. The system prompt is "You are a helpful assistant that classifies the following sentence into one of the following categories: science/technology, travel, politics, sports, health, entertainment, geography. Do not add any explanations."

The user prompt is "Is this a piece of news regarding "science, technology, travel, politics, sports, health, entertainment, or geography"? TEXT.", where TEXT is the sentence we are classifying. For Tower we did not use the system prompt.

Machine translation. The system prompt is You are a helpful assistant that translates the following sentence from SRG to TGT. Do not add any explanations.

The user prompt is Translate the following SRC source text to TGT:\nSRC: SENTENCE\nTGT: ". SRG and TGT are the name of the source and target language. SENTENCE is the sentence to translate. For the pivot experiments, the first step translates to Italian, while the second translates from Italian.

B Classification Results

Tables 5 to 8 show the scores on the text classification task.

Model	Metric: F ₁			
	Pms	Ita	Fra	Eng
Llama	0.778 (0.047)	0.849 (0.041)	0.822 (0.044)	0.849 (0.040)
Qwen	0.671 (0.054)	0.772 (0.047)	0.778 (0.046)	0.750 (0.051)
Gemma	0.767 (0.050)	0.849 (0.042)	0.877 (0.037)	0.863 (0.041)
Tower	0.589 (0.057)	0.781 (0.049)	0.781 (0.049)	0.795 (0.046)
EuroLLM	0.579 (0.058)	0.524 (0.058)	0.479 (0.057)	0.671 (0.054)
Gemini	0.822 (0.044)	0.904 (0.034)	0.877 (0.037)	0.918 (0.031)
Gpt	0.772 (0.048)	0.808 (0.045)	0.822 (0.045)	0.836 (0.042)

Table 5: F₁ of the different models on the classification task. In parenthesis the STD of the score.

Model	Metric: Precision			
	Pms	Ita	Fra	Eng
Llama	0.789 (0.047)	0.849 (0.041)	0.822 (0.044)	0.849 (0.040)
Qwen	0.686 (0.054)	0.778 (0.047)	0.789 (0.046)	0.761 (0.051)
Gemma	0.767 (0.050)	0.849 (0.042)	0.877 (0.037)	0.863 (0.041)
Tower	0.589 (0.057)	0.781 (0.049)	0.781 (0.049)	0.795 (0.046)
Eurollm	0.583 (0.058)	0.528 (0.058)	0.479 (0.057)	0.671 (0.054)
Gemini	0.822 (0.044)	0.904 (0.034)	0.877 (0.037)	0.918 (0.031)
Gpt	0.778 (0.048)	0.808 (0.045)	0.822 (0.045)	0.836 (0.042)

Table 6: Precision of the different models on the classification task. In parenthesis the STD of the score.

Model	Metric: Recall			
	Pms	Ita	Fra	Eng
Llama	0.767 (0.048)	0.849 (0.041)	0.822 (0.044)	0.849 (0.040)
Qwen	0.658 (0.054)	0.767 (0.047)	0.767 (0.047)	0.740 (0.052)
Gemma	0.767 (0.050)	0.849 (0.042)	0.877 (0.037)	0.863 (0.041)
Tower	0.589 (0.057)	0.781 (0.049)	0.781 (0.049)	0.795 (0.046)
Eurollm	0.575 (0.058)	0.521 (0.058)	0.479 (0.057)	0.671 (0.054)
Gemini	0.822 (0.044)	0.904 (0.034)	0.877 (0.037)	0.918 (0.031)
Gpt	0.767 (0.048)	0.808 (0.045)	0.822 (0.045)	0.836 (0.042)

Table 7: Recall of the different models on the classification task. In parenthesis the STD of the score.

Model	Metric: Accuracy			
	Pms	Ita	Fra	Eng
Llama	0.767 (0.048)	0.849 (0.041)	0.822 (0.044)	0.849 (0.040)
Qwen	0.658 (0.054)	0.767 (0.047)	0.767 (0.047)	0.740 (0.052)
Gemma	0.767 (0.050)	0.849 (0.042)	0.877 (0.037)	0.863 (0.041)
Tower	0.589 (0.057)	0.781 (0.049)	0.781 (0.049)	0.795 (0.046)
Eurollm	0.575 (0.058)	0.521 (0.058)	0.479 (0.057)	0.671 (0.054)
Gemini	0.822 (0.044)	0.904 (0.034)	0.877 (0.037)	0.918 (0.031)
Gpt	0.767 (0.048)	0.808 (0.045)	0.822 (0.045)	0.836 (0.042)

Table 8: Accuracy of the different models on the classification task. In parenthesis the STD of the score.

C Machine Translation Results

Tables 9 to 20 show the scores for the direct MT task (standard deviation in parenthesis), while Tables 21 to 26 show the scores with pivoting.

Model	Pms → Ita			
	BLEU	chrF++	TER	COMET
Llama	42.910 (3.720)	63.800 (2.890)	44.080 (3.650)	0.933 (0.003)
Qwen	37.490 (3.730)	59.360 (2.850)	48.890 (3.810)	0.932 (0.003)
Gemma	42.090 (4.260)	62.920 (3.110)	45.960 (4.300)	0.935 (0.003)
Tower	34.450 (4.100)	57.450 (3.090)	56.690 (5.850)	0.931 (0.003)
Eurollm	32.290 (3.690)	56.140 (3.110)	56.130 (4.300)	0.909 (0.009)
Gemini	46.290 (4.310)	66.640 (3.130)	42.450 (5.390)	0.937 (0.002)
Gpt	42.290 (4.170)	63.040 (3.230)	44.300 (4.060)	0.932 (0.004)

Table 9: Translation results from Pms to Ita.

Model	Pms → Fra			
	BLEU	chrF++	TER	COMET
Llama	21.700 (2.400)	47.070 (2.160)	70.330 (3.470)	0.908 (0.005)
Qwen	20.420 (2.500)	45.510 (2.160)	71.000 (2.920)	0.905 (0.005)
Gemma	21.550 (2.640)	47.170 (2.270)	70.680 (3.490)	0.908 (0.005)
Tower	17.770 (2.310)	43.320 (2.280)	77.660 (3.690)	0.906 (0.005)
Eurollm	16.670 (2.500)	41.550 (2.630)	79.670 (3.580)	0.901 (0.006)
Gemini	21.550 (2.690)	47.850 (2.160)	71.460 (4.840)	0.908 (0.005)
Gpt	20.670 (2.410)	45.980 (2.200)	69.420 (3.110)	0.906 (0.005)

Table 10: Translation results from Pms to Fra.

Model	Pms → Eng			
	BLEU	chrF++	TER	COMET
Llama	20.270 (2.760)	48.640 (2.490)	71.780 (4.460)	0.901 (0.004)
Qwen	19.180 (2.690)	46.060 (2.380)	74.230 (3.790)	0.900 (0.004)
Gemma	19.970 (2.810)	48.200 (2.550)	73.510 (4.220)	0.900 (0.004)
Tower	17.540 (2.620)	45.230 (2.570)	80.070 (4.900)	0.900 (0.004)
Eurollm	16.080 (2.750)	44.180 (2.420)	82.440 (4.930)	0.894 (0.006)
Gemini	22.800 (2.680)	51.240 (2.280)	68.290 (3.800)	0.899 (0.004)
Gpt	20.940 (2.720)	48.770 (2.380)	71.950 (3.760)	0.899 (0.005)

Table 11: Translation results from Pms to Eng.

Model	Ita → Pms			
	BLEU	chrF++	TER	COMET
Llama	6.980 (1.290)	34.780 (1.640)	83.220 (2.550)	0.842 (0.008)
Qwen	5.250 (1.540)	31.100 (1.760)	93.290 (7.820)	0.843 (0.007)
Gemma	6.270 (1.310)	34.090 (1.560)	84.580 (2.560)	0.841 (0.008)
Tower	4.560 (1.240)	31.200 (1.850)	99.970 (12.440)	0.843 (0.007)
Eurollm	6.250 (1.830)	33.870 (2.020)	88.590 (12.490)	0.844 (0.007)
Gemini	7.280 (1.470)	33.930 (1.820)	85.080 (3.030)	0.833 (0.009)
Gpt	6.800 (1.370)	35.830 (1.590)	82.320 (2.530)	0.848 (0.006)

Table 12: Translation results from Ita to Pms.

Model	Fra → Pms			
	BLEU	chrF++	TER	COMET
Llama	3.000 (1.120)	27.160 (1.360)	95.790 (2.510)	0.835 (0.008)
Qwen	2.410 (1.110)	24.890 (1.400)	101.020 (6.610)	0.838 (0.007)
Gemma	2.960 (0.860)	27.450 (1.290)	94.370 (2.300)	0.831 (0.009)
Tower	2.930 (1.120)	27.080 (1.530)	97.820 (7.040)	0.836 (0.008)
Eurollm	2.810 (1.100)	26.590 (1.370)	95.590 (4.060)	0.836 (0.008)
Gemini	3.350 (0.990)	26.840 (1.360)	96.200 (2.670)	0.829 (0.009)
Gpt	3.280 (1.000)	27.080 (1.290)	95.270 (2.170)	0.835 (0.008)

Table 15: Translation results from Fra to Pms.

Model	Ita → Fra			
	BLEU	chrF++	TER	COMET
Llama	33.500 (2.550)	58.890 (1.840)	54.750 (2.930)	0.943 (0.003)
Qwen	32.380 (2.510)	58.060 (1.750)	55.300 (2.940)	0.943 (0.003)
Gemma	34.260 (2.720)	59.610 (1.890)	53.520 (3.150)	0.942 (0.003)
Tower	34.100 (2.660)	59.110 (1.900)	53.900 (2.900)	0.943 (0.003)
Eurollm	34.700 (2.450)	60.190 (1.810)	53.000 (2.990)	0.944 (0.003)
Gemini	33.100 (2.710)	58.830 (1.750)	55.880 (3.090)	0.943 (0.003)
Gpt	32.400 (2.320)	58.600 (1.700)	55.270 (2.710)	0.942 (0.003)

Table 13: Translation results from Ita to Fra.

Model	Fra → Ita			
	BLEU	chrF++	TER	COMET
Llama	27.890 (2.270)	55.010 (1.660)	58.320 (2.660)	0.952 (0.003)
Qwen	26.810 (2.450)	54.430 (1.750)	59.620 (2.710)	0.952 (0.003)
Gemma	29.270 (2.530)	55.710 (1.770)	57.120 (2.650)	0.952 (0.003)
Tower	29.680 (2.690)	55.890 (1.910)	57.400 (2.800)	0.952 (0.003)
Eurollm	29.200 (2.500)	55.650 (1.660)	57.270 (2.670)	0.952 (0.003)
Gemini	29.520 (2.650)	55.840 (1.790)	57.980 (3.020)	0.952 (0.003)
Gpt	27.680 (2.390)	55.140 (1.740)	57.640 (2.700)	0.951 (0.003)

Table 16: Translation results from Fra to Ita.

Model	Ita → Eng			
	BLEU	chrF++	TER	COMET
Llama	32.520 (2.910)	60.930 (2.050)	52.970 (3.430)	0.941 (0.003)
Qwen	32.050 (2.790)	60.750 (1.920)	52.360 (3.110)	0.942 (0.003)
Gemma	33.410 (2.860)	61.440 (1.950)	51.580 (3.100)	0.941 (0.003)
Tower	33.900 (2.800)	61.890 (2.020)	51.610 (3.160)	0.942 (0.003)
Eurollm	32.970 (2.810)	60.850 (1.920)	52.460 (3.140)	0.941 (0.003)
Gemini	32.400 (2.670)	61.170 (1.850)	53.240 (3.350)	0.941 (0.003)
Gpt	32.730 (2.620)	61.390 (1.900)	52.430 (3.090)	0.941 (0.003)

Table 14: Translation results from Ita to Eng.

Model	Fra → Eng			
	BLEU	chrF++	TER	COMET
Llama	46.510 (3.530)	69.120 (2.250)	37.150 (3.270)	0.947 (0.002)
Qwen	45.320 (3.510)	68.170 (2.130)	38.340 (3.400)	0.947 (0.002)
Gemma	46.410 (3.390)	69.290 (2.150)	36.160 (3.170)	0.947 (0.002)
Tower	45.030 (3.330)	68.350 (2.190)	37.830 (3.170)	0.945 (0.002)
Eurollm	47.380 (3.810)	69.520 (2.370)	36.880 (3.480)	0.945 (0.002)
Gemini	48.790 (3.640)	70.670 (2.140)	34.530 (3.290)	0.947 (0.002)
Gpt	46.460 (3.550)	69.410 (2.150)	36.770 (3.100)	0.947 (0.002)

Table 17: Translation results from Fra to Eng.

Model	Eng → Pms			
	BLEU	chrF++	TER	COMET
Llama	3.170 (1.140)	28.010 (1.390)	94.430 (2.370)	0.824 (0.009)
Qwen	2.340 (1.200)	24.970 (1.280)	98.110 (3.870)	0.828 (0.008)
Gemma	2.460 (0.840)	27.470 (1.300)	94.370 (2.370)	0.825 (0.008)
Tower	3.170 (1.140)	27.620 (1.680)	99.740 (9.360)	0.830 (0.008)
Eurollm	3.050 (1.070)	27.860 (1.420)	94.830 (6.580)	0.828 (0.008)
Gemini	3.130 (0.890)	26.980 (1.470)	95.730 (4.030)	0.818 (0.010)
Gpt	2.290 (0.920)	27.990 (1.220)	92.130 (1.940)	0.830 (0.007)

Table 18: Translation results from Eng to Pms.

Model	Eng → Ita			
	BLEU	chrF++	TER	COMET
Llama	29.850 (2.110)	56.230 (1.650)	54.900 (2.390)	0.953 (0.003)
Qwen	29.920 (2.560)	56.340 (1.930)	56.200 (2.790)	0.952 (0.003)
Gemma	30.770 (2.510)	57.210 (1.760)	54.380 (2.430)	0.953 (0.003)
Tower	34.510 (3.070)	59.740 (1.990)	51.570 (2.610)	0.953 (0.003)
Eurollm	33.590 (2.530)	58.860 (1.740)	52.000 (2.500)	0.953 (0.003)
Gemini	31.660 (2.320)	58.100 (1.550)	53.300 (2.430)	0.953 (0.003)
Gpt	31.240 (2.340)	57.750 (1.620)	52.740 (2.320)	0.952 (0.003)

Table 19: Translation results from Eng to Ita.

Model	Eng → Fra			
	BLEU	chrF++	TER	COMET
Llama	52.800 (3.640)	71.170 (2.420)	33.280 (2.970)	0.949 (0.002)
Qwen	48.010 (3.130)	68.150 (1.980)	38.380 (2.820)	0.948 (0.002)
Gemma	52.420 (3.510)	71.090 (2.360)	33.750 (2.920)	0.947 (0.003)
Tower	53.020 (3.360)	71.410 (2.130)	34.360 (2.800)	0.949 (0.002)
Eurollm	55.550 (3.680)	73.030 (2.280)	32.350 (3.040)	0.950 (0.002)
Gemini	55.030 (3.690)	72.550 (2.380)	32.150 (3.060)	0.945 (0.003)
Gpt	53.010 (3.260)	71.610 (2.110)	33.280 (2.890)	0.950 (0.002)

Table 20: Translation results from Eng to Fra.

Model	Pms → Fra, Pivot			
	BLEU	chrF++	TER	COMET
Llama	23.580 (2.450)	49.090 (2.220)	68.000 (3.290)	0.927 (0.002)
Qwen	19.890 (2.460)	45.140 (2.180)	72.630 (3.190)	0.927 (0.002)
Gemma	23.740 (2.820)	49.450 (2.290)	68.140 (3.570)	0.927 (0.002)
Tower	20.380 (2.460)	46.070 (2.150)	73.850 (3.860)	0.927 (0.002)
Eurollm	20.210 (2.850)	44.900 (2.540)	74.080 (3.840)	0.927 (0.002)
Gemini	25.600 (2.410)	50.780 (2.150)	64.740 (2.930)	0.927 (0.002)
Gpt	23.930 (2.270)	49.320 (2.090)	66.070 (2.940)	0.927 (0.002)

Table 21: Translation results with pivoting from Pms to Fra.

Model	Pms → Eng, Pivot			
	BLEU	chrF++	TER	COMET
Llama	21.780 (2.910)	49.550 (2.450)	68.830 (3.870)	0.917 (0.002)
Qwen	19.340 (2.730)	47.160 (2.400)	70.800 (3.610)	0.917 (0.002)
Gemma	23.630 (3.050)	50.770 (2.500)	66.010 (4.070)	0.917 (0.002)
Tower	18.650 (2.740)	47.080 (2.430)	75.450 (4.590)	0.917 (0.002)
Eurollm	18.140 (2.940)	45.500 (2.570)	77.720 (4.310)	0.917 (0.002)
Gemini	25.320 (2.730)	53.280 (2.360)	64.010 (3.740)	0.917 (0.002)
Gpt	23.490 (2.950)	50.760 (2.440)	66.860 (3.950)	0.917 (0.002)

Table 22: Translation results with pivoting from Pms to Eng.

Model	Fra → Pms, Pivot			
	BLEU	chrF++	TER	COMET
Llama	3.990 (1.180)	28.780 (1.430)	92.600 (2.440)	0.836 (0.008)
Qwen	2.720 (1.120)	26.220 (1.430)	99.710 (7.940)	0.838 (0.007)
Gemma	3.760 (1.060)	28.640 (1.430)	92.540 (2.410)	0.834 (0.008)
Tower	2.460 (1.040)	26.440 (1.590)	107.140 (15.430)	0.837 (0.008)
Eurollm	2.870 (1.060)	26.960 (1.480)	95.120 (5.920)	0.835 (0.008)
Gemini	3.880 (1.110)	28.000 (1.550)	94.160 (2.800)	0.829 (0.009)
Gpt	3.580 (1.080)	29.110 (1.400)	93.090 (2.280)	0.838 (0.007)

Table 23: Translation results with pivoting from Fra to Pms.

Model	Fra → Eng, Pivot			
	BLEU	chrF++	TER	COMET
Llama	42.730 (3.390)	66.800 (2.190)	39.970 (3.580)	0.952 (0.001)
Qwen	42.940 (3.490)	65.990 (2.240)	40.340 (3.410)	0.952 (0.001)
Gemma	42.950 (3.260)	66.630 (2.020)	39.080 (3.130)	0.952 (0.001)
Tower	40.260 (3.210)	64.990 (2.100)	42.240 (3.450)	0.952 (0.001)
Eurollm	41.720 (3.450)	65.570 (2.250)	42.000 (3.750)	0.952 (0.001)
Gemini	43.330 (3.650)	67.260 (2.150)	39.590 (3.400)	0.952 (0.001)
Gpt	42.820 (3.490)	66.820 (2.320)	39.970 (3.490)	0.952 (0.001)

Table 24: Translation results with pivoting from Fra to Eng.

Model	Eng → Pms, Pivot			
	BLEU	chrF++	TER	COMET
Llama	3.770 (1.050)	28.910 (1.410)	92.680 (2.490)	0.826 (0.008)
Qwen	2.200 (0.980)	26.200 (1.370)	96.920 (3.500)	0.828 (0.008)
Gemma	3.560 (1.040)	29.210 (1.420)	91.640 (2.340)	0.826 (0.008)
Tower	2.850 (1.110)	27.250 (1.710)	104.880 (11.890)	0.828 (0.008)
Eurollm	3.250 (1.180)	28.190 (1.410)	91.060 (1.830)	0.827 (0.008)
Gemini	3.970 (1.170)	28.850 (1.570)	92.650 (2.720)	0.822 (0.009)
Gpt	3.650 (1.100)	29.600 (1.380)	90.910 (2.120)	0.828 (0.008)

Table 25: Translation results with pivoting from Eng to Pms.

Model	Eng → Fra, Pivot			
	BLEU	chrF++	TER	COMET
Llama	47.630 (3.130)	67.830 (2.170)	37.390 (2.810)	0.956 (0.001)
Qwen	42.230 (2.970)	64.330 (1.970)	42.110 (2.750)	0.956 (0.001)
Gemma	47.460 (3.390)	67.640 (2.240)	38.090 (2.950)	0.956 (0.001)
Tower	47.240 (3.190)	67.840 (2.090)	38.470 (2.730)	0.956 (0.001)
Eurollm	48.710 (3.400)	68.500 (2.310)	37.040 (3.090)	0.956 (0.001)
Gemini	50.750 (3.590)	69.800 (2.440)	35.350 (3.080)	0.956 (0.001)
Gpt	48.130 (3.350)	68.090 (2.330)	37.010 (3.030)	0.956 (0.001)

Table 26: Translation results with pivoting from Eng to Fra.

D Parity Results

Tables 27 to 30 show the tokenizer parity scores with respect to the different languages. Note that scores with respect to different languages are not comparable.

Model	Parity w.r.t Pms		
	Ita	Fra	Eng
Llama	0.905	0.898	0.569
Qwen	0.909	0.897	0.576
Gemma	0.816	0.850	0.609
Tower	0.806	0.830	0.616
Eurollm	0.784	0.858	0.639
Bpe	0.783	0.847	0.687
Unigram	0.752	0.817	0.664
Gemini	0.820	0.853	0.618
Gpt	0.883	0.833	0.618

Table 27: Parity scores with respect to Piedmontese.

Model	Parity w.r.t Ita		
	Pms	Fra	Eng
Llama	1.140	0.998	0.634
Qwen	1.136	0.993	0.638
Gemma	1.267	1.041	0.749
Tower	1.284	1.033	0.767
Eurollm	1.324	1.097	0.818
Bpe	1.320	1.088	0.881
Unigram	1.373	1.090	0.887
Gemini	1.258	1.040	0.757
Gpt	1.162	0.945	0.703

Table 28: Parity scores with respect to Italian.

Model	Parity w.r.t Fra		
	Pms	Ita	Eng
Llama	1.164	1.028	0.640
Qwen	1.167	1.034	0.649
Gemma	1.243	0.983	0.726
Tower	1.269	0.993	0.751
Eurollm	1.232	0.935	0.753
Bpe	1.235	0.941	0.816
Unigram	1.286	0.941	0.823
Gemini	1.235	0.984	0.734
Gpt	1.258	1.086	0.751

Table 29: Parity scores with respect to French.

Model	Parity w.r.t Eng		
	Pms	Ita	Fra
Llama	1.848	1.633	1.602
Qwen	1.831	1.622	1.584
Gemma	1.740	1.380	1.416
Tower	1.721	1.347	1.371
Eurollm	1.653	1.255	1.357
Bpe	1.532	1.166	1.250
Unigram	1.581	1.158	1.245
Gemini	1.707	1.363	1.398
Gpt	1.699	1.470	1.366

Table 30: Parity scores with respect to English.

E Alignment Metrics

We do not use possible reference alignments, so $|P| = |S|$. Assuming that $|A \cap S|$, $|A|$, and $|S|$ are not empty, F_1 can be rewritten as:

$$\begin{aligned}
 F_1 &= \frac{2\text{prec} \cdot \text{rec}}{\text{prec} + \text{rec}} = \frac{2 \frac{|A \cap P|}{|A|} \cdot \frac{|A \cap S|}{|S|}}{\frac{|A \cap P|}{|A|} + \frac{|A \cap S|}{|S|}} = \\
 &= \frac{2 \frac{|A \cap S|}{|A|} \cdot \frac{|A \cap S|}{|S|}}{\frac{|A \cap S|}{|A|} + \frac{|A \cap S|}{|S|}} = \\
 &= \frac{2|A \cap S|}{|A||S|} \cdot \frac{|A||S|}{|A| + |S|} = \frac{2|A \cap S|}{|S| + |A|}
 \end{aligned}$$

And AER as:

$$\text{AER} = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|} = 1 - \frac{2|A \cap S|}{|S| + |A|}$$

Therefore $\text{AER} = 1 - F_1$

F Hyper-parameters

For the translation task and the classification, we use greedy decoding and we generate at most 100 tokens, which is sufficient for the ground-truth labels. The closed source models do not use reasoning. The open-weight model are used with the Transformers 4.57.1 text generation pipeline, while the closed models are used through OpenRouter.

We run eflomal (version 2.0.0) with its default parameters, then we symmetrize the alignment with fast align atools with *grow-diag-final-and*. SimAlign is run with the following arguments:

- Model: xlm-roberta-base
- Tokenizer type: bpe
- Distortion: 0

- Layer: 8
- Matching method: itermax

G Computational Resources and Costs

The total cost for Gemini was \$0.62 and \$0.17 for GPT. The provider has a zero data retention policy. The other experiments were run on a local cluster with up to 2 NVIDIA H100 with 95GB VRAM and 60GB RAM.

H Questionnaire

The questionnaire is in Italian. Here we show the original version and in brackets the English translation. The questionnaire introduction explains to the user the goal of the project and emphasizes that it is not about evaluating the user and that it is anonymous.

Quale lingua utilizzi di più quotidianamente (scuola, lavoro, in giro, ecc.)?

[Which language do you use most on a daily basis?]

Seleziona una lingua dall'elenco. Se scegli "Altro", specifica la lingua nel campo di testo.

[Select a language from the list. If you choose "Other," specify the language in the text field.]

{Seleziona una lingua: Italiano, Francese, Spagnolo, Tedesco, Inglese, Rumeno, Arabo, Macedone, Albanese, Piemontese, Preferisco non rispondere, Altro}

{Select a language: Italian, French, Spanish, German, English, Romanian, Arabic, Macedonian, Albanian, Piedmontese, Prefer not to reply, Other}

Quanto bene parli il piemontese?

[How well do you speak Piedmontese]

Seleziona una delle opzioni che meglio descrive la tua conoscenza del piemontese.

[Select one of the options that best describes your knowledge of Piedmontese.]

- Niente o quasi, solo qualche parola [Nothing or almost nothing, just a few words]
- Poco, conosco alcune espressioni, ma faccio fatica a esprimere frasi nuove [Not much, I know some expressions, but I struggle to express new sentences]
- Abbastanza, ma a volte lo mischio con l'italiano (o la lingua che uso principalmente) [Quite a bit, but sometimes I mix it with Italian (or whatever language I use mostly)]
- Perfettamente o quasi, riesco a esprimere praticamente tutto [Perfectly or almost perfectly, I can express practically everything]

Secondo te, il piemontese ha una grammatica e ortografia ben definita ("questa parola si scrive così", "questo verbo si coniuga così")?

[In your opinion, does Piedmontese have a well-defined grammar and spelling ("this word is written like this", "this verb is conjugated like this")?]

Seleziona una delle opzioni che meglio descrive la tua opinione.

[Select one of the options that best describes your opinion.]

{D'accordo, Neutrale, In disaccordo}

Quando le persone scrivono in piemontese usano questa grammatica?

[When people write in Piedmontese, do they use this grammar?]

Seleziona una delle opzioni che meglio descrive la tua opinione.

[Select one of the options that best describes your opinion.]

{D'accordo, Neutrale, In disaccordo}

Da chi hai imparato il piemontese?

[Where did you learn Piedmontese from?]

Puoi selezionare più opzioni. Se selezioni "Altro", puoi specificare.

[You can select multiple options. If you select "Other," you can specify.]

- Nonni [Grand parents]
- Genitori [Parents]
- Parenti [Relatives]
- Amici o colleghi [Friends or colleagues]
- Altro [Other]

Qual è la tua fascia d'età?

[What is your age range?]

Fai 30 anni tra 4 giorni? Seleziona "Tra 20 e 30" — Hai compiuto 40 l'altro ieri? Seleziona "Tra 40 e 50"

[Are you turning 30 in 4 days? Select "Between 20 and 30." — Did you turn 40 the day before yesterday? Select "Between 40 and 50."]

- Meno di 20 [Less than 20]
- Tra 20 e 30 [Between 20 and 30]

- Tra 30 e 40 [*Between 30 and 40*]
- Tra 40 e 50 [*Between 40 and 50*]
- Tra 50 e 60 [*Between 50 and 60*]
- Più di 60 [*More than 60*]
- Preferisco non rispondere [*I prefer not to answer*]

TRADUZIONE

[*Translation*]

Come scriveresti questa frase in piemontese?

[*How would you write this sentence in Piedmontese?*]

Linee guida:

- Se non sai come tradurla non scrivere nulla. Se vuoi puoi riprovare e il questionario dovrebbe proporre una frase casuale diversa. [*If you don't know how to translate it, don't write anything. If you want, you can try again and the questionnaire should suggest a different random sentence.*]
- Non usare traduttori automatici (Google Translate, ecc.). [*Do not use automatic translators (Google Translate, etc.).*]
- Non aggiungere spiegazioni (no "la traduzione è:", "... (vuole anche dire ...)") o diverse traduzioni possibili (no: "... (che vuole dire ...)", "... opzione 1/opzione 2 ..."). [*Do not add explanations (no "the translation is:", "... (also means ...)") or multiple possible translations (no: "... (which means ...)", "... option 1/option 2 ...").*]
- Può essere che alcune parole siano difficilmente traducibili. Scrivile come le scriveresti tu. [*Some words may be difficult to translate. Write them as you would.*]
- Puoi chiedere aiuto ai nonni. [*You can ask your grandparents for help.*]
- Accentuati e simboliche magari non sono sulla tastiera ('a' come esempio). Da telefono puoi tenere premuta una lettera per vedere le opzioni disponibili. [*Accents and symbols may not be on the keyboard ('a' as an example). On your phone, you can press and hold a letter to see the available options.*]

à: /'a, á: /"a, â: /^a, ã: ~a, ä: /:a, å: /,a, â: /.a, å: /°a, ä: /=a, ø: //o

In italiano [*In Italian*]

Sample

In piemontese [*In Piedmonetese*]

Text field

VALUTAZIONE

[*Evaluation*]

Come valuteresti la seguente traduzione?

[*How would you rate the following translation?*] Considera possibili variazioni del piemontese (ad esempio di qualcuno di Torino o di Verduno). La traduzione è stata fatta da un'altro utente e presentata senza alcuna modifica.

[*Consider possible variations in Piedmontese (for example, someone from Turin or Verduno). The translation was done by another user and presented unchanged.*]

In italiano [*In Italian*]

Sample

In piemontese [*In Piedmonetese*]

Sample

- Interamente corretta o quasi [*Completely correct or almost*]

- Probabilmente corretta, l'avrei scritta in altro modo [*Probably correct, I would have written it differently*]
- Parzialmente corretta [*Partially correct*]
- Totalmente sbagliata o quasi [*Totally wrong or almost*]
- Non lo so [*I do not know*]
- Risposta mancante, offensiva o non pertinente [*Missing, offensive or irrelevant response*]

German-English Code-Switching in Large Language Models

Firat Cem Aksüt¹, Stefan Hillmann¹, Pia Knoeferle², Sebastian Möller¹

¹Technische Universität Berlin, ²Humboldt-Universität zu Berlin,

Correspondence: aksuet@campus.tu-berlin.de, stefan.hillmann@tu-berlin.de

Abstract

Code-Switching (CS) is common in multilingual communication, yet it is unclear how well current Large Language Models (LLMs) reproduce naturally occurring switching patterns. This paper studies German–English CS (“Denglisch”) generated by GPT-4o-mini and LLaMA-3.3, using Reddit data from the Denglisch Corpus as a reference. Model outputs are compared to authentic posts using established CS metrics (M-Index, I-Index, CE-SAR), an analysis of Shared Lexical Items (SLIs) as switch triggers, and a human evaluation of Perceived Naturalness and Perceived Fluency. Both models approximate global CS characteristics but differ in the diversity and complexity in comparison to real data. LLaMA-3.3 more closely matches corpus-level metrics, whereas GPT-4o-mini produces more conservative switching that is rated as significantly more natural and fluent. In addition, GPT-4o-mini reproduces SLI-triggered switching patterns similar to those found in authentic data, while this effect is weaker for LLaMA-3.3.

1 Introduction

Large Language Models (LLMs) such as GPT-4 are increasingly being used in everyday applications, such as chatbots, translation systems and writing assistants. Since the release of ChatGPT in 2022, they have quickly become widespread (Hu, 2023). At the same time, multilingualism is commonplace in digital communication spaces. According to estimates, around 3.5 billion people speak at least two languages (Grosjean, 2021). A key phenomenon here is Code-Switching (CS), i.e. switching between two or more languages within a conversation, sentence or even a word (Riehl, 2019a). CS fulfils both linguistic economy and identity-related functions, especially in informal contexts and social media (Riehl, 2019b).

Despite their strong multilingual capabilities, there has been limited research into how well mod-

ern LLMs actually replicate such naturally occurring CS patterns. Studies show that even large multilingual LLMs often lag behind specialised systems in CS-related tasks and are ‘not (yet) Code-Switchers’ (Zhang et al., 2023). At the same time, linguistic studies suggest that so-called *Shared Lexical Items (SLIs)*, i.e. words that occur in both languages, can trigger language switching in a targeted manner by increasing the probability of a switch in their environment (triggering hypothesis) (Broersma and de BOT, 2006).

Within this context, this paper examines the extent to which current LLMs can realistically imitate German-English CS from Reddit posts. Based on the Denglisch Corpus (Osmelak and Wintner, 2023), we have two models (GPT-4o-mini and LLaMA-3.3) generate new Denglisch posts and compare them with authentic posts using established CS metrics, an analysis of SLIs and a human evaluation.

Specifically, this paper addresses the following research questions:

- **Research Question (RQ) 1:** To what extent can current LLMs reproduce the global CS patterns of Denglisch Reddit posts?
- **Research Question (RQ) 2:** Do LLMs exhibit SLI-triggered CS patterns similar observed to those in real Denglisch data?

2 Background

2.1 Code-Switching

Code-switching (CS) generally refers to the switch between two (or more) languages within a single utterance, sentence, or conversation (Riehl, 2019a; Muysken, 2011). The matrix language, which determines the grammatical structure and the main part of the expression, is supplemented by the embedded language, which inserts individual words, phrases, or expressions from another language

(Myers-Scotton, 2017). The phenomenon typically occurs among bilingual speakers and multilingual communities and requires a corresponding level of linguistic competence. Typically, the grammatical structures of the languages are not violated. CS is influenced by sociocultural, situational, or contextual factors. Poplack (1980) and Kootstra (2015) provide a psycholinguistic view on it.

The literature distinguishes between different types of CS. Traditionally, a distinction is made between **intra-sentential CS**, **inter-sentential CS** and **tag switching** (Poplack, 1980). In intra-sentential CS, the language switch occurs within a sentence, typically at syntactically compatible boundaries. In inter-sentential CS, the switch occurs between sentences or larger units of speech. Tag switching, on the other hand, refers to the insertion of short fixed expressions (e.g. discourse markers) from another language. In addition, **intra-word CS** (Myers-Scotton, 1989) is often described, in which the switch occurs within a word (e.g. through the combination of morphemes).

2.2 Code-Switching Metrics

In order to quantify CS, there are a number of established measures that capture different aspects of language switching. The metrics used in this study are primarily the M-Index, I-Index and CSR.

The normalized (Chi et al., 2024) **M-Index** (Mixing Index) (Barnett et al., 1999, p. 202) measures the balance of the languages involved based on relative token frequency. Values close to 0 indicate a strong dominance of one language, while values close to 1 indicate an even distribution.

The **I-Index** (Integration-Index) quantifies the density of switch points in the text (Guzmán et al., 2017). It measures how often two neighbouring tokens are assigned to different languages. A value of 0 means no language switch, while a value of 1 means a switch occurs at every token.

The **CSR** (Code-Switching Rate) corresponds to a simplified form of the I-Index. It describes the proportion of token boundaries at which a language change occurs and can be interpreted as the probability of a change at a random position.

Finally, **CESAR** measures the degree of CS relative to a reference language. It combines the occurrence, i.e. the proportion of tokens that contain tokens from another language, with the burstiness, i.e. how strongly these foreign-language tokens occur in clusters within the units (Abidi and Smaïli, 2021). A value of 0 means that the text is exclu-

sively in the reference language, while a value of 1 means that it does not occur at all.

2.3 Shared Lexical Items

The *triggering hypothesis* (Clyne, 2003) assumes that certain lexical elements that occur in both languages of a speaker’s repertoire increase the probability of a language switch. These ‘shared lexical items’ (SLIs) include proper names, culturally specific terms, and terms for which there is no direct translation. The presence of such a trigger in an utterance appears to cause the brain to switch from one language to another (BROERSMA and DE BOT, 2006).

Wintner et al. (2023) confirm this assumption in a large-scale analysis of several language pairs. They have demonstrated that the probability of a switch increases significantly when an SLI appears shortly before the switch and is slightly lower when it occurs afterwards. They distinguish SLIs according to origin, such as shared English, shared German or shared other. These findings form the basis for the present study, in which we examine whether modern LLMs show similar SLI-triggered patterns in German-English CS.

2.4 LLMs and Code-Switching

Although modern LLMs have strong multilingual capabilities, they still show limitations in CS. Studies show that even large models often lag behind specialised systems in CS-related tasks (Zhang et al., 2023). Challenges arise in particular from correct token-level language identification and other token-level processing, dealing with ambiguous word forms, and adhering to grammatical CS patterns (Çetinoğlu et al., 2016). For the German-English language pair in particular, both CS research in general and work on LLM behaviour under CS are still relatively underrepresented and only few annotated resources are available. Osmelak and Wintner (2023) address this gap by introducing the Denglisch Corpus, a German-English CS corpus that serves as an empirical basis for our study.

3 Data and Methodology

3.1 Denglisch Corpus

For our experiments, we use the Denglisch Corpus of Osmelak and Wintner (2023). This is an extensive corpus of German-English CS posts from Reddit. The corpus was created specifically to reflect

naturally occurring Denglisch in informal online texts and contains around 31,500 posts. It contains various forms of CS, such as intra/inter-sentential, tag-based and hybrid forms of CS. The posts come from different subreddits and cover a wide range of topics, making the corpus a representative basis for the analysis of mixed-language online communication.

To generate and compare the LLM responses, we take a random sample of 50 posts (Input-Data) that represent the entire corpus as well as possible in terms of length, CS characteristics and stylistic variation. There are several reasons for limiting the sample to 50 examples. On the one hand, processing very large amounts of data can lead to instability (Bender et al., 2021). At the same time, a small, carefully selected input allows for a more precise analysis of model behaviour, especially with regard to few-shot scenarios.

3.2 Annotation

Osmelak and Wintner (2023) also provide a fine-grained annotation scheme and word-level classifier as part of their work. In addition to clear language labels (German/English), the scheme also includes categories for loanwords, ambiguous forms and mixed constructions, which enables precise identification of CS points. Part of the corpus was annotated manually and used to develop a classifier, which was then applied to the complete data set. They report an average tagging accuracy of 0.764 at the sentence level (i.e., across all tags within a sentence), and, for instance, F_1 scores of 0.96 and 0.97 for German and English tags, respectively (Osmelak and Wintner, 2023, Table 5).

For our experiments, we reuse this classifier to consistently annotate all LLM-generated texts. To make the labels comparable across corpus and model outputs, we map the original fine-grained tags onto a compact tagset with four language categories. All German-specific labels are collapsed into a single GERMAN class, all English-specific labels into ENGLISH, mixed labels (e.g. hybrid forms) into MIX, and discourse marker, interjections and similar neutral items into a TAG/OTHER class. In addition, we exploit the original shared-item labels of the classifier. All tokens tagged as shared (e.g. shared English, shared German, shared other) are marked with an SLI flag and their origin. This annotation allows us to automatically compute all CS metrics and to identify both different CS types and SLI-triggered switches across the

three sources.

3.3 Models

We compare the GPT-4o-mini and LLaMA-3.3-70B-Instruct (LLaMA-3.3) models. The selection is based on criteria for CS evaluation scenarios. These include multilingualism, model size, accessibility, and usage costs. According to Zhang et al. (2023), also the MMLU¹ benchmark is suitable for this purpose although it is not explicitly designed as a CS benchmark.

While GPT-4o-mini is a proprietary API model with paid token rates and an MMLU score of 82 % (OpenAI, 2024), LLaMA-3.3 is a freely usable open-source model that can be run locally or in cloud environments and an MMLU score of 86 % (Hugging Face, 2024). Furthermore, the models differ in parameter count (GPT-4o-mini: 8B²; LLaMA-3.3: 70B) and training methods. Since model size and training regime have been shown to influence the quality of generative tasks (Brown et al., 2020), this model comparison provides cross-validation of the CS patterns across parameter and training differences.

3.4 Generation Setup

To generate the LLM outputs, we use a standardised prompt design that is identical for both models. The prompts given to the LLMs and discussed here are provided in Appendix A. Each of the 50 input posts is given to the respective models together with a brief introduction to CS. The respective CSR is also provided with each of the sample posts so that the models can recognise the typical CS intensity of the data. Our pilot tests with and without this additional information showed a positive affect on the models' performance regarding CS generation. The structure of the prompt follows the few-shot principle, which has proven effective for controlling linguistic generation tasks (Brown et al., 2020).

In the instructions, the models are asked to generate a 3-sentence Reddit post that contains natural German-English CS and varies in content and style. In addition, the generated post should have a CSR close to the median value of the examples provided.

¹MMLU (Massive Multitask Language Understanding) is a multiple-choice test that evaluates LLMs in 57 different subject areas to measure their cross-domain knowledge coverage and problem-solving skills.

²A pre-print article (<https://arxiv.org/abs/2412.19260>) discloses the parameter size (8B), the final article does not include absolute numbers (Ben Abacha et al., 2025).

To promote stylistic diversity, we use a temperature of 0.9. Lower values led to predominantly monolingual or repetitive responses in pilot runs. All generations are performed under identical conditions for both models. The outputs are then saved and further processed for annotation and metric calculation. This controlled setup mitigates prompt bias and found differences in the CS patterns are can be attributed to the models themselves.

3.5 Human Evaluation Setup

To supplement the automatic metrics, we conducted a human evaluation to assess the Perceived Naturalness and Perceived Fluency of the posts. The survey was implemented via a crowdsourcing platform and targeted participants located in the German-speaking DACH region (Germany, Austria, Switzerland). Eligibility criteria required participants to have German as a (co-)native language and to report at least intermediate proficiency in English, ensuring that they were able to understand and evaluate German–English CS posts.

After excluding two participants based on a control question, we could include responses of 133 individuals in our analysis. The age distribution is predominantly in the 25–44 age range. Younger (under 18) and older participants (over 65) are clearly under-represented. The vast majority of participants stated German as their native language ($\approx 91\%$), with only a few naming English or other languages as their native language. Among the German-speaking participants, more than three-quarters rated themselves as ‘advanced’ or ‘fluent’ in English, suggesting a high level of bilingual competence and good conditions for assessing German–English CS patterns.

Each participant completed one questionnaire containing five Reddit posts with German–English CS, randomly sampled from the three sources GPT-4o-mini, LLaMA-3.3, and the Input-Data. In total, 14 posts per source (42 posts overall) were included in the evaluation, each of which was rated 15 times, resulting in 630 individual ratings. The posts were randomly assigned to questionnaire versions and randomised in order to reduce position and anchoring effects. The exact wording of all questions and items discussed here and in the following are provided in Appendix B.

The central scales for the assessment were Perceived Naturalness and Fluency. Each of them were measured on a five-point Likert scale ranging from ‘very unnatural/unsmooth’ (1) to ‘very

natural/smooth’ (5). After the naturalness rating, participants who selected ‘very unnatural’ or ‘rather unnatural’ were presented with a follow-up question asking to select from pre-defined reasons why the post seemed unnatural (e.g. unnatural mixture of German and English, unusual word choice, grammar errors, lack of natural flow). Conversely, participants who selected ‘very natural’ or ‘rather natural’ received an analogous question asking for reasons why the post seemed natural (e.g. authentic language switching, natural mixture of German and English, typical Reddit phrases, natural word choice). If ‘neutral’ was selected, no follow-up question on reasons was shown.

In addition, the questionnaire collected demographic information (age group, native language, self-assessed proficiency in the other language) as well as self-reported usage patterns. This means, how often participants read and post in online forums such as Reddit, and how often they themselves use CS in online texts. These variables serve to contextualise the ratings and to identify potential factors that might influence the perception of CS.

The collected data was then evaluated descriptively. We compared the mean values of the evaluations across models and corpora and examined whether the Perceived Naturalness or Fluency correlated with objective CS metrics (CSR, M-Index, I-Index, CESAR).

4 Quantitative Results

4.1 Global CS Patterns

In a first step, we examine the global properties of the CS patterns generated by the models using the three metrics presented (see subsection 2.2). The M-Index describes the linguistic diversity within a post, the I-Index describes the frequency of language changes, and CESAR describes the complexity and burstiness of the CS patterns.

As shown in Figure 1, LLaMA-3.3 achieves higher mean values than GPT-4o-mini (0.247 vs. 0.196) in the M-Index, which indicates greater linguistic diversity in the LLaMA-3.3 responses. However, the Input-Data show higher values (0.369), meaning that none of the models reach the level of real Denglish posts. A similar picture emerges for the I-Index. With a mean of 0.165, LLaMA-3.3 is clearly above GPT-4o-mini (0.099) and the difference to the Input-Data (0.174) is remarkable smaller. LLaMA-3.3 thus tends towards more frequent language changes and a more inte-

grated language mix overall, while GPT-4o-mini remains more conservative.

The differences between the two models are less pronounced in the CESAR score. GPT-4o-mini achieves a mean of 0.273, while LLaMA-3.3 achieves slightly higher values of 0.283. However, none of both match the CESAR score of the Input-Data (0.336), indicating that neither the complexity nor the burstiness of the CS patterns in the corpus are fully replicated. Figure 1 shows that LLaMA-3.3 is closer to the corpus mean values across all metrics than GPT-4o-mini. Overall, these results only partially support Hypothesis 1. Both models imitate global properties of the CS structure, but underestimate the degree of linguistic diversity, alternation frequency and complexity, with LLaMA-3.3 approximating the Input-Data more consistently than GPT-4o-mini.

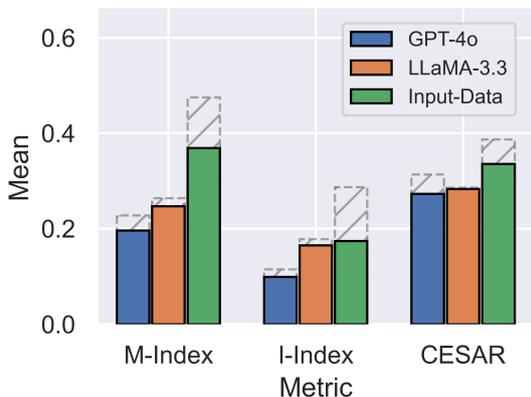


Figure 1: Mean values and standard deviation of M-Index, I-Index, and CESAR for GPT-4o-mini, LLaMA-3.3, and the Input-Data.

4.2 SLI-triggered CS

Building on the concept of Shared Lexical Items (SLIs) and the triggering hypothesis from subsection 2.3, we now examine whether SLIs occur preferentially near CS points in our data (Hypothesis 2). Specifically, we examine the extent to which LLMs reproduce the SLI trigger effects observed in the Input-Data.

For each of the three sources (GPT-4o-mini, LLaMA-3.3, Input-Data), we construct a 2×2 contingency table that compares SLIs vs. non-SLIs and ‘near to CS’ vs. ‘not near to CS’. A token is considered to be near to CS if a switch point occurs within a range of six words. This range captures both upstream and downstream trigger effects. We

use the corresponding frequencies to calculate the Relative Switching Propensity (RSP), i.e. the ratio of the CS probability near SLIs to the CS probability near non-SLIs. Values greater than 1 indicate a positive correlation between SLIs and CS. The model values are averaged over 50 generation runs (2,500 posts per model), with the Input-Data corresponding to the 50 original posts. The central results are summarised in Table 1.

	GPT-4o-mini	LLaMA-3.3	Input-Data
% SLIs near CS	63.00	87.50	67.06
% non-SLIs near CS	32.81	66.17	37.84
RSP	1.92	1.32	1.77

Table 1: Shared lexical items (SLIs) and their association with CS: percentage of SLIs and non-SLIs near CS and relative switching propensity (RSP) for GPT-4o-mini, LLaMA-3.3, and the Input-Data.

The Input-Data shows a clear SLI trigger effect. Of the 85 SLIs in the sample, 57 occur near CS (67.06 %), while this is only true for 1,077 of 2,845 non-SLIs (37.84 %). This results in an RSP of 1.77. This means that CS is 1.77 times more likely to be near an SLI than near a non-SLI. For the Input-Data, Fisher’s exact test shows that SLIs occur near CS significantly more often than non-SLIs ($p < 0.05$).

GPT-4o-mini clearly reflects this pattern. On average, 63.00 % of SLIs are close to CS, while this applies to only 32.81% of non-SLIs. The resulting RSP is 1.92, which is slightly above the value of the Input-Data (1.77). For GPT-4o-mini, SLIs occur near CS significantly more often than non-SLIs according to Fisher’s exact test ($p < 0.05$).

LLaMA-3.3 shows a slightly different profile. Although 87.50 % of SLIs are close to CS, the proportion of non-SLIs close to CS is also significantly higher at 66.17 %. Accordingly, the RSP is lower at 1.32, as the baseline CS probability is also high for non-SLIs. For LLaMA-3.3, Fisher’s exact test does not indicate a significant association ($p \geq 0.05$).

Overall, Input-Data and GPT-4o-mini exhibit a clear SLI trigger effect with comparable RSP values, whereas LLaMA-3.3 shows a weaker association, reflecting a generally higher CS rate across tokens. From a quantitative perspective, the results thus provide partial confirmation of Hypothesis 2. SLIs are associated with an increased CS probability, and GPT-4o-mini reproduces this pattern closer to the Input-Data than LLaMA-3.3.

4.3 Human Evaluation

The results of the automated analysis were supplemented by a human evaluation, in which the Perceived Naturalness and Fluency of the posts were assessed.

To assess **Perceived Naturalness**, participants rated all posts on a five-point Likert scale from ‘very unnatural’ (1) to ‘very natural’ (5). Overall, GPT-4o-mini was rated significantly more positively than LLaMA-3.3 and the Input-Data. The categories ‘rather natural’ or ‘very natural’, were selected in around 47 % of the ratings for GPT-4o-mini fall into the categories ‘rather natural’ or ‘very natural’, compared to 33 % for LLaMA-3.3 and 30 % for the Input-Data. Conversely, the Input-Data was most frequently rated as (rather) unnatural (56 %), followed by LLaMA-3.3 (50%) and GPT-4o-mini (36 %). Interestingly, in terms of mean scores, GPT-4o-mini achieves the highest Perceived Naturalness (3.19), followed by LLaMA-3.3 (2.77) and the Input-Data (2.56). The standard deviations are very similar across all three sourced at around 1.2 points on the five-point scale. Mann-Whitney U tests with Bonferroni correction indicate that the distributions of naturalness ratings differ significantly between all three sources (corrected $p < 0.01$).

The **Perceived Fluency** of the language switches was also measured on a five-point scale from ‘very unsmooth’ (1) to ‘very smooth’ (5). Here, too, the LLMs performed better overall than the Input-Data. For GPT-4o-mini, about half of the ratings fall into the top two categories (‘rather fluent’ and ‘very fluent’), while for LLaMA-3.3 the proportion is slightly lower, and the Input-Data is perceived as (rather) non-fluent significantly more often. The mean fluency scores again show a clear ordering. GPT-4o-mini obtains the highest mean rating (3.29), LLaMA-3.3 scores 3.02 and the Input-Data 2.60, with standard deviations close to 1.2 for all three sources. Mann-Whitney U tests indicate significant differences between both LLMs and the Input-Data (corrected $p < 0.001$), while the difference between GPT-4o-mini and LLaMA-3.3 in fluency ratings is not statistically significant. [Figure 2](#) summarises the mean ratings for both Perceived Naturalness and Fluency and illustrates the consistent advantage of GPT-4o-mini over LLaMA-3.3 and the Input-Data.

The open-ended responses regarding the reasons for (un)naturalness provide additional insights into

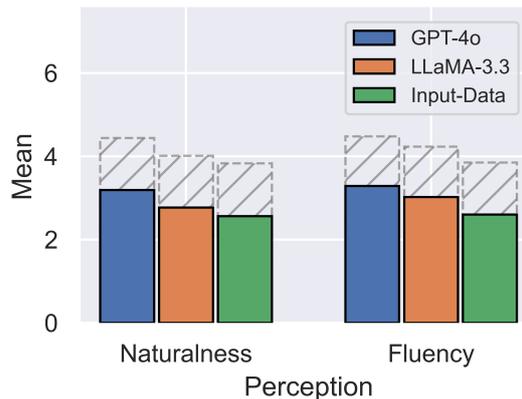


Figure 2: Mean values and standard deviation of Perceived Naturalness and Perceived Fluency of GPT-4o-mini, LLaMA-3.3, and the Input-Data (5-point Likert scale).

the perception of the models. The most frequently cited reason for unnaturalness is an ‘Unnatural mixture of German and English’. This reason occurs most frequently with GPT-4o-mini, followed by LLaMA-3.3 and the Input-Data. ‘Unusual choice of words’ is mentioned with similar frequency in all three groups. It is striking that a ‘Lack of natural speech flow’ is mentioned particularly frequently in LLaMA-3.3 and the Input-Data, while this reason appears less frequently in GPT-4o-mini. Grammatical errors, on the other hand, are mainly attributed to the Input-Data and significantly less frequently to the LLM outputs, which suggests that the models usually produce relatively clean grammar.

To qualitatively illustrate the differences between GPT-4o-mini, LLaMA-3.3, and the Input-Data, [Table 2](#) shows representative three-sentence posts from each source used in our experiments. For each source, we show one post with low, medium and high CS rates to illustrate variation across CS intensity levels.

Overall, and considering only the numerical results, the human evaluation indicates that the two LLMs employed, with —GPT-4o-mini in particular, outperform the authentic Reddit posts with respect to perceived naturalness and fluency of language. We attribute this primarily to the models’ capacity to produce code-switching, combined with their tendency to generate text that adheres closely to standard grammatical and orthographic conventions, in contrast to typical Reddit contributions. Had the LLMs been able to systematically imitate the non-standard, error-prone style characteristic

GPT-4o-mini	LLaMA-3.3	Input-Data
Post 1 (low CS)		
Die letzten Tage waren einfach <i>crazy</i> . Ich habe so viel Zeit damit verbracht, mich durch die neuen Marvel-Serien zu <i>binge-wachen</i> und es ist als ob ich in einer anderen Welt bin. Die Storylines sind so gut, dass ich ständig über die Charakterentwicklung nachdenke und dabei dachte ich: <i>Are they really going to pull this off?</i> Hoffentlich geht das nicht in die Hose wie bei manchen Filmen.	Das Problem mit den neuen Datenschutzbestimmungen ist, dass sie oft sehr komplex sind und man sich durch die kleinen Druckerzeichen und den <i>fine print</i> kämpfen muss, um zu verstehen, was man da eigentlich akzeptiert, und ich denke, das ist ein bisschen wie ein <i>Catch-22</i> , weil man entweder seine Privatsphäre opfern oder sich durch einen <i>terms of service</i> kämpfen muss, der länger ist als ein Harry-Potter-Buch.	<i>There is no glory in prevention</i> ja, aber ob die absolute Anzahl der Kontakte durch eine Ausgangssperre wirklich reduziert wurde, ist schon zu hinterfragen.
Post 2 (medium CS)		
Das Wetter hier ist wirklich verrückt. Letzte Woche waren wir bei 30 Grad und jetzt liegt Schnee. <i>Just when I thought</i> , wir könnten endlich den Frühling genießen, kommt der Winter zurück. Ich meine, <i>seriously</i> , was soll das? Es ist, als würde <i>Mother Nature</i> mit uns spielen.	Ich bin gerade total <i>frustrated</i> mit dem aktuellen Stand der Dinge, es ist einfach <i>too much</i> , wenn man sieht, wie die Politiker mit dem Thema Umsatzsteuer umgehen, <i>total unfair</i> , und ich denke, es ist Zeit für einen <i>change</i> , vielleicht sollte man einfach das System hinterfragen und nach neuen Lösungen suchen, wie zum Beispiel eine <i>flat tax</i> , aber das ist natürlich ein komplexes Thema, das mehr Diskussion und Expertenwissen braucht, um es richtig zu verstehen und umzusetzen.	Amerikanische <i>mandatory sentences</i> und auch <i>three strike laws</i> , grade in Kombination mit der Todesstrafe, sind extremst menschenunwürdig und von der Art und Weise wie Menschen zu unfairen <i>plea-deals</i> genötigt werden weil das Justizsystem einfach komplett zusammen brechen würde wenn auch nur 40% aller Menschen auf ne <i>trial</i> bestehen würden will ich gar nicht erst anfangen.
Post 3 (high CS)		
Ich habe neulich eine Doku über die amerikanische Kultur gesehen und <i>wow</i> , die Unterschiede sind echt krass. In den USA feiern sie <i>Thanksgiving</i> mit <i>Turkey</i> und <i>Family</i> , während wir hier in Deutschland einfach unser Weihnachtessen haben. Das klingt fast wie ein Wettbewerb, welcher Feiertag die meisten Kalorien bringt. Und dann die <i>Black Friday Sales</i> – <i>what’s up with that?</i> Ich kann mir nicht vorstellen, dass wir hier so etwas machen würden.	Ich habe gerade ein <i>super cooles</i> Video auf <i>YouTube</i> gesehen, das war ein totaler <i>game changer</i> , weil es zeigte, wie man mit einem einfachen Hack das Leben leichter machen kann, zum Beispiel mit einem <i>life hack</i> , um den Alltag zu vereinfachen, und ich dachte, das wäre ein interessanter <i>topic</i> für eine Diskussion, weil es zeigt, dass man mit einem <i>little trick</i> und einem bisschen <i>creativity</i> ein <i>lot</i> erreichen kann.	das problem dabei ist, wenn man <i>browser</i> erweiterungen wie <i>Reddit Enhancement Suite</i> benutzt steht hinter jedem namen den man jemals <i>ge-up</i> oder <i>downvoted</i> hat die summe der <i>votes</i> . Deine art zu <i>voten</i> ist dann wie wenn du auf <i>youtube</i> einmal nen <i>rap video upvotest</i> weil der text besonders kreativ ist oder aus welchem grund auch immer obwohl du <i>rap</i> musik hasst und dann hast du die nächsten 2 monate auf deiner <i>youtube</i> startseite nur noch <i>rap videos / kanäle</i> .

Table 2: Representative three-sentence posts from GPT-4o-mini, LLaMA-3.3, and the Input-Data illustrating typical German–English CS patterns observed in our experiments. Posts are ordered from low to high CS rate.

of many Reddit posts, it is likely that the evaluators would have judged the LLM-generated content less favorably than the original dataset messages.

A further perspective on these findings arises from the composition of our evaluator sample. The participants are not representative of the general population of German Reddit users who frequently engage with the platform. This limitation and its implications are discussed in more detail in the following section.

Crucially, the primary aim of the human evaluation was not to assess the LLMs’ ability to perfectly replicate prototypical Reddit texts—or, more broadly, highly irregular user-generated content—but rather to determine whether the models can

produce code-switching in a manner that human readers find convincing.

5 Discussion

This study examined the extent to which current LLMs can realistically imitate German-English CS in Reddit-like texts. The focus was on global CS patterns in comparison to the Denglisch Corpus (RQ 1) and SLI-triggered language switching (RQ 2). This was supplemented by a human evaluation of Perceived Naturalness and Fluency. In the following, we discuss the findings in terms of these questions and possible technical causes.

With regard to RQ 1, the models only partially

reproduce the global CS characteristics. In terms of the M-Index and I-Index, LLaMA-3.3 shows smaller deviations from the Denglisch corpus than GPT-4o-mini, indicating a higher degree of linguistic mixing and more frequent language switching. For CESAR, both models exhibit similarly small differences relative to the Input-Data. From a purely metric point of view, LLaMA-3.3 initially appears to be more ‘English-like’ than GPT-4o-mini. However, the human evaluation shows that greater mixing and higher alternation frequency are not automatically perceived as more natural or fluent. GPT-4o-mini achieves significantly higher ratings for Perceived Naturalness and Fluency than LLaMA-3.3 and the Input-Data, even though it switches more conservatively overall. This supports the assumption that natural Denglisch is characterised more by selective, well-embedded language switching than by ‘as much CS as possible’. This is a pattern that GPT-4o-mini realises more strongly than LLaMA-3.3.

RQ 2 concerned SLI-triggered language switching. Here, the Input-Data and GPT-4o-mini show very similar patterns. In both cases, the probability of a switch in the vicinity of SLIs is significantly increased. The RSP in these cases is in a comparable range. Although LLaMA-3.3 also has an RSP value above 1, the effect is weaker and statistically not significant. Considering the triggering hypothesis, this can be interpreted as GPT-4o-mini apparently having learned to use certain shared lexemes as natural ‘anchor points’ for language switching, while LLaMA-3.3 uses CS in a much more widespread manner. Due to the overall high CS rate, the switch probability also increases in the vicinity of non-SLIs, which weakens the relative trigger effect. RQ 2 is thus also partially confirmed. SLIs clearly act as triggers in both the Input-Data and the GPT-4o-mini outputs, but this mechanism is significantly weaker in LLaMA-3.3.

The observed differences can be plausibly explained by technical and linguistic factors. GPT-4o-mini belongs to the GPT-4 family and was trained on a broad, multilingual database, followed by an *Reinforcement Learning from Human (RLHF)* phase with a strong focus on dialogical and user-oriented scenarios (OpenAI, 2024). OpenAI states that GPT-4 shows significantly improved performance in non-English languages compared to previous versions (OpenAI, 2024). Given this background, the strong performance of GPT-4o-mini in a German-English context can be understood as

a result of diversified training. In addition, RLHF tuning may have contributed to GPT-4o-mini tending to respond in the language of the input and only changing languages when stylistically or contextually motivated, as indirectly evidenced by the studies of (Zhou et al., 2024). This assumption is supported by the results of this work regarding the selective use of CS in GPT-4o-mini outputs.

LLaMA-3.3-70B-Instruct is based on Meta’s LLaMA-3 architecture, which was pre-trained on approximately 15 trillion publicly available tokens (Grattafiori et al., 2024). According to the model card, the corpus contains texts in over 30 languages but is heavily dominated by English. Only about 5 % of the training data is non-English content (Hugging Face, 2024), so the proportion of explicitly German texts is probably well below 1 %. Instruction fine-tuning was also predominantly carried out on English-language prompts (Hugging Face, 2024). This configuration is consistent with the findings of Meeus et al. (2024) and Chang et al. (2023), which show that models trained primarily in English and highly multilingual systematically lose quality in underrepresented languages (‘curse of multilinguality’). This offers a plausible explanation for the weaker performance of LLaMA-3.3 in the Denglisch context.

In substance, this has several consequences for the evaluation of CS in LLMs. First, it becomes clear that surface metrics such as M-Index, I-Index and CESAR provide important information, but without human evaluation and psycholinguistically motivated measures such as RSP, they can easily be misleading. One model may be closer to the Input-Data in terms of metrics, yet appear less natural than another. Second, the SLI findings show that LLMs can internalise subtle, cognitively motivated patterns of bilingual language production under suitable training conditions. GPT-4o-mini closely replicates the SLI trigger effect of the Input-Data, while LLaMA-3.3 only replicates it to a lesser extent. Thirdly, the diverging profiles of GPT-4o-mini and LLaMA-3.3 underscore that CS competence is highly model-dependent and that statements about the CS ability of LLMs must always be interpreted against the background of specific training data and fine-tuning regimes.

6 Conclusion

This study examined the extent to which current LLMs can replicate German-English CS in Red-

dit posts. Compared to the Denglisch Corpus, both models show that they only approximate global CS patterns. LLaMA-3.3 is closer to the Input-Data values from the corpus in terms of M-Index and I-Index, while GPT-4o-mini is perceived as significantly more natural and fluent in the human evaluation.

With regard to SLI-triggered CS, GPT-4o-mini shows a pattern that is very similar to the Input-Data, while the relative trigger effect is significantly weaker in LLaMA-3.3. Overall, this suggests that LLMs are indeed capable of internalising more refined, psycholinguistically motivated CS structures, but that their expression is highly model-dependent.

7 Future Work

This study provides relevant insights into the ways in which LLMs emulate German–English code-switching. At the same time, the identified limitations and the obtained results suggest several avenues for further research that can be pursued in subsequent studies. In addition to the question of the extent to which LLMs are capable of generating irregular user-generated content, the following code-switching-related topics may be addressed:

- By fine-tuning selected LLMs on annotated CS corpora, it is possible to investigate whether the models generate more realistic language switches as a result. This would demonstrate the extent to which CS can be learned and controlled by LLMs.
- A comparison of different language combinations, such as Turkish-German, can provide information about whether the observed model behaviour is stable across languages or specific to the German-English language pair. This would allow for a better differentiation between language pair- and model-specific influences.
- The use of alternative corpora, such as spoken dialogue data or other platforms beyond Reddit, could provide deeper insights into how the characteristics of the Input-Data influence CS behaviour in LLMs.
- A focus on prompting could help to understand the extent to which this affects CS behaviour or what remains shaped by internal training patterns.

- The development of a standardised benchmark for evaluating CS in LLMs would create a basis for reproducible research and enable systematic model comparisons.
- Finally, it could also be investigated to what extent the models use CS in a discourse-specific manner. This means, for example, whether they use CS for emphasis, group identification, or to mark contrast.

Limitations

Like any study, this work has several limitations that should be taken into account when interpreting the results. These mainly concern the data basis, the selection of models, the analysis tools used, and the design of the human evaluation. We consider the following the most important.

One key limitation concerns the data used. The Denglisch Corpus is based on Reddit posts, which reflect authentic and informal language use but represent a very specific type of Code-Switching. This is characterised by written online communication, platform culture and specific socio-demographic user groups. Verbal Code-Switching in everyday life, for example in spontaneous conversations, can be significantly different and was not taken into account in this work.

In addition, the data set could partially be considered outdated. Some of the subreddits analysed originate from earlier years (e.g. de_2013), which may have a negative impact on the representativeness of the data. The limited scope of the Input-Data used also constitutes a limitation since only a small, randomly selected portion (50 posts) of the total Denglisch Corpus was used. As a result, extreme cases and marginal phenomena of Code-Switching may have been under-represented.

Another limitation concerns the annotation of the data. Automatic categorisation using the CRF-based classifier carries the risk of misclassification, especially in the case of complex or ambiguous words. Since many evaluations in this work are based on these annotations, even minor errors in labeling can have an impact on the calculated metrics. For practical reasons, it was not possible to perform a complete manual check in the context of this work.

The transferability of the results is also limited in terms of content. The study only examined the German–English language pair. However, Code-Switching is a language-specific and culturally in-

fluenced phenomenon that can vary greatly in structure, frequency, and motivation in other language constellations. The findings obtained here cannot, therefore, be easily transferred to other language pairs or communication contexts.

Finally, the selection of LLMs is also limited. Only two models were used in the evaluation. They differ in terms of architecture, training data, and licensing, but do not represent the full range of the existing LLM spectrum. The results, therefore, cannot be generalized to other model classes.

Ethical Considerations

The study raises no ethical concerns regarding data collection or analysis, as it relies exclusively on publicly available Reddit data and anonymized, automatically generated text, without involving personal or sensitive information. A potential ethical risk lies in the possible misuse of the findings to generate synthetic posts for online forums such as Reddit, which are increasingly difficult to distinguish from human-authored content. Such applications could facilitate the spread of misinformation or manipulative content. At the same time, more realistic and context-sensitive language generation can also have positive societal effects, for example, by enabling more inclusive and natural interactions in advisory, support, or recommendation scenarios. We therefore emphasize that the results should be interpreted and applied with care, and that responsible use of code-switching-capable language models is essential.

All participants in the human evaluation of the generated and Reddit data used a crowdsourcing platform, participated voluntarily, and were reimbursed at least the German minimum wage.

References

Karima Abidi and Kamel Smaïli. 2021. [CESAR: A new metric to measure the level of code-switching in corpora -Application to Maghrebian dialects](#). In *Springer series "Advances in Intelligent Systems and Computing"*, Springer series "Advances in Intelligent Systems and Computing", Amsterdam, Netherlands.

Ruthanna Barnett, Eva Codó, Eva Eppler, Montse Forcadell, Penelope Gardner-Chloros, Roeland van Hout, Melissa Moyer, Maria Carme Torras, Maria Teresa Turell, Mark Sebba, Marianne Starren, and Sietse Wensing. 1999. [The lides coding manual: A document for preparing and analyzing language interaction data](#). Technical Report Version 1.1, Tilburg University, Tilburg, The Netherlands.

Asma Ben Abacha, Wen-wai Yim, Yajuan Fu, Zhaoyi Sun, Meliha Yetisgen, Fei Xia, and Thomas Lin. 2025. [MEDEC: A benchmark for medical error detection and correction in clinical notes](#). In *Proc. ACL 2025*, pages 22539–22550, Vienna, Austria. ACL.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. [On the dangers of stochastic parrots: Can language models be too big?](#) In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21*, page 610–623, New York, NY, USA. Association for Computing Machinery.

Mirjam Broersma and Kees de BOT. 2006. [Triggered codeswitching: A corpus-based evaluation of the original triggering hypothesis and a new alternative](#). *Bilingualism: Language and Cognition*, 9(1):1–13.

MIRJAM BROERSMA and KEES DE BOT. 2006. [Triggered codeswitching: A corpus-based evaluation of the original triggering hypothesis and a new alternative](#). *Bilingualism: Language and Cognition*, 9(1):1–13.

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, and 12 others. 2020. [Language models are few-shot learners](#). *Preprint*, arXiv:2005.14165.

Özlem Çetinoğlu, Sarah Schulz, and Ngoc Thang Vu. 2016. [Challenges of computational processing of code-switching](#). In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 1–11, Austin, Texas. Association for Computational Linguistics.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. 2023. [When is multilinguality a curse? language modeling for 250 high- and low-resource languages](#). *Preprint*, arXiv:2311.09205.

Jie Chi, Electra Wallington, and Peter Bell. 2024. [Characterizing code-switching: Applying linguistic principles for metric assessment and development](#). In *Proceedings of Interspeech 2024*, Proceedings of Interspeech, pages 7–11. ISCA. The 25th Interspeech Conference, Interspeech 2024 ; Conference date: 01-09-2024 Through 05-09-2024.

Michael Clyne. 2003. [Dynamics of language contact: English and immigrant languages](#). *Dynamics of Language Contact: English and Immigrant Languages*, pages 1–282.

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, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.

- François Grosjean. 2021. *Life as a Bilingual: Knowing and Using Two or More Languages*. Cambridge University Press.
- Gualberto Guzmán, Joseph Ricard, Jacqueline Serigos, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2017. [Metrics for modeling code-switching across corpora](#). In *Proceedings of Interspeech 2017*, pages 67–71, Stockholm, Sweden. ISCA.
- Krystal Hu. 2023. [Chatgpt sets record for fastest-growing user base – analyst note](#). *Reuters*.
- Hugging Face. 2024. Llama-3.3-70B-Instruct. <https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct>. Accessed on June 11, 2025.
- Gerrit Jan Kootstra. 2015. [A psycholinguistic perspective on code-switching: Lexical, structural, and socio-interactive processes](#). In Gerald Stell and Kofi Yakpo, editors, *Code-switching Between Structural and Sociolinguistic Perspectives*, pages 39–64. DE GRUYTER.
- Matthieu Meeus, Anthony Rathé, François Remy, Pieter Delobelle, Jens-Joris Decorte, and Thomas De-meester. 2024. [Chocollama: Lessons learned from teaching llamas dutch](#). *Preprint*, arXiv:2412.07633.
- Pieter Muysken. 2011. [Code-switching](#). In Rajend Mesthrie, editor, *The Cambridge Handbook of Sociolinguistics*, 1 edition, pages 301–314. Cambridge University Press.
- Carol Myers-Scotton. 1989. [Codeswitching with english: types of switching, types of communities](#). *World Englishes*, 8(3):333–346.
- Carol Myers-Scotton. 2017. *Code-Switching*, chapter 13. John Wiley & Sons, Ltd.
- OpenAI. 2024. GPT-4o mini: Advancing cost-efficient intelligence. <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>. Accessed on June 11, 2025.
- OpenAI. 2024. Hallo gpt-4o. <https://openai.com/de-DE/index/hello-gpt-4o/>. Accessed on July 8, 2025.
- Doreen Osmelak and Shuly Wintner. 2023. [The denglich corpus of german-english code-switching](#). In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2023)*, pages 42–51, Dubrovnik, Croatia. Association for Computational Linguistics.
- Shana Poplack. 1980. [Sometimes i’ll start a sentence in spanish y termino en español: toward a typology of code-switching 1](#). *Linguistics*, 18:581–618.
- Claudia Maria Riehl. 2019a. [Code-switching](#). Online publication, LMU Munich.
- Claudia Maria Riehl. 2019b. Sprachkontaktforschung, soziolinguistik und code-switching. In Hans Goebel, Peter H. Nelde, Zdeněk Starý, and Wolfgang Wölck, editors, *Sprachkontaktforschung: Ein internationales Handbuch zur Sprache in Kontaktzonen*, volume 30.3 of *Handbücher zur Sprach- und Kommunikationswissenschaft*, pages 1408–1418. De Gruyter Mouton, Berlin / Boston.
- Shuly Wintner, Safaa Shehadi, Yuli Zeira, Doreen Osmelak, and Yuval Nov. 2023. [Shared lexical items as triggers of code switching](#). *Transactions of the Association for Computational Linguistics*, 11:1471–1484.
- Ruochen Zhang, Samuel Cahyawijaya, Jan Christian Blaise Cruz, Genta Winata, and Alham Fikri Aji. 2023. [Multilingual large language models are not \(yet\) code-switchers](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12567–12582, Singapore. Association for Computational Linguistics.
- Jiayi Zhou, Jiaming Ji, Juntao Dai, and Yaodong Yang. 2024. [Sequence to sequence reward modeling: Improving rlhf by language feedback](#). *Preprint*, arXiv:2409.00162.

A Prompts

“In linguistics, code-switching refers to a process whereby a speaker switches from one language to another within a single utterance or within a text or dialogue. The more dominant language is called the ‘matrix language’, while the inserted language is called the ‘embedded language’. Here are some examples of code-switching between German and English from Reddit posts:”

Abbildung A.1: First part of the prompt given to the LLMs to introduce Code-Switching; English version

...
input_37 (CS-Rate: 0.037): There is no glory in prevention ja, aber ob die absolute Anzahl der Kontakte durch eine Ausgangssperre wirklich reduziert wurde, ist schon zu hinterfragen.
input_38 (CS-Rate: 0.0345): Also ich bezahle gerne steuern wenn dadurch menschen wie du entlastet werden.

From each according to his ability, to each according to their needs.
...

Abbildung A.2: Second part of the prompt given to the LLMs providing the Input-Data (exemplary extract)

“You are a Reddit user writing a 1-3 sentence post. The dominant language is German (matrix language), embedded expressions, words or phrases are in English (embedded language). Write a single Reddit post with 1-3 sentences that reads authentic, human and not generic. The post must contain natural-sounding code-switching. The code-switching rate is defined as the number of code-switching points relative to the total number of tokens in the post. It is extremely important that the generated posts contain an average code-switching rate of approximately {median_cs_rate: .3f} (several short English insertions per post are desirable). Each time, choose a new topic that realistically appears in Reddit posts, e.g. emotional, factual, political, funny, serious, trivial or profound. Consciously vary the style with each run: sometimes introspective, sometimes sarcastic, sometimes angry, sometimes ironic, sometimes sober. Don’t use recurring introductions such as ‘I think’ or ‘I have’, but start freely, spontaneously and in a variety of styles. The text should read as if it were written by a real Reddit user with concrete examples, details, facts or personal impressions. Make sure that no post is too similar to a previous example or deals with the same topic. Also, be careful not to use introductions. Just give me the post. Do not use special characters or symbols to mark English words. Individual sentences should not be too long or convoluted.”

Abbildung A.3: Third part of the prompt given to the LLMs providing the instruction; English version

B Survey Questions

Survey Block 1: Main Section

Question 1 (Perceived Naturalness): How natural does the post seem to you?
a) Very unnatural b) Rather unnatural c) Neutral d) Rather natural e) Very natural

Question 2: What makes the post feel unnatural?
a) Unusual choice of words b) Unnatural mixture of German and English
c) Grammar errors d) Lack of natural speech flow e) Other

Question 3: What makes the post feel natural?
a) Authentic language switching b) Natural mixture of German and English
c) Typical Reddit phrases d) Natural word choice e) Other

Question 4 (Perceived Fluency): How smoothly does the text switch between English and German?
a) Very unsmooth b) Rather unsmooth c) Neutral d) Rather smooth e) Very smooth

Abbildung B.1: Survey Block 1, evaluation of five Reddit posts (from LLMs and Input-Data) in terms of linguistic features

Survey Block 2: Control Question

Question 5: This is a control question to ensure that you are reading carefully. Please select answer option 3 here.
a) Strongly disagree b) Rather Disagree c) Neutral (correct answer) d) Rather agree
e) Strongly agree

Abbildung B.2: Survey Block 2, control question to ensure the attention of the participants

Survey Block 3: Demography

Question 6: Which age group do you fall into?
a) Under 18 b) 18-24 c) 25-34 d) 35-44 e) 45-54 f) 55-65 g) 65 or older

Question 7: What is your native language?
a) German b) English c) Both

Question 8: How would you rate your level in the other language?
a) Beginner b) Intermediate c) Advanced d) Fluent

Abbildung B.3: Survey Block 3, demographic questions to record the age and language skills of the participants

Survey Block 4: Online Forums and CS

Question 9: How often do you read in online forums such as Reddit?

- a) Less than once/month b) Once/month c) Once/week d) Every few days e) Every day
-

Question 10: How often do you post in online forums such as Reddit?

- a) Less than once/month b) Once/month c) Once/week d) Every few days e) Every day
-

Question 11: How often do you use Code-Switching in online texts?

- a) 0% - 25% of cases b) 26% - 50% of cases c) 51% - 75% of cases d) 76% - 100% of cases

Abbildung B.4: Survey Block 4, questions about usage of online forums and CS

Perplexity as a Metric for Dialectal Distance: A Computational Study of Greek Varieties

Stergios Chatzikyriakidis^{1*}, Erofilis Psaltaki^{2*},
Dimitrios Papadakis¹, Erik Henriksson², Veronika Laippala²

¹University of Crete, ²TurkuNLP, University of Turku
stergios.chatzikyriakidis@uoc.gr, erofili.psaltaki@utu.fi,
philp0961@philology.uoc.gr, mavela@utu.fi, erik.henriksson@utu.fi

Abstract

In this paper, we use LLM perplexity as a measure to assess Greek dialectal distance¹. We test seven models on Standard Modern Greek (SMG) and eight dialects, namely Heptanesian, Cypriot, Maniot, Pontic, Northern, Cretan, Tsakonian, and Griko. Using samples of 5k, 15k, and 25k tokens from the GRDD+ corpus for each variety, we find a consistent dialect ranking across models, with Heptanesian closest to SMG, and Griko most distant (perplexity ratio 3.6–14.5× depending on model). These results are largely in agreement with theoretical dialectological knowledge. For example, Tsakonian consistently appears distant in all measures, reflecting its status as the sole Doric descendant, while Heptanesian appears closer by all metrics, pointing to its status as one of the dialects used to shape the official variety. Perplexity correlates strongly with Bits-Per-Character (mean $r = 0.94$) and Normalized Compression Distance (mean $r = 0.87$, range 0.76–0.93), providing support for its use as a dialectometric tool. However, a number of important confounds are also found. First, tokenization effects compress Llama 2’s perplexity range. Second, genre artifacts seem to inflate the results for Cretan. Third, potential training data contamination likely reduces perplexity for Cypriot and Pontic. Lastly, we find that Greek-specific models like Meltemi and Krikri do not consistently outperform general models.

1 Introduction

Modern Greek dialects provide an interesting case study for distance studies due to their rich dialectal diversity, ranging from varieties closely related to Standard Modern Greek (SMG), such as Heptanesian, to highly divergent varieties like Tsakonian and Griko, the latter spoken in Southern Italy and

heavily influenced by Romance languages. Trudgill (2003) claims that Tsakonian, Southern Italian Greek, Pontic, and Cappadocian are the varieties that are linguistically very different from Standard Greek. While most research on Greek dialects has focused on their geographical distribution and classification, there remains a gap in computationally assessing their linguistic distances and classifying them based on these distances. Even in traditional dialectology, there is no classification of Greek dialects based on their linguistic distance.

Recently, the computational assessment of linguistic distance between language varieties has gained renewed interest with the advent of large language models (LLMs). In this context, automatic metrics such as LLM perplexity can be used to approximate linguistic distances—both inter-dialect distance and distance from the standard variety—regardless of how much the model has been exposed to each variety during training. Employing these metrics contributes both to our understanding of dialect relationships and to the development of low-cost methods in computational dialectology.

This study examines Greek dialectal distances using LLM perplexity as a primary metric, complemented by information-theoretic measures (bits per character, BPC) and compression-based distances (normalized compression distance, NCD). Together, these metrics offer complementary insights into dialectal distance: LLM perplexity captures divergence from patterns learned by the model, bits per character (BPC) provides a measure of predictability and normalized compression distance (NCD) quantifies similarity based on shared redundancy.

2 Related Work

A foundational work on Modern Greek dialects is that by Newton (1972), who established that high vowel loss and vowel raising define the northern

*These authors contributed equally.

¹The full code for the dialect metrics and the dataset are available at the following anonymous link: https://github.com/TurkuNLP/Greek_dialects_perplexity.

dialects. Another foundational work is [Trudgill \(2003\)](#). His cartographical representation identifies fifteen dialect areas based on six phonological criteria. [Kontossopoulos \(2001\)](#) distinguishes *di-alektoi* (Tsakonian, Pontic, Cappadocian, Southern Italian Greek) from *idiomata* (all other varieties), a taxonomy predicting that the former should show substantially greater distance from SMG. Among divergent varieties, Tsakonian is the sole Modern Greek descendant of Doric rather than Koine Greek ([Nicholas, 2019](#)), while Griko preserves the infinitive, lost elsewhere by the Medieval period ([Holton and Manolesou, 2010](#)) alongside extensive Romance contact effects ([Manolesou, 2005](#)). Pontic retains distinct features including infinitive constructions and distinctive across-the-board enclitic placement ([Sitaridou, 2014](#); [Condoravdi and Kiparsky, 2002](#)). Cypriot exhibits phonological differences in terms of gemination ([Arvaniti, 2001](#)) plus a number of morphosyntactic differences including clefts and differences in clitic positioning.

Previous research in NLP has explored the use of smaller models for perplexity-based pruning ([Ankner et al., 2024](#)), along with analyses of dialectal biases in LLMs ([Pan et al., 2025](#)). Other studies have assessed the fairness and robustness of LLMs in handling dialects across canonical reasoning tasks ([Lin et al., 2025](#)) and have investigated tokenization and representation biases in multilingual models for dialectal NLP tasks ([Kanjirang et al., 2025](#)). Additional work includes the creation of a library for exploring where a particular language model is perplexed ([Cooper and Scholak, 2024](#)), in addition to analyses of low-perplexity sequences—high-probability text spans generated by language models ([Wuhrmann et al., 2025](#)). Furthermore, perplexity has also been used as a stylistic signal for authorship attribution through author-specific language models ([Huang et al., 2025](#)).

There is also research on dialect metrics, such as the application of compression-based similarity measures to the quantification of distances among Bulgarian dialects ([Simov and Osenova, 2007](#)), and model training with optimal tokenization levels in datasets with consistent and inconsistent writing practices ([Kanjirang et al., 2023](#)). Other studies have attempted to predict dialectal features at the token level for Norwegian dialects ([Barnes et al., 2023](#)).

Previous research on the classification of Modern Greek dialects has largely focused on their geographical distribution, with the most recent clas-

sification proposed by [Trudgill](#). There have also been classification attempts for Athenian Greek and Cypriot Greek based on vowel acoustic parameters ([Themistocleous, 2017](#)), as well as on measures of temporal and spectral information from selected consonants ([Themistocleous, 2019](#)). More recent work has quantified dialectal distances among Asia Minor Greek dialects using dialectometric techniques ([Bompolas and Melissaropoulou, 2025](#)). To the best of our knowledge, this study is the first attempt to computationally measure Greek dialectal distances using multiple metrics.

3 Data

We utilize the GRDD+ dataset ([Chatzikyriakidis et al., 2025](#)), an extended Greek dialectal dataset that includes data from ten varieties of Greek. Four of these dialects—Cretan, Cypriot, Pontic, Northern—and Standard Modern Greek were part of the original GRDD dataset ([Chatzikyriakidis et al., 2023](#)) and have since been expanded in terms of coverage by the dataset’s authors. Six additional varieties have been incorporated into the GRDD+: Greco-Corsican, Griko, Heptanesian, Tsakonian, Maniot, and Katharevusa. For the purposes of the present study, we use all available varieties of the GRDD+ except Greco-Corsican, as it is extinct, and Katharevusa, since it represents a historical formal register.

Standard Modern Greek (SMG) is the official language of both Greece and Cyprus ([Mackridge, 1985](#)). It is based primarily on Peloponnesian dialects and was later enriched by features from the dialects of Istanbul and the Ionian Islands.

Cypriot Greek is the native language of most Greek Cypriots, both on the island of Cyprus and in the diaspora, although the official language of the Republic of Cyprus is Standard Modern Greek (SMG) ([Newton, 2013](#); [Tsiplakou, 2014](#)).

Pontic Greek is a Modern Greek dialect that ultimately derives from Koine Greek and is historically associated with the Pontus region on the southern Black Sea coast ([Sitaridou and Chatzikyriakidis, 2012](#); [Schreiber, 2018](#)).

Cretan Greek is the variety spoken on the island of Crete. Like many other Modern Greek dialects, it descends from Koine Greek ([Mackridge, 1985](#)).

Northern Greek is the form of Greek spoken in Thessaly, Thrace, and the Northeast Aegean.

Griko is a Greek variety spoken in Grecia Salentina in southern Italy, where it is recognized as

a minority language (Salminen, 1999). The dialect is written with the Latin alphabet (Chatzikyriakidis, 2010).

Heptanesian is a Modern Greek dialect spoken in the Ionian Islands. The dialect shows strong Venetian and Italian influence due to centuries of Venetian rule (Ralli, 2012).

Tsakonian, still spoken in the eastern Peloponnese, is a highly divergent Modern Greek variety considered by many to be a separate language, exceptional for descending directly from the ancient Doric dialect (Joseph et al., 1987; Mackridge, 2010).

Maniot is the variety spoken in the region of Laconian Mani in the southern Peloponnese (Trudgill, 2003).

4 Models Evaluated

We evaluate seven open-weight model variants, primarily drawn from the Llama family together with two Greek-specialized checkpoints. Open-weight models are required because perplexity computation relies on access to token-level log probabilities, which closed API models (closed GPT models, Claude, Gemini) do not expose. Although several systems share the same underlying architecture, they differ substantially along dimensions that are directly relevant to dialect modeling: training generation, parameter scale, and language specialization.

All models are evaluated strictly in a zero-shot setting. We use the publicly available checkpoints as released, without additional fine-tuning, continued pretraining, or exposure to Standard Modern Greek or dialectal evaluation data. Consequently, perplexity differences reflect the models’ inherent pretrained linguistic coverage rather than task-specific adaptation.

Our selection spans three dimensions. First, model generation: Llama 2 represents an earlier architecture with limited Greek tokenization, while Llama 3.x models show improved multilingual coverage. Second, model scale: we compare 7–8B parameter models with 70B variants to assess whether scale improves dialectal discrimination. Third, language specialization: Krikri-8B and Meltemi 7B are explicitly trained on Greek data, allowing us to test whether Greek-focused pretraining improves dialectal coverage.

- **General-purpose:** Llama 2 7B, Llama 3 8B, Llama 3.1 8B, Llama 3.1 70B, Llama 3.3 70B

- **Greek-specialized:** Krikri-8B², Meltemi 7B³

5 Metrics and Diagnostics

We incorporate multiple complementary metrics to capture different aspects of linguistic complexity and structural similarity. These measures fall into two groups: metrics that rely on LLM outputs and metrics computed directly from the raw text and tokenization, without any LLM inference. The only LLM-based metric is perplexity, computed across seven open-weight models using token-level log probabilities. The remaining metrics, Bits-Per-Character (BPC), Normalized Compression Distance (NCD), and Tokens per Word, are calculated solely from the input strings and their tokenization. These metrics capture different aspects of linguistic complexity and structural similarity. Each metric is described below.

- **Perplexity (PPL).** Measures how “surprised” a language model is by the text, defined as the exponential of cross-entropy loss. Lower PPL means the text is closer to the model’s training distribution. Perplexity is not an intrinsic property of a dialect, rather, it measures distance from what the model has learned as “Greek”. It is therefore model-dependent: a model trained on Pontic would show Pontic as closer and SMG as more distant. PPL is also confounded by genre, register and topic.
- **Bits-Per-Character (BPC).** An information-theoretic normalization of perplexity, defined as: $BPC = \log_2(PPL) \times \frac{\text{tokens}}{\text{characters}}$. BPC accounts for tokenization differences across models. A model with poor Greek tokenization (i.e., many tokens per word) would otherwise exhibit artificially high perplexity. BPC enables cross-model comparison. It shares the same theoretical limitations as PPL but is more robust to tokenizer effects.
- **Normalized Compression Distance (NCD).** Approximates Kolmogorov complexity by measuring shared information between texts using compression algorithms. NCD is model-agnostic and does not depend on LLM training data. While theoretically well-grounded, the NCD values in our dataset are tightly clustered (range 0.988–1.020), making it difficult

²Built on Llama 3.1 8B and further trained on Greek data.

³Developed by the same team as Krikri; they recommend Krikri 8B as their newest model.

to distinguish closely related dialects. It performs better for more distant languages, and is therefore included primarily for completeness, as PPL and BPC are more discriminative in this setting.

- **Tokens per Word.** A diagnostic metric measuring tokenizer coverage. Lower values indicate better vocabulary match. For SMG, values are 1.46 for Greek-specialized models (Krikri/Meltemi) versus 6.12 for Llama 2. The near-6 tokens-per-word score of Llama 2 reflects character-level fallback, indicating that the tokenizer fails to recognize Greek words.
- **Characters per Token.** The inverse of fragmentation, with higher values indicating more efficient subword units. SMG shows 4.4 characters per token for Greek-specialized models versus 1.05 for Llama 2. Values near 1.0 signal byte-level tokenization. Greek-specialized tokenizers (Krikri, Meltemi) achieve over four characters per token for SMG, dropping to approximately 2.3 for Griko, showing recognition of SMG morphemes but fragmentation of dialectal vocabulary.
- **Unique Token Ratio.** The proportion of unique tokens in the corpus serves as an indicator of lexical diversity, with higher values reflecting more diverse vocabulary usage. For SMG, this proportion is 21% for Krikri, compared to only 0.5% for Llama 2. Such low values indicate that the model repeatedly uses a very limited set of tokens, which is a typical signature of character-level fallback, where tokenization relies primarily on individual characters rather than lexical units. This measure therefore functions as a diagnostic for tokenizer quality.
- **Coefficient of Variation (CV) of PPL.** Defined as the standard deviation divided by the mean of window-level perplexity, measuring text homogeneity. Low CV indicates consistent text (e.g., Cretan rhymed folkloric material: 0.22–0.24), while high CV reflects a heterogeneous corpus (e.g., Pontic: 0.81–0.90). The low CV for Cretan confirms genre homogeneity, meaning that all windows are similarly surprising because they belong to the same poetic style.

- **Inter-metric Correlations.** We examined the relationships between our different metrics to assess whether they provide consistent signals about dialect similarity. The high correlation between PPL and BPC (mean $r = 0.94$ across models) is partly definitional, as BPC is derived from PPL. The correlation between PPL and NCD (mean $r = 0.87$, range 0.76–0.93) is more informative, indicating convergence between independent methods. The agreement of multiple metrics on the same dialect ordering (Heptanesian closest, Griko and Tsakonian most distant) provides stronger evidence than any single metric alone.

PPL–NCD correlations were computed between model perplexities and a model-agnostic Normalized Compression Distance vector calculated relative to Standard Modern Greek (NCD(dialect, SMG)); correlations are r unless otherwise noted.

6 Results

6.1 Perplexity Across Models

Table 1 presents perplexity measurements across all seven models. As expected, Standard Modern Greek obtains the lowest perplexity for all models, reflecting its dominance in training corpora. Perplexity increases systematically for dialects that diverge more substantially from SMG, with the highest values for Tsakonian and Griko, both highly divergent varieties. The dialect ranking is quite consistent: Heptanesian appears closest to SMG, followed by Cypriot and Maniot, then Pontic and Northern, with Cretan, Tsakonian, and Griko showing the highest distances.

Model-wise, even the largest models (Llama 3.1 70B and Llama 3.3 70B) show substantial gaps between SMG and the most divergent dialects. The Greek-specialized models (Krikri and Meltemi) offer small gains for some dialects but do not consistently outperform general-purpose models and still yield high perplexity overall. This suggests that Greek-focused training data does not ensure broad dialectal coverage, likely due to limited dialectal data. An anomaly appears in Meltemi, where Heptanesian scores unexpectedly lower than SMG. This finding may be due to the close similarity to SMG, given that the model has not been trained on dialectal data.

The Griko/SMG ratio ranges from 3.6× (Llama 2 7B) to 14.5× (Llama 3.1 8B), reflecting differences

in tokenization and Greek variety coverage, and underscoring the extreme difficulty these highly distinct dialects pose for the models.

6.2 Tokenization Effects

Table 2 shows clear differences in tokenization efficiency across both dialects and models. SMG consistently receives the fewest tokens per word, while more distinct varieties like Northern, Pontic, and Griko are segmented less efficiently. General-purpose models (Llama variations) show limited sensitivity to Greek morphological structure, whereas Greek-specialized tokenizers (Krikri, Meltemi) produce both lower overall token counts and a distribution that aligns with the expected dialectal distances.

The patterns help contextualize the perplexity results. With 6.12 tokens per word for SMG, compared to 1.46 for Greek-specialized models, Llama 2 operates at near-character level. The model predicts characters rather than linguistic units, compressing the perplexity range and reducing sensitivity to dialectal structure. However, even in this way, the perplexity measures here follow the same pattern as the rest of the models in terms of dialectal distance ranking.

Greek-specialized tokenizers (Krikri, Meltemi) show efficient tokenization for SMG (1.46–1.47 tok/word) that degrades for more distant dialects (2.40–2.44 for Griko). This 1.67× decrease in tokenization efficiency contributes directly to the larger perplexity gaps observed in these models, reflecting their greater sensitivity to genuine linguistic divergence. Although tokens per word provide insight into tokenization efficiency, we did not compute a quantitative baseline such as token overlap between dialects. Incorporating such a measure in future work could more precisely relate tokenization patterns to genuine linguistic divergence.

6.3 Model Comparison

Table 3 summarizes model performance and presents a comparison between the variety with the lowest perplexity (SMG) and the highest perplexity (Griko), based on a sample of 25,000 words. The results show a consistent and substantial performance gap between the two varieties for every model.

Across the board, SMG yields consistently lower perplexity values, reflecting its status as the standard variety and its strong representation in training data. Griko, on the other hand, systematically ex-

hibits the highest perplexity, confirming its position as the most challenging dialect for all models. Despite the transliteration of its Latin alphabet into Greek, this heavily Italian-influenced variety differs significantly from the other varieties.

The PPL ratio quantifies this contrast, revealing that modeling Griko is between 3.6× and 14.5× more difficult than modeling SMG, depending on the architecture. Llama 3.1 8B shows the highest sensitivity (14.5× ratio, 2.47 BPC range), while Llama 2 7B shows the lowest (3.6×, 1.83). The 70B models show slightly lower ratios than their 8B counterparts, a fact that might point to memorization effects that smooth over dialectal irregularities. In addition, there is a substantial gap in overall perplexity between the smallest Llama model and the larger variants, with the 7B model consistently yielding higher scores across conditions. This difference is likely attributable to model scale, as smaller models have more limited representational capacity, resulting in weaker probability estimates overall. The pattern aligns with well-established scaling effects in language modeling, where increases in parameter count systematically improve perplexity.

The Greek-specific models (Krikri-8B, Meltemi 7B) still display large performance gaps between SMG and Griko, since they exhibit very high perplexity on Griko and large PPL ratios, indicating limited robustness to this minority variety. While Meltemi 7B shows a comparatively lower ratio, this is probably driven by its unusually high perplexity on SMG rather than improved modeling of Griko. Overall, these results suggest that specialization in Greek does not automatically translate into effective coverage of all the varieties, confirming the underrepresentation of such dialects even in language-specific models.

Bits-Per-Character (BPC) normalizes perplexity for tokenization differences, enabling fairer cross-model comparison. The BPC range (Griko minus SMG) is relatively consistent across models (1.83–2.47) compared to raw perplexity ratios, indicating that part of the observed variation indeed stems from differences in tokenization. However, the stability of the BPC range across architectures also suggests that the persistent performance gap between SMG and Griko points to deeper lexical, morphological, and distributional mismatches between the two varieties.

Dialect	General-Purpose Models					Greek-Specialized	
	Llama 2 7B	Llama 3 8B	Llama 3.1 8B	Llama 3.1 70B	Llama 3.3 70B	Krikri 8B	Meltemi 7B
SMG	2.91	5.17	6.15	4.69	4.81	8.83	39.43
Heptanesian	3.76	8.87	11.74	8.41	8.83	12.36	25.38*
Cypriot	4.79	18.18	25.40	17.50	18.48	24.74	51.86
Maniot	5.12	18.66	25.91	18.62	20.07	24.73	55.59
Pontic	6.47	23.09	32.05	23.57	24.55	37.15	48.95
Northern	7.53	33.32	49.62	36.13	37.47	41.63	57.67
Cretan [†]	6.72	41.35	63.02	45.96	48.08	62.07	123.92
Tsakonian	7.98	46.81	67.61	51.41	52.75	96.54	192.55
Griko [‡]	10.46	59.66	89.32	59.45	61.86	121.57	208.76
Ratio (G/S)	3.6×	11.5×	14.5×	12.7×	12.9×	13.8×	5.3×

Table 1: Perplexity (PPL) by Model and Dialect (25,000 words).

[†]Primarily rhymed folkloric material. [‡]Transliterated from Latin. *Lower than SMG (anomaly). Ratio = Griko/SMG.

Dialect	Llama 2	Llama 3.x	Krikri	Meltemi
SMG	6.12	2.38	1.46	1.47
Heptanesian	5.79	2.41	1.63	1.63
Cypriot	6.04	2.62	1.85	1.86
Maniot	5.78	2.60	1.94	1.91
Cretan	5.33	2.40	1.83	1.79
Pontic	5.44	2.84	2.17	2.24
Northern	5.86	2.99	2.38	2.45
Tsakonian	5.39	2.73	2.12	2.11
Griko	5.51	3.00	2.44	2.40
Ratio (G/S)	0.90×	1.26×	1.67×	1.63×

Table 2: Tokenization: Tokens per Word (25,000 words).

Llama 3.x = shared tokenizer across Llama 3, 3.1, 3.3.

6.4 Inter-Metric Correlations

Table 4 shows correlations between PPL, BPC, and NCD across all seven models. PPL and BPC are strongly correlated (mean $r = 0.94$), which is partly definitional since BPC is derived from PPL. The PPL–NCD correlation (mean $r = 0.87$, range 0.76–0.93) is more informative, as it links two independent methods: neural language modeling and compression-based distance. This convergence suggests that both metrics capture similar distributional properties. Greek-specialized models (Krikri, Meltemi) show somewhat lower PPL–NCD correlations, possibly reflecting their different tokenization strategies.

6.5 Stability Across Sample Sizes

We evaluated all models at 5,000, 15,000, and 25,000 words. Table 5 presents the coefficient of

Model	SMG PPL	Griko PPL	PPL Ratio	BPC Range
Llama 2 7B	2.91	10.46	3.6×	1.83
Llama 3 8B	5.17	59.66	11.5×	2.25
Llama 3.1 8B	6.15	89.32	14.5×	2.47
Llama 3.1 70B	4.69	59.45	12.7×	2.30
Llama 3.3 70B	4.81	61.86	12.9×	2.32
Krikri-8B	8.83	121.57	13.8×	2.28
Meltemi 7B	39.43	208.76	5.3×	2.06

Table 3: Model Comparison Summary (25,000 words).
BPC Range = Griko BPC – SMG BPC.

Model	PPL–BPC	PPL–NCD	BPC–NCD
Llama 2 7B	.99***	.93***	.96***
Llama 3 8B	.95***	.90**	.96***
Llama 3.1 8B	.94***	.90**	.96***
Llama 3.1 70B	.95***	.90**	.96***
Llama 3.3 70B	.94***	.90**	.96***
Krikri-8B	.93***	.83*	.94***
Meltemi 7B	.89**	.76*	.93***
Mean	.94	.87	.95

Table 4: Inter-Metric Correlations by Model ($n = 8$ dialects).

* $p < .05$, ** $p < .01$, *** $p < .001$.

variation (CV) across sample sizes for all models. The dialect ranking remains stable across sample sizes, with most CV values below 10%.

The most distant varieties show the highest stability: Griko (CV 0.1–2.7%) and Tsakonian (0.2–3.2%) yield consistent estimates across all models. Cretan also shows low variance (1.0–3.9%), reflecting the stylistic homogeneity of a corpus, a substantial part of which consists

mainly of rhymed folkloric material. Northern shows the highest variance across most models (CV 9.7–12.4%), pointing to a more heterogeneous corpus. Similarly, Pontic shows elevated variance for Greek-specialized models (8.2–12.3%), likely reflecting the diversity of sources: Wikipedia articles, theatrical plays, jokes, and songs.

Meltemi shows anomalously high CV for SMG (22.9%), consistent with the corpus composition issues noted earlier. General-purpose models show lower variance for SMG (2.2–4.2%) than Greek-specialized models, possibly reflecting more stable SMG representation in their training data.

Table 6 confirms that the Griko/SMG ratio is stable across sample sizes for most models. Llama 2 7B shows consistent low sensitivity (3.3–3.6 \times), while Llama 3.x models cluster around 11–14 \times . The notable exception is Meltemi, whose ratio drops from 9.4 \times at 5k to 5.3 \times at 25k. The 25,000-word results represent the most reliable estimates, as increasing sample size reduces sampling variability and leads to the stabilization of frequency-based measures.

7 Discussion

7.1 Dialect-Specific Findings

Tsakonian is shown to be the second most distant variety in our findings (PPL 47–193 across models). This is in line with its status in Greek dialectology. For example, this aligns with Nicholas’s (2019) lexicostatistical analysis that dates the divergence of Tsakonian back to approximately 800 CE and, furthermore, seems to reflect fundamental structural differences stemming from this divergence. Notably, Tsakonian shows better tokenization efficiency than Northern or Griko, indicating that its extreme perplexity reflects true divergence rather than tokenization artifacts.

Griko shows the highest distance across all models (PPL 59–209), reflecting centuries of complete geographic isolation following Byzantine decline in Southern Italy combined with extensive Romance contact (Manolessou, 2005). Transliteration from Latin script may introduce additional artifacts, though the consistency of Griko’s extreme distance across models suggests this effect is secondary. Additional experiments using the original Latin script texts yielded almost identical results, indicating that script choice does not substantially affect the measurements.

Cretan shows anomalously high perplexity (PPL

41–124) despite efficient tokenization close to SMG levels. We attribute this to corpus composition rather than linguistic distance: the Cretan data consists primarily of rhymed folkloric material, such as mantinades, traditional rhyming couplets with fixed meter and formulaic expressions. The models are surprised by poetic structure, not dialectal features. Supporting this interpretation, Cretan shows the lowest coefficient of variation across all models (0.22–0.24), indicating homogeneous text that is consistently surprising.

Cypriot appears surprisingly close to SMG (PPL 18–52) despite preserving substantial structural distinctiveness, including phonemic geminates (Arvaniti, 2001) and morphosyntactic conservatism (Newton, 1972). These features may be underrepresented in orthographically normalized text, and Cypriot’s substantial online presence as the native variety of Cyprus, likely reduces model surprise through training exposure. This contamination hypothesis requires empirical verification.

Heptanesian is closest to SMG across all models (PPL 8–25), confirming its historical contribution to standardization. Trudgill (2003) notes that the Ionian Islands supplied most of the input into Standard Greek, and the Heptanesian School of literature directly shaped the emerging Demotic standard. The Meltemi result (Heptanesian < SMG) potentially reflects corpus composition issues rather than genuine linguistic relationships, but needs further investigation.

Pontic occupies a mid-range position despite traditional classification as highly divergent. The variety preserves infinitive constructions (Sitariidou, 2014), SOV vestiges, distinctive clitic placement (Condoravdi and Kiparsky, 2002), and archaic negation (Drettas, 1997), yet shows only moderate perplexity. Training data exposure offers a plausible explanation: Pontic has its own Wikipedia likely included in LLM training, and its poor tokenization efficiency (which typically inflates perplexity) fails to push scores higher, consistent with contamination reducing apparent distance.

Northern varieties show mid-range distance consistent with their classification as *idiomata* rather than *dialektoi* (Kontossopoulos, 2001). The defining vowel phenomena (Newton, 1972; Topintzi and Baltazani, 2012) represent phonological rather than morphosyntactic divergence, and in orthographically normalized text these differences may be partially obscured.

Dialect	General-Purpose Models					Greek-Specialized	
	Llama 2 7B	Llama 3 8B	Llama 3.1 8B	Llama 3.1 70B	Llama 3.3 70B	Krikri 8B	Meltemi 7B
SMG	3.9	2.6	4.2	2.2	2.2	8.9	22.9
Heptanesian	5.0	7.9	8.8	7.5	7.7	3.6	4.2
Cypriot	4.0	5.5	6.2	4.3	4.7	1.8	5.1
Maniot	3.9	6.2	5.4	8.0	7.7	8.6	5.5
Pontic	0.8	5.1	5.0	5.6	5.4	8.2	12.3
Northern	4.9	9.7	11.4	10.7	10.8	11.0	12.4
Cretan [†]	1.0	2.5	2.4	3.9	3.9	1.9	1.0
Tsakonian	0.2	1.9	2.0	3.2	2.9	3.2	2.2
Griko [‡]	0.1	1.2	0.6	2.5	1.9	2.5	2.7

Table 5: Coefficient of Variation (%) Across Sample Sizes (5k, 15k, 25k words).

[†]Rhymed folkloric material. [‡]Transliterated. Lower CV = more stable estimates.

Model	5k	15k	25k
Llama 2 7B	3.3×	3.6×	3.6×
Llama 3 8B	10.9×	10.7×	11.5×
Llama 3.1 8B	13.1×	13.5×	14.5×
Llama 3.1 70B	13.3×	12.7×	12.7×
Llama 3.3 70B	13.4×	12.7×	12.9×
Krikri-8B	16.2×	12.3×	13.8×
Meltemi 7B	9.4×	5.5×	5.3×

Table 6: Griko/SMG Perplexity Ratio Across Sample Sizes.

Meltemi variance reflects SMG corpus issues.

Maniot consistently occupies an intermediate position, sharing the archaic $v \rightarrow /u/$ with Tsakonian but deriving from Koine Greek (Kontossopoulos, 2001).

7.2 Methodological Implications

Three findings have implications for perplexity as a dialectometric tool. First, tokenization effects dominate cross-model comparison: Llama 2’s compressed range (3.6× ratio vs. 11–14× for others) reflects character-level tokenization rather than reduced dialectal sensitivity, and bits-per-character normalization following Mielke et al. (2018) enables fairer comparison. Second, genre confounds the dialect signal; as the Cretan case demonstrates, corpus composition can inflate perplexity independent of linguistic distance, and future work should control for this through stratified sampling. Third, training data exposure reduces perplexity in ways that may mask linguistic distance; Cypriot’s and Pontic’s moderate scores despite substantial traditional divergence may reflect model familiarity, and contamination detection methods (Shi et al.) could help disentangle these effects.

Our results produced rankings that are to a large degree in agreement with traditional dialectologi-

cal knowledge and classification. Tsakonian’s large distance can be seen as evidence of earlier branch divergence (Nicholas, 2019), while for Griko, it points to Romance contact as well as isolation (Manolessou, 2005). The minimal distance we find in Heptanesian confirms its contribution to language standardization (Trudgill, 2003). However, Pontic’s mid-range position somewhat contradicts traditional classification, possibly reflecting training contamination. Cretan also shows surprisingly high perplexity. However, this likely reflects genre rather than dialect.

Lastly, the interesting finding that Greek-specialized models do not consistently outperform more general-purpose models suggests that these models have minimal dialectal knowledge, and that Greek-focused pre-training does not guarantee better results on dialectal varieties. Furthermore, this is commensurate with the observation by Fleisig et al. (2024) that LLMs exhibit standard language ideology, and thus produce degraded performance on non-standard varieties.

7.3 Validation Against Literature

Our results seem to align largely, but not fully, with the knowledge we have about these dialects from a theoretical standpoint. Heptanesian Greek is indeed considered to have proximity to SMG, given its historical contribution to the standard language. Tsakonian’s distance comes as no surprise and confirms its uniqueness as the only dialect with direct descentance from Doric. Likewise, Griko’s distance reflects its heavy Romance borrowing and contact. The convergence of multiple metrics (PPL, BPC, NCD) on rankings consistent with expert linguistic assessment validates perplexity as a tool for dialectometry, with appropriate caveats about

genre and training data exposure.

Limitations and Future Work

The Cretan corpus consists primarily of rhymed folkloric material, such as the rhymed couplets called *mantinades*, introducing genre confounds that inflate perplexity independent of linguistic distance. The Griko corpus required transliteration from Latin script, potentially affecting results. Corpus sizes are limited to 25,000 words maximum in order to be more representative, given that for some dialects we do not have considerably more data than this number. In future work, we can test larger corpus sizes, at least for the dialects that do have available data. Additionally, perplexity measurements may be confounded by training data exposure: well-documented varieties like Cypriot may appear artificially close to SMG. Finally, Fleisig et al. (2024) document that LLMs show systematic bias against non-standard varieties, and thus our rankings may conflate true dialectal divergence with standard language ideology.

In future work, we aim to test the contamination hypothesis by training fine-grained Dialect Identification classifiers and examine data contamination issues by sampling data that some of our used models have been trained on, looking for dialectal contamination. Furthermore, we plan to extend the size of the datasets where possible and also create genre-controlled protocols. To complement perplexity-based metrics, future work could also explore combining LLMs with other NLP approaches, such as word embeddings or machine learning classifiers, to provide more robust measures of dialectal similarity.

Acknowledgments

Stergios Chatzikiyriakidis gratefully acknowledges funding from Amazon (project: Neural-Symbolic Integration for Enhanced Natural Language Processing (NIELS)) that provided computational support for the fine-tuning experiments described in the paper. Stergios Chatzikiyriakidis is also partially funded by the European Union (ERC ADG, PhylProGramm, 101096554). 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 Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Erofilis Psaltaki received funding from the Euro-

pean Union’s Horizon Europe research and innovation program under the Marie Skłodowska-Curie grant agreement No 101177564—HAIF. Cofunded 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 (REA). Neither the European Union nor the granting authority can be held responsible for them.

Veronika Laippala and Erik Henriksson received funding from the Research Council of Finland through “FIN-CLARIAH research infrastructure” (project 358720, which has also received funding from the European Union – NextGenerationEU instrument) and “Mechanisms of register variation in massively multilingual web-scale corpora” (project 362459); Erik Henriksson additionally received funding through “Green NLP - controlling the carbon footprint in sustainable language technology” (project 353167). Furthermore, We also wish to acknowledge CSC – IT Center for Science Ltd. for providing computational resources.

References

- Zachary Ankner, Cody Blakeney, Kartik Sreenivasan, Max Marion, Matthew L Leavitt, and Mansheej Paul. 2024. Perplexed by perplexity: Perplexity-based data pruning with small reference models. *arXiv preprint arXiv:2405.20541*.
- Amalia Arvaniti. 2001. Cypriot Greek and the phonetics and phonology of geminates. *MIT Working Papers in Linguistics*, 41:19–36.
- Jeremy Barnes, Samia Touileb, Petter Mæhlum, and Pierre Lison. 2023. Identifying token-level dialectal features in social media. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 146–158.
- Stavros Bompolas and Dimitra Melissaropoulou. 2025. Understanding dialectal variation in contact scenarios through dialectometry: Insights from inner asia minor greek. *Languages*, 10(1):13.
- Stergios Chatzikiyriakidis. 2010. *Clitics in four dialects of Modern Greek: A dynamic account*. Ph.D. thesis, University of London.
- Stergios Chatzikiyriakidis, Dimitris Papadakis, Sevastioanna Papaioannou, and Erofilis Psaltaki. 2025. Grdd+: An extended greek dialectal dataset with cross-architecture fine-tuning evaluation. *arXiv preprint arXiv:2511.03772*.
- Stergios Chatzikiyriakidis, Chatrine Qwaider, Ilias Kolokousis, Christina Koula, Dimitris Papadakis,

- and Efthymia Sakellariou. 2023. Grdd: A dataset for greek dialectal nlp. *arXiv preprint arXiv:2308.00802*.
- Cleo Condoravdi and Paul Kiparsky. 2002. Clitics and clause structure. *Journal of Greek Linguistics*, 2:1–39.
- Nathan Cooper and Torsten Scholak. 2024. Perplexed: Understanding when large language models are confused. *arXiv preprint arXiv:2404.06634*.
- Georges Drettas. 1997. *Aspects pontiques*. Association de recherches pluridisciplinaires, Paris.
- Eve Fleisig, Suchin Gururangan, and Noah A. Smith. 2024. Standard language ideology in NLP. *arXiv preprint arXiv:2404.11968*.
- David Holton and Io Manolessou. 2010. Medieval and early modern Greek. *Cambridge Encyclopedia of the Language Sciences*, pages 481–484.
- Weihang Huang, Akira Murakami, and Jack Grieve. 2025. Attributing authorship via the perplexity of authorial language models. *PloS one*, 20(7):e0327081.
- Brian D Joseph, Irene Philippaki-Warbuton, and Irene Philippaki-Warbuton. 1987. *Modern Greek*. Croom Helm London.
- Vani Kanjirangat, Tanja Samardzic, Ljiljana Dolamic, and Fabio Rinaldi. 2023. Optimizing the size of subword vocabularies in dialect classification. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 14–30.
- Vani Kanjirangat, Tanja Samardzic, Ljiljana Dolamic, and Fabio Rinaldi. 2025. Tokenization and representation biases in multilingual models on dialectal nlp tasks. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24003–24021.
- Nikolaos Kontossopoulos. 2001. *Dialektoi kai Idiomata tis Neas Ellinikis*. Gregoris, Athens. In Greek.
- Fangru Lin, Shaoguang Mao, Emanuele La Malfa, Valentin Hofmann, Adrian de Wynter, Xun Wang, Si-Qing Chen, Michael J Wooldridge, Janet Pierrehumbert, and Furu Wei. 2025. Assessing dialect fairness and robustness of large language models in reasoning tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6317–6342.
- Peter Mackridge. 1985. *The Modern Greek Language: A Descriptive Analysis of Standard Modern Greek*. Oxford University Press, Oxford.
- Peter Mackridge. 2010. Modern greek. *A Companion to the Ancient Greek Language*, pages 564–587.
- Io Manolessou. 2005. The Greek dialects of Southern Italy: An overview. In *Cambridge Papers in Modern Greek*, volume 13, pages 103–125.
- Sabrina J. Mielke, Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner. 2018. Spell once, summon anywhere: A two-level open-vocabulary language model. In *Proceedings of NAACL-HLT*, pages 1661–1671.
- Brian Newton. 1972. *The Generative Interpretation of Dialect: A Study of Modern Greek Phonology*. Cambridge University Press.
- Brian Newton. 2013. *Cypriot Greek: Its phonology and inflections*, volume 121. Walter de Gruyter.
- Nick Nicholas. 2019. A critical lexicostatistical examination of Ancient and Modern Greek and Tsakonian. *Journal of Applied Linguistics and Lexicography*, 1(1):18–68.
- Eileen Pan, Anna Seo Gyeong Choi, Maartje Ter Hoeve, Skyler Seto, and Allison Koenecke. 2025. Analyzing dialectal biases in llms for knowledge and reasoning benchmarks. *arXiv preprint arXiv:2510.00962*.
- Angela Ralli. 2012. Verbal loanblends in griko and heptanesian: a case study of contact morphology. *L’Italia Dialettale*, 73:111–132.
- Tapani Antero Salminen. 1999. *UNESCO red book on endangered languages: Europe*. Helsingin Yliopisto [Host].
- Laurentia Schreiber. 2018. 6.4. romeyka. In *The Languages and Linguistics of Western Asia*, pages 892–934. De Gruyter Mouton.
- Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer.
- Kiril Simov and Petya Osenova. 2007. Applying a normalized compression metric to the measurement of dialect distance. *Serdica Journal of Computing*, 1(1):73–86.
- Ioanna Sitaridou. 2014. The Romyka infinitive: Continuity, contact and change in the Hellenic varieties of Pontus. *Diachronica*, 31(1):31–73.
- Ioanna Sitaridou and Stergios Chatzikyriakidis. 2012. Cultural survival shifts focus: The case of pontic greek. *When empires clash: Modern-day outcomes of historical Greek and Turkish language encounters*, 4:29.
- Charalambos Themistocleous. 2017. Dialect classification using vowel acoustic parameters. *Speech Communication*, 92:13–22.
- Charalambos Themistocleous. 2019. Dialect classification from a single sonorant sound using deep neural networks. *Frontiers in Communication*, 4:64.
- Nina Topintzi and Mary Baltazani. 2012. The acoustics of high-vowel loss in a Northern Greek dialect and typological implications. In *Consonant Clusters and Structural Complexity*, pages 369–398. De Gruyter Mouton, Berlin.

Peter Trudgill. 2003. Modern greek dialects: A preliminary classification. *Journal of Greek linguistics*, 4(1):45–63.

Stavroula Tsiplakou. 2014. How mixed is a ‘mixed’ system?: The case of the cypriot greek koiné. *Linguistic Variation*, 14(1):161–178.

Arthur Wuhrmann, Andrei Kucharavy, and Anastasiia Kucherenko. 2025. Low-perplexity llm-generated sequences and where to find them. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 774–783.

A Results at Different Sample Sizes

Tables 7 and 8 present full perplexity results at 5,000 and 15,000 word sample sizes.

Dialect	General-Purpose Models					Greek-Specialized	
	Llama 2 7B	Llama 3 8B	Llama 3.1 8B	Llama 3.1 70B	Llama 3.3 70B	Krikri 8B	Meltemi 7B
SMG	3.15	5.49	6.82	4.71	4.83	7.59	22.16
Heptanesian	4.07	9.67	12.93	9.15	9.64	12.65	26.53
Cypriot	5.25	20.59	29.23	19.26	20.56	23.99	46.39
Maniot	4.91	17.79	25.17	17.33	18.71	24.10	54.90
Pontic	6.35	25.63	35.12	26.42	27.41	43.07	65.41
Northern	8.44	40.36	62.31	44.84	46.85	50.20	67.49
Cretan [†]	6.88	43.98	66.78	50.63	52.84	65.04	123.72
Tsakonian	8.02	44.73	64.53	47.60	49.26	89.16	182.70
Griko [‡]	10.48	59.77	89.61	62.63	64.60	123.16	208.61
Ratio (G/S)	3.3×	10.9×	13.1×	13.3×	13.4×	16.2×	9.4×

Table 7: Perplexity (PPL) by Model and Dialect (5,000 words).

[†]Primarily rhymed folkloric material. [‡]Transliterated from Latin.

Dialect	General-Purpose Models					Greek-Specialized	
	Llama 2 7B	Llama 3 8B	Llama 3.1 8B	Llama 3.1 70B	Llama 3.3 70B	Krikri 8B	Meltemi 7B
SMG	2.90	5.44	6.56	4.92	5.05	9.43	35.99
Heptanesian	3.61	7.97	10.42	7.60	7.97	11.62	23.92
Cypriot	4.88	18.58	26.13	17.73	18.85	23.72	51.67
Maniot	5.40	20.61	28.50	21.01	22.49	29.10	61.97
Pontic	6.44	23.00	31.29	23.45	24.40	35.72	53.92
Northern	8.25	41.91	65.14	46.49	48.19	54.51	78.23
Cretan [†]	6.85	42.98	65.81	48.27	50.37	63.23	126.52
Tsakonian	8.01	45.35	65.29	49.09	50.33	93.00	185.65
Griko [‡]	10.46	58.24	88.43	62.70	64.23	116.06	197.16
Ratio (G/S)	3.6×	10.7×	13.5×	12.7×	12.7×	12.3×	5.5×

Table 8: Perplexity (PPL) by Model and Dialect (15,000 words).

[†]Primarily rhymed folkloric material. [‡]Transliterated from Latin.

A Subword Embedding Approach for Variation Detection in Luxembourgish User Comments

Anne-Marie Lutgen, Alistair Plum, Christoph Purschke

University of Luxembourg, Esch-sur-Alzette, Luxembourg

{anne-marie.lutgen, alistair.plum, christoph.purschke}@uni.lu

Abstract

This paper presents an embedding-based approach to detecting variation without relying on prior normalisation or predefined variant lists. The method trains subword embeddings on raw text and groups related forms through combined cosine and n-gram similarity. This allows spelling and morphological diversity to be examined and analysed as linguistic structure rather than treated as noise. Using a large corpus of Luxembourgish user comments, the approach uncovers extensive lexical and orthographic variation that aligns with patterns described in dialectal and sociolinguistic research. The induced families capture systematic correspondences and highlight areas of regional and stylistic differentiation. The procedure does not strictly require manual annotation, but does produce transparent clusters that support both quantitative and qualitative analysis. The results demonstrate that distributional modelling can reveal meaningful patterns of variation even in “noisy” or low-resource settings, offering a reproducible methodological framework for studying language variety in multilingual and small-language contexts.

1 Introduction

Variation in language is often treated as noise in NLP pipelines (Eisenstein, 2013; Al Sharou et al., 2021). Spelling differences, orthographic inconsistencies, and regional forms are typically normalised or removed to simplify token space, which can erase sociolinguistic signal (Baron and Rayson, 2008). Work in sociolinguistics and large-corpus dialectology shows that such variation is systematic and informative for geography and social structure (Grieve et al., 2019). Subword modelling has long been used to handle non-standard forms in practice and improves classification in noisy settings (Munro and Manning, 2010).

For under-researched languages and varieties, identifying and extracting language variation re-

mains challenging. Pre-processing tools such as VARD insert modern equivalents for historical spellings to aid search and tagging, relying on lexicons and edit-distance style methods, and target normalisation rather than discovery (Baron and Rayson, 2008). Embedding-based studies indicate that distributional models encode many types of spelling variation and near-orthographic similarity, though evaluations are typically based on curated sets of variant pairs and focus on representation quality (Nguyen and Grieve, 2020). Closely related work in context-sensitive spelling correction uses word and character n-gram embeddings to map misspellings to canonical forms, again optimising correction and not mining new variants (Fivez et al., 2017). Research on dialectal change detection has brought models for geographic differences, but does not allow for directly extracting unconstrained orthographic families from raw text (Jiang et al., 2020; Pham et al., 2024). Broader multilingual analyses of surface-form overlap highlight that form-level variation carries structure that models can exploit (Kallini et al., 2025).

There are methods to represent variation and to correct it, and there are resources that label dialectal differences. What is less supported are methods that detect and mine previously unlisted variant families directly from raw text without seed lexicons, and that do so beyond strictly dialectal contrasts. We propose the methodology laid out in this paper to address this gap. With this methodology, we are able to discover candidate variation families from distributional evidence and provide transparent scores for downstream qualitative and quantitative analysis.

The main contributions of the research carried out and presented in this paper are:

- (1) A reproducible methodology¹ for inducing lexical and orthographic variation from raw

¹<https://github.com/plumaj/vadamt>

text using subword embeddings, similarity-based grouping, and controlled pruning without relying on predefined variant lists or normalisation rules.

- (2) A large-scale empirical study of variation in Luxembourgish user comments, showing that the automatically induced families capture systematic patterns and provide a structured basis for qualitative linguistic analysis.

2 Background

Luxembourgish is a small language situated in a dense multilingual environment, with extensive contact to both German and French. Its grammatical structure derives from Moselle Franconian (Gilles, 2019), while sustained contact with French has shaped its lexicon, borrowing patterns, and code-switching practices. Written Luxembourgish displays considerable orthographic and lexical diversity, especially in informal online settings, which makes it a challenge for NLP.

Initial work on computational methods for Luxembourgish was limited but established an initial foundation. Adda-Decker et al. (2008) introduced the first tools and corpora for automatic processing. Subsequent studies examined characteristic orthographic phenomena (Snoeren et al., 2010) and provided early annotated resources for mixed-language processing (Lavergne et al., 2014).

Recently, research activity has increased noticeably. Work has expanded to sentiment analysis (Sirajzade et al., 2020; Gierschek, 2022), orthographic correction (Purschke, 2020), syntactic annotation (Plum et al., 2024), topic classification (Philippy et al., 2024), comment moderation (Ranasinghe et al., 2023), and automatic normalisation (Lutgen et al., 2025). A broader set of classification tasks, including named entity recognition, was provided by Lothritz et al. (2022), and the generative benchmark LuxGen was introduced by Plum et al. (2025). These works illustrate the rapid growth of Luxembourgish NLP but also reveal gaps in coverage, consistency, and domain diversity.

Model development reflects a similar trajectory. Strategies range from cross-lingual transfer from German, as in LuxGPT (Bernardy, 2022), to data augmentation with synthetic Luxembourgish text in LuxemBERT (Lothritz et al., 2022), and balanced multilingual pretraining for LuxT5 (Plum et al., 2025). Other models include ENRICH4ALL (Anastasiou, 2022) for administrative-domain chat-

bots and the LUX-ASR speech recognition models (Gilles et al., 2023a,b). Together, these efforts demonstrate progress, yet available datasets remain fragmented and vary widely in size, annotation schemes, and linguistic phenomena.

One area that has received little explicit attention is lexical and orthographic variation. Lutgen et al. (2025) develop a qualitative performance test to evaluate normalisation models for specific orthographic variants. In linguistics, the Variation Atlas by Gilles (2021) represents the most comprehensive overview of phonological, lexical, grammatical and regional variants in Luxembourgish. This atlas is constructed by using an app to collect users’ speech inputs for specific phenomena and socio-demographic data, which is then transcribed, analysed, and published (Entringer et al., 2021).

3 Methodology

The methodology adopted in this study combines semi-supervised modelling with targeted qualitative analysis to identify lexical and orthographic variation directly from raw text, without relying on predefined dictionaries or normalisation rules. Throughout, spelling diversity is treated as a source of linguistic information rather than noise, allowing the unsupervised detection of previously unrecorded orthographic and mixed variants while ensuring transparency and reproducibility. This design supports large-scale induction alongside qualitative interpretation, and aligns with recent work arguing that normalisation can obscure meaningful patterns in non-standard and partly standardised varieties (Grieve et al., 2019; Kallini et al., 2025).

First, subword embeddings² are trained on the raw corpus to obtain distributional representations that preserve orthographic detail. Second, these embeddings are used to induce groups of related forms through a combination of cosine and n-gram similarity, followed by controlled pruning and aggregation across relevant *dimensions* such as users, time periods, or domains. Third, the automatically identified groups are examined manually to assess their linguistic coherence and to trace patterns that are not fully captured by numerical criteria.

3.1 Distributional Embeddings

Before outlining the methodology in more detail, we briefly characterise what word- and subword-

²In this work, we use the term subword embeddings to refer to embeddings constructed from fixed character n-grams, rather than learned segmentation-based subword vocabularies.

level embeddings encode. Distributional embeddings represent lexical items based on their patterns of co-occurrence in context, such that similarity in embedding space reflects shared semantic content, syntactic behaviour, and usage environments (Turney and Pantel, 2010; Levy and Goldberg, 2014). Subword models extend this principle by incorporating character-level information, which allows orthographically related forms to be represented closely even when token frequencies are low or surface forms differ (Bojanowski et al., 2017). As a consequence, embedding similarity reflects semantic relatedness, morphosyntactic similarity, and orthographic overlap. This makes clustering in embedding space a suitable operation for identifying candidate groups of lexical variants, particularly in settings where variation manifests through both form and contextual usage (Munro and Manning, 2010; Nguyen and Grieve, 2020).

3.2 Stage 1: Training Subword Embeddings

Embeddings are trained with FastText (Bojanowski et al., 2017) using the configuration values `vector_size`, `window`, `min_count`, `epochs`, `min_n`, `max_n`, `sg`. These values are estimated in accordance with the size of the corpus, as well as with some testing of the variant families (as detected in the following stage).

The input is a JSON file containing a required `text_field`. Optional fields specify the comparison dimension, such as `user_id` or `date`. The corpus is streamed to manage memory, and basic token statistics are collected. Cleaning behaviour is minimal: Mentions beginning with @ are removed before tokenisation, lowercasing is controlled by the `lowercase` flag.

3.3 Stage 2: Identifying Variant Families

After training, a candidate lexicon V is created from all tokens that meet the `min_count` threshold. For each seed $w \in V$, the method retrieves the top neighbours based on the values `open_TOPN` or `strict_TOPN`. Cosine similarity is computed as

$$\cos(\mathbf{w}, \mathbf{v}) = \frac{\mathbf{w} \cdot \mathbf{v}}{\|\mathbf{w}\| \|\mathbf{v}\|}.$$

Pairs are filtered according to the associated similarity threshold (`open_TH` or `strict_TH`). Then character n-gram Jaccard overlap is computed:

$$J(w, v) = \frac{|G(w) \cap G(v)|}{|G(w) \cup G(v)|},$$

where $G(\cdot)$ contains all n-grams in the range `min_n` to `max_n`. Cosine and Jaccard values jointly determine whether two tokens belong to the same group.

We implement two modes to help identify variant families. The *open* mode forms a local star around each seed. The *strict* mode builds an undirected graph and extracts connected components. Graph growth is limited by `DEGREE_CAP`. Groups that do not reach `SNN_MIN` members are removed. For the analysis presented in subsequent sections of this paper, we used *strict* mode.

Scoring and Pruning For each group F , we compute its size and the mean values of cosine similarity and Jaccard overlap. A cohesion score is the harmonic mean of these two averages. Groups are removed if they fail to reach the minimum size or if relative frequencies exceed the bound set by `MAX_FREQ_RATIO`. All pairwise scores are retained for inspection.

Dimension-Based Aggregation If a dimension field is provided, the method counts in how many distinct dimensions each variant appears and records the frequency of the most common dimension. For each variant we store its coverage, its top dimension, and the share that this dimension represents of its total frequency. These values feed into filters such as `MIN_USERS` and `MAX_FREQ_RATIO`. The summary CSV lists these quantities for all groups.

Parameter	Key	Default
Lowercasing	<code>lowercase</code>	<code>true</code>
Comparison dimension	<code>dimension</code>	<code>user_id</code>
Vector size	<code>vector_size</code>	100
Context window	<code>window</code>	5
Min frequency	<code>min_count</code>	10
Epochs	<code>epochs</code>	10
Skip-gram model	<code>sg</code>	1
Character n-gram range	<code>min_n-max_n</code>	3-7
Neighbours/seed (open)	<code>open_TOPN</code>	30
Similarity thr. (open)	<code>open_TH</code>	0.75
Neighbours/seed (strict)	<code>strict_TOPN</code>	100
Similarity thr. (strict)	<code>strict_TH</code>	0.73
Min family size	<code>SNN_MIN</code>	2
Degree cap	<code>DEGREE_CAP</code>	200
Min token length	<code>MIN_LEN</code>	3
Min users per variant	<code>MIN_USERS</code>	3
Max frequency ratio	<code>MAX_FREQ_RATIO</code>	25

Table 1: Main configuration parameters.

Configuration and Output Table 1 presents an overview of the parameters used and their defaults used for the purposes of this study. The method iterates through the vocabulary, computes cosine

and Jaccard scores where needed, and constructs the final groups. The output consists of a JSONL file containing all groups with their members and a CSV summary with the main statistics. As this is an experimental study, the configuration of the values is based mainly on trial and error, by checking the variant families manually after each run. In contrast to normalisation tools such as VARD (Baron and Rayson, 2008), the procedure retains surface forms and measures their similarity instead of mapping them to canonical variants.

3.4 Stage 3: Qualitative Analysis

The model outputs variant families with their cosine and Jaccard values and the frequency of each member of the group. The first step in the qualitative analysis is to manually go over the families and approach the analysis with a bottom-up method. Based on the chosen dimension (i.e. user, time, etc.), the families represent for instance user variants in a similar context or variants over time in a similar context. Then, adopting a bottom-up approach, classifying the families is a straightforward way to analyse the families based on the type of variation (orthographic, morphological, lexical, stylistic, regional, etc.). The families could also be semantically or functionally related, or the relation is not identifiable, which could be a category in itself. With the help of categorisation, the identification of patterns in the data is more feasible.

4 Luxembourgish User Comments

We demonstrate the use of our methodology in the context of Luxembourgish user comments. The comments are part of the online media platform RTL³, the main news broadcaster in Luxembourg and the only news broadcaster completely in Luxembourgish. The comments span from 2008 to 2024 and total roughly 1,42 million comments. As the use of Luxembourgish has been expanding in the written domain in the past 25 years and formal grammar teaching in school is still not properly regulated, we observe a high amount of variation in written texts. This is especially visible in informal domains, like online user comments. An in-depth analysis of variation in this domain in Luxembourgish has not been conducted yet. However, with our methodology, we can analyse a wide number of comments and classify the occurring variation.

³<https://rtl.lu>

After applying stage 1 and 2, using the users as the comparative dimension, we start with the qualitative analysis. Using a bottom-up approach, we classify the families into 7 distinct categories, which are depicted in Table 2. As the families have a different number of words and also different phenomena appearing in one family, we decided for a multi-label approach. One family can have up to 3 categories. We illustrate the frequency of each category in Table 2. In the following, we describe each category and highlight findings related to language variation in Luxembourgish.

Category	Frequency
Orthographic	394
Morphological	222
Lexical	115
Collocation	21
Tokenisation	14
Regional	8
Other	242

Table 2: Categories and frequency

Orthographic The orthographic category describes spellings that are not part of the official orthography (Zenter fir d’Lëtzebuurger Sprooch, 2019). This includes the use of different graphemes to express the same word, which are often based on the word’s phonological properties. Since Luxembourgish has a high phoneme-grapheme correspondence in addition to an ideology that you can write how you speak, the language presents a wide range of orthographic variation. One example is *laang* (lb. *long*) where one family includes the orthographic variant *lang*. The single vowel violates the quantity rule in the orthography (Zenter fir d’Lëtzebuurger Sprooch, 2019), as a long vowel is written doubly when more than one consonant follows. However, since the orthography is not well known, both variants appear frequently. Additionally, this category also includes families that encompass different lexemes that are spelled incorrectly. One example for this case is *krng*, *srng*, *dng*, *êng*, *öng* (lb. *none*, *his*, *yours*, *one*, *one*). The correct spelling is *keng*, *seng*, *deng*, *eng*, *eng*. In this instance, we can see two different spellings for *eng* which has two different sources. The first one *êng* represents more of a typical misspelling in comparison to *öng* which represents a phonological variant expressed with the choice of <ö>. The phonological variant of *eng* is pronounced with a rounded vowel which is then written as <ö> by some authors. This is still

#	Family	English	Standard
(1)	Zäit, Zeit, Zait, Zéit, Zaït	time	Zäit
(2)	mir, mer, mier, mär, maer, miir, mäer	we	mir
(3)	mat, matt, maat	with	mat
(4)	mecht, mécht, mescht, mëcht, mëscht	to do	mécht
(5)	wäit, weit, wait, wéit	far	wäit
(6)	sech, sëch, séch	himself	sech
(7)	Numm, Num	name	Numm
(8)	laang, lang	long	laang
(9)	Fehler, Feeler	mistake	Feeler

Table 3: Families in the category orthographic variation (One lexeme, different variants).

classified as orthographic variation since it violates the official orthography (Zenter fir d’Lëtzebuenger Sprooch, 2019).

Morphological The morphological category classifies all morphological variation. This encompasses conjugated verb forms, and inflections of nouns and adjectives in case, number, and gender. This also includes compounding nouns, clippings and conversions. One example is the family *fillen*, *fillt*, *fille* (lb. *to feel*) which includes the conjugated form *fillt* and the deletion of the final <-n> before specific characters, known as the n-rule, in *fille*.

Lexical This category includes different lexemes that are semantically related. This includes synonyms and antonyms like the family *méi*, *manner* (lb. *more*, *less*) and lexical variants like *dass*, *datt* (lb. *that*).

Collocation The collocation category describes families with lexical items that form conventionalised combinations in daily use. For this category, the distinct words in a family form a collocation together. For instance, *Gott*, *säi*, *Dank*⁴ (lb. *god*, *be*, *thanks*) forms one family of distinct items that frequently appear in the same context. Together, they form the collocation *Gott säi Dank* meaning *thank god*.

Tokenisation The tokenisation category describes families where the same word appears twice, but in one instance without the definite article *d’* (lb. *the*) attached to it and once with it attached to the word. One example for this is the family *Zukunft*, *d’Zukunft* (lb. *the future*).

Regional The regional category encompasses categories where regional varieties are visible in

the graphemic representation. This category overlaps with the orthographic category, as it violates the official orthography (Zenter fir d’Lëtzebuenger Sprooch, 2019). However, in this case the regional influence is visible and can be verified in the official Luxembourgish Variation Atlas (Gilles, 2021).

Other The other category includes families that have no distinct features or are not identifiable as meaningful words. This includes pragmatic expressions that are often used in online discourse like *bla*, *blabla*, *ehm*, *hoho*, *tzzz* or non-identifiable words like *fisk*, *ragnax*, *har*, *sed*.

4.1 Orthographic Variation: Insights

In this section, we present new insights from the orthographic category revealed by the clustering method. We examine several representative examples in more detail and discuss their impact.

One Lexeme, Different Variants We start with examples, where one lexeme has several orthographic varieties in one single family. As the families are constructed based on similar neighbouring tokens in the embedding space, this grouping is straightforward. Interestingly, often not only variants of the same lexeme are part of the family but also the correctly spelled variant. This indicates that at least for these families, the lexemes of the orthographic Luxembourgish and the non-orthographic variants of Luxembourgish are aligned in the embedding space (Cao et al., 2020). Additionally, this shows that the model is able to capture orthographic variants. In these cases, we either see multiple variants for one lexeme in one family, or only two variants, where one item is usually spelled correctly as indicated in Table 3. We observe that these categories include the most common orthographic variants in Luxembourgish.

⁴For readability, nouns in families are capitalized.

#	Family	English	Standard
(10)	déi, wéi, méi, ewéi	that, as, more, as	déi, wéi, méi, ewéi
(11)	dèi, wèi, mèi	that, as, more	déi, wéi, méi
(12)	dat, wat	this, which	dat, wat
(13)	dad, wad	this, which	dat, wat
(14)	weéi, deéi, eweéi, meéi	as, that, as, more	wéi, déi, ewéi, méi
(15)	méh, déh, wéh, ewéh	more, that, as, as	méi, déi, wéi, ewéi

Table 4: Families in the category orthographic variation (Different lexemes, one variant).

Especially in the families with only two variants, one variant is often the most common spelling variant.

Different Lexemes, One Variant The second insight for orthographic variation is the clustering of families of the same type of variation for different words across multiple families. For these families, we observe similar words, for instance *dèi, wèi, mèi* (Table 4 (11)) or *dad, wad* (Table 4 (13)) clustering together but with the same variant pattern in every word. A variant pattern is the same type of grapheme writing in different words, for example instead of the correct spelling of the diphthong <éi>, the graphemes <èi> are used consistently to represent the diphthong. Instances for these patterns are shown in Table 4. Additionally, not only variants cluster together but also the standard variants of these function words as shown in the instances (10) and (12) in Table 4. Since the instances are mostly function words, the clustering in itself is evident, however, the clustering of the identical variation pattern in different families indicates that these variants have social meaning which is represented in the embedding space. Therefore, the clustering shows that these words appear in similar topics or similar sentence structures that are linked via this variant. Another option would be a higher frequency for multiple authors writing about a similar topic. A more in-depth corpus analysis would give more insights into the social meaning of these variants.

4.2 Regional Variation: Insights

Overall, our method only made 8 instances of regional variation visible. Due to an advanced state of dialect levelling in Luxembourgish, regional dialects have evolved into a national variety with some remaining lexical, phonological, and grammatical features (Gilles, 1999). Due to the informal nature of the comments, some phonological

regional features are visible in the grapheme representation of the written lexeme. As these written forms differ from the official orthography, they also classify as orthographic variants. There are also instances of regional variants that are not orthographic variants but are part of the official dictionary for Luxembourgish⁵ like *mar* which is a regional variant of *muer* (lb. *tomorrow*). With the Variation Atlas (Gilles, 2021) we can verify specific variants that are part of the 811 variant maps included in the atlas.

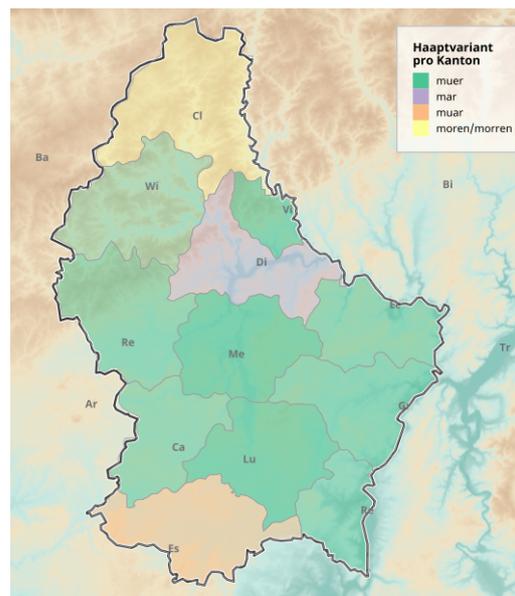


Figure 1: Map of *muer* variants (Gilles, 2021)

One of those variants is the family *muer, muar, moar* (lb. *tomorrow*). Figure 1 shows the map of the variant *muer* and the regional variants *mar, muar, moren*. Not shown in the map due to a low frequency of instances is the variant *moar*, which is also part of the family in our analysis. The variant *muar* is prevalent in the south of Luxembourg, whereas *muer* is the most common variant in Lux-

⁵<https://lod.lu>

embourg and considered the standard variant⁶. The variant *moar* is mostly localised in the east of the country, but only a few participants (19) of the variation atlas survey have used that variant. However, in the comments 338 instances are recorded in comparison to 604 instances of *muar*, and 6577 instances of *muer*. This shows that *moar* is still the least used variant in contrast to the most frequent variant *muer*. However, it is still a common variant used in the comments, which was not clearly recorded before as the findings of the Variation Atlas did not indicate this. Additionally, the Jaccard values show a low overlap between authors using these variants, which indicates a consistent use of a variant instead of switching between variants in different contexts.

4.3 Lexical Variation: Insights

The lexical variation category is one of the most heterogeneous categories in our analysis, as this category includes every family that encompasses different lexemes which are semantically related to each other. In addition to synonyms and antonyms, we also found common lexical variants in the data that are included in the Variation Atlas (Gilles, 2021).

Two interesting families are *séier*, *schnell*, *seier* (lb. *fast*) and *säit*, *seit*, *sait*, *zanter*, *zenter*, *zënter*, *séit* (lb. *since*). These families do not only show lexical variation, but also orthographic variation. In this section, we focus on the lexical side. Figure 2 illustrates the regional distribution of the use of *séier* and *schnell*. Overall, the frequency of use of both variants is nearly identical, at an almost equal split. The variant map illustrates some preference for the *séier* variant in the north of the country, and some for *schnell* in the south. However, the statistical analysis shows that only the north is significantly favouring the *séier* variant (Gilles, 2021). One factor that influences the use of this variant significantly is age. The older the participants of the variation atlas survey are, the more the variant *séier* is preferred.

Similarly, we can observe these tendencies with *säit*, *zënter* and *zanter*. Gilles (2021) illustrates in the Variation Atlas that the variant *säit* is overall favoured and regionally the most used variant. However, if we look at the factor age, the variant *zënter* and *zanter* are more popular with an older age group. It needs to be noted that *zënter* and *zanter* are phonological variants but belong the same

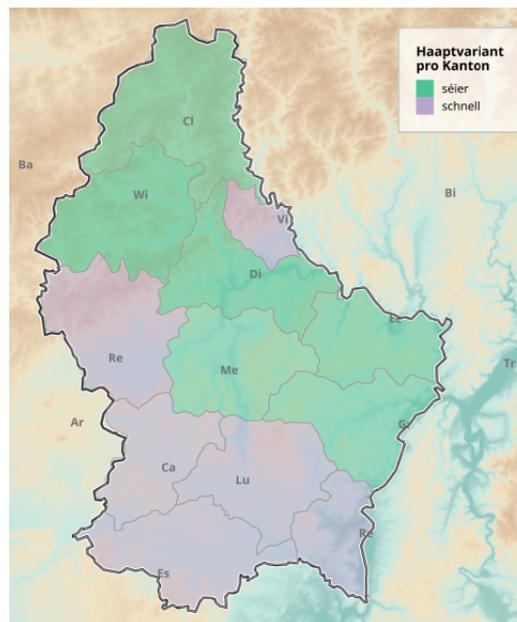


Figure 2: Map of *schnell* variants (Gilles, 2021)

lexical variant.

Further, we also found families that are topically related. The family *chat*, *chatgpt*, *gpt*, *chatbot* illustrates the more recent advances in AI and how these entered into the discourse in the user comments. Another example is the family *Grexit*, *Frexit*, *Brexit*, *iteers*, *Lexit*, *Nexit*, *Luxexit* which encompasses word formations inspired by the expression *Brexit*. By combining country names with the word *exit* (e.g. *France* = *Frexit*, *Netherlands* = *Nexit*, *Luxembourg* = *Luxexit*) we get a family of variants that is topically related to a withdrawal from the European Union for different countries. Interestingly, different countries cluster together, indicating a similar discourse on the topic for different countries.

5 Discussion

Previous work has demonstrated that specific lexical and orthographic variants tend to cluster in distributional space. Studies such as Nguyen and Grieve's (2020) show that subword-based embeddings encode systematic spelling variation, and related work in dialectology and sociolinguistics has used clustering to analyse similarity between known varieties or regional forms. Crucially, much of this research starts from predefined units, such as known variants, dialect labels, or geographic groupings, and then examines how these cluster together. The analytical focus is typically on confirming that related forms or varieties occupy nearby positions in the embedding space.

⁶Verified via the official dictionary for Luxembourgish.

The present study reverses this perspective. Instead of beginning with predefined variants, it detects and mines variants directly from co-occurrence and similarity patterns in the data. Clustering is thus used not as a validation tool, but as an exploratory mechanism that makes candidate forms visible without prior assumptions.

This methodological change is particularly relevant for Luxembourgish. Written Luxembourgish exhibits high orthographic diversity, partial standardisation, frequent borrowing, and code-switching, which make manual enumeration of variants difficult and incomplete. These properties create favourable conditions for unsupervised discovery of variation. This also applies to the dataset at the centre of our analysis. User comments provide dense, repetitive, and informal language use across many contributors, which supports distributional modelling of variation while retaining socially meaningful diversity.

Beyond identifying variant families, the approach opens up several directions for further analysis. One avenue concerns the role of different dimensions in shaping cluster structure. Future work will examine which types of variation cluster along which dimensions, for example whether certain families are associated with particular groups of users or whether historical and more recent forms of Luxembourgish can be distinguished through their distributional profiles. Exploring these questions will help clarify how different sources of variation interact and which analytical perspectives are best supported by this methodology.

The resulting clusters reveal patterns that are well aligned with linguistic intuition, but are not explicitly encoded in existing resources. This suggests that substantial lexical and orthographic variation remains undocumented. Resources such as the one analysed here therefore warrant further investigation, both to enrich descriptive accounts of Luxembourgish and to inform the development of computational models that better reflect actual language use.

Building on these considerations, we emphasise that the full codebase is released publicly to encourage reuse, scrutiny, and extension by other researchers. At the same time, our findings and limited further testing suggest that the proposed approach is not suited to highly standardised languages, where orthographic variation is limited and many relevant distinctions are already captured by existing lexical resources and normali-

sation pipelines. Its strengths instead lie in settings characterised by non-standardisation, activate variation, or incomplete codification, where spelling diversity encodes sociolinguistic and contextual information rather than noise. We therefore hope that this work enables and motivates applications to similar language varieties and research situations, including under-resourced or emerging standards. Beyond language-specific use cases, the method is also applicable to other domains and previously unexplored corpora, such as large web crawls or collections with limited metadata. In this sense, the contribution is less a universal solution for all languages than a transferable framework for studying variation in contexts where standard assumptions do not hold.

6 Conclusion

In this paper, we have made two main contributions. First, we describe a transparent and reproducible methodology for inducing lexical and orthographic variation directly from raw text, without relying on predefined variant lists or normalisation. Second, we present a large-scale empirical analysis of Luxembourgish user-generated text that documents variation as it is used in practice. Across seven analytically defined categories, the method identifies around 800 variant families, revealing systematic patterns of spelling and lexical diversity. The findings confirm earlier observations that related variants cluster in distributional space (Nguyen and Grieve, 2020), while extending this insight by showing how such clusters can be mined directly from data rather than used only for validation.

Looking ahead, we plan to apply this methodology to additional Luxembourgish corpora in order to compare domains and writing contexts. This includes exploring whether user-level variation patterns can be characterised more systematically and assessing how well different available corpora reflect everyday language use. At the same time, initial experiments suggest that the approach is less effective for highly standardised languages, where orthographic variation is limited and distributional signals are weaker. This highlights that the method is particularly suited to languages and domains with active variation, and that its applicability depends on the sociolinguistic properties of the data.

Limitations

The findings in this study need to be interpreted with care. The corpus consists of user comments from a single online platform, which represents only a small portion of the Luxembourgish-speaking population. Patterns observed in this dataset therefore do not necessarily generalise to the wider speech community, nor do they capture the full range of regional, social, or stylistic variation present in Luxembourgish. The method identifies orthographic and lexical families based on distributional and subword similarity, which makes it sensitive to corpus composition and frequency effects. Rare variants may be missed, and high-frequency items can dominate neighbourhood structures. While the induced families provide useful candidates for analysis, their linguistic validity still depends on qualitative assessment. The results should thus be seen as a structured starting point for investigating Luxembourgish variation rather than a comprehensive account of the language.

A further limitation concerns the interpretation of the induced clusters. While the method identifies groups of closely related forms, it cannot by itself determine whether these patterns reflect linguistic variation, author-specific preferences, or temporal effects. In practice, these sources of variation are often intertwined in user-generated text, and distributional similarity alone does not allow them to be disentangled with certainty. Although dimension-based aggregation provides partial insight into how variants are distributed across users or time periods, the clustering process itself is agnostic to the underlying cause of similarity. As a result, the identified families should be interpreted as candidates for linguistic variation that require contextual and qualitative analysis to establish their nature.

Ethical Considerations

This study uses publicly accessible user comments, but they remain sensitive textual data. All processing follows the terms of use of the platform from which the comments were collected. No attempt is made to identify individual users, and the analysis relies only on aggregated patterns such as variant frequencies and distribution across dimensions. Even though user identifiers are present in the raw data, they are treated only as categorical variables and are not (and could not be) interpreted as personal attributes. The dataset represents a self-selected set of online participants whose linguistic

behaviour may differ from that of the wider population, and care should be taken not to attribute group-level characteristics to individual users. Finally, automatically induced variant families can reflect social or regional differences, but these patterns should be interpreted with caution to avoid reifying stereotypes or overgeneralising from limited data. The methodological framework is intended for linguistic analysis rather than profiling or prediction of individuals.

Acknowledgments

This research was supported by the Luxembourg National Research Fund (Project code: C22/SC/117225699). The experiments reported in this paper were conducted on the MeluXina high-performance computing infrastructure, made available through an allocation granted by the University of Luxembourg on the EuroHPC supercomputer hosted by LuxProvide. We would like to thank the TRAVOLTA project partners for their invaluable advice and guidance on this project.

References

- Martine Adda-Decker, Thomas Pellegrini, Eric Bilinski, and Gilles Adda. 2008. [Developments of “Lëtzebuergesch” Resources for Automatic Speech Processing and Linguistic Studies](#). In *Proceedings of LREC*.
- Khetam Al Sharou, Zhenhao Li, and Lucia Specia. 2021. [Towards a Better Understanding of Noise in Natural Language Processing](#). In *Proceedings of RANLP*.
- Dimitra Anastasiou. 2022. [ENRICH4ALL: A First Luxembourgish BERT Model for a Multilingual Chatbot](#). In *Proceedings of SIGUL*. ELRA.
- Alistair Baron and Paul Rayson. 2008. [WARD2: A tool for dealing with spelling variation in historical corpora](#).
- Laura Bernardy. 2022. [A Luxembourgish GPT-2 Approach Based on Transfer Learning](#). Master’s thesis, University of Trier.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. [Enriching word vectors with subword information](#). In *Transactions of ACL*.
- Steven Cao, Nikita Kitaev, and Dan Klein. 2020. [Multilingual alignment of contextual word representations](#). In *Proceedings of ICLR*.
- Jacob Eisenstein. 2013. [What to do about bad language on the internet](#). In *Proceedings of NAACL-HLT*.
- Nathalie Entringer, Peter Gilles, Sara Martin, and Christoph Purschke. 2021. [Schnëssen. surveying](#)

- language dynamics in luxembourgish with a mobile research app. *Linguistics Vanguard*, 7.
- Pieter Fivez, Simon Šuster, and Walter Daelemans. 2017. Unsupervised Context-Sensitive Spelling Correction of English and Dutch Clinical Free-Text with Word and Character N-Gram Embeddings. *Computational Linguistics in the Netherlands Journal*.
- Daniela Gierschek. 2022. *Detection of Sentiment in Luxembourgish User Comments*. Ph.D. thesis, University of Luxembourg.
- Peter Gilles. 1999. *Dialektausgleich im Lëtzebuergesch: Zur phonetisch-phonologischen Fokussierung einer Nationalsprache*. Niemeyer, Tübingen, Germany.
- Peter Gilles. 2019. 39. Komplexe Überdachung II: Luxemburg. Die Genese Einer Neuen Nationalsprache. In Joachim Herrgen and Jürgen Erich Schmidt, editors, *Sprache und Raum - Ein internationales Handbuch der Sprachvariation. Volume 4 Deutsch*. De Gruyter Mouton, Berlin, Boston.
- Peter Gilles. 2021. Variationsatlas vum lëtzebuergesch. <https://infolux.uni.lu/variatiounsatlas>. Accessed: 15.10.2025.
- Peter Gilles, Léopold Edem Ayité Hillah, and Nina Hosseini Kivanani. 2023a. ASRLUX: Automatic Speech Recognition for the Low-Resource Language Luxembourgish. In *Proceedings of the International Congress of Phonetic Sciences*.
- Peter Gilles, Nina Hosseini Kivanani, and Léopold Edem Ayité Hillah. 2023b. LUX-ASR: Building an ASR system for the Luxembourgish language. In *Proceedings of IEEE Spoken Language Technology Workshop*.
- Jack Grieve, Chris Montgomery, Andrea Nini, Akira Murakami, and Diانشeng Guo. 2019. Mapping Lexical Dialect Variation in British English Using Twitter. *Frontiers in AI*, 2.
- Hang Jiang, Haoshen Hong, Yuxing Chen, and Vivek Kulkarni. 2020. DialectGram: Automatic Detection of Dialectal Changes with Multi-geographic Resolution Analysis. In *Proceedings of the Society for Computation in Linguistics*.
- Julie Kallini, Dan Jurafsky, Christopher Potts, and Martijn Bartelds. 2025. False Friends Are Not Foes: Investigating Vocabulary Overlap in Multilingual Language Models. *arXiv preprint*. ArXiv:2509.18750 [cs].
- Thomas Lavergne, Gilles Adda, Martine Adda-Decker, and Lori Lamel. 2014. Automatic language identity tagging on word and sentence-level in multilingual text sources: a case-study on Luxembourgish. In *Proceedings of LREC*.
- Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Proceedings of NeurIPS*.
- Cedric Lothritz, Bertrand Lebichot, Kevin Allix, Lisa Veiber, Tegawende Bissyande, Jacques Klein, Andrey Boytsov, Clément Lefebvre, and Anne Goujon. 2022. LuxemBERT: Simple and Practical Data Augmentation in Language Model Pre-Training for Luxembourgish. In *Proceedings of LREC*.
- Anne-Marie Lutgen, Alistair Plum, Christoph Purschke, and Barbara Plank. 2025. Neural text normalization for Luxembourgish using real-life variation data. In *Proceedings of VarDial (COLING)*.
- Robert Munro and Christopher D. Manning. 2010. Subword Variation in Text Message Classification. In *Proceedings of NAACL-HLT*.
- Dong Nguyen and Jack Grieve. 2020. Do Word Embeddings Capture Spelling Variation? In *Proceedings of ICCL*.
- Nhi Pham, Lachlan Pham, and Adam Meyers. 2024. Towards Better Inclusivity: A Diverse Tweet Corpus of English Varieties. In *Proceedings of LAW (ACL)*.
- Fred Philippy, Shohreh Haddadan, and Siwen Guo. 2024. Forget NLI, Use a Dictionary: Zero-Shot Topic Classification for Low-Resource Languages with Application to Luxembourgish. In *Proceedings of SIGUL (LREC-COLING)*.
- Alistair Plum, Caroline Döhmer, Emilia Milano, Anne-Marie Lutgen, and Christoph Purschke. 2024. LuxBank: The first Universal Dependency treebank for Luxembourgish. In *Proceedings of TLT*.
- Alistair Plum, Tharindu Ranasinghe, and Christoph Purschke. 2025. Text generation models for Luxembourgish with limited data: A balanced multilingual strategy. In *Proceedings of VarDial (COLING)*.
- Christoph Purschke. 2020. Attitudes Toward Multilingualism in Luxembourg. A Comparative Analysis of Online News Comments and Crowdsourced Questionnaire Data. *Frontiers in AI*, 3.
- Tharindu Ranasinghe, Alistair Plum, Christoph Purschke, and Marcos Zampieri. 2023. Publish or Hold? Automatic Comment Moderation in Luxembourgish News Articles. In *Proceedings of RANLP*.
- Joshgun Sirajzade, Daniela Gierschek, and Christoph Schommer. 2020. An Annotation Framework for Luxembourgish Sentiment Analysis. In *Proceedings of SLTU-CCURL (LREC)*.
- Natalie D. Snoeren, Martine Adda-Decker, and Gilles Adda. 2010. The Study of Writing Variants in an Under-resourced Language: Some Evidence from Mobile N-Deletion in Luxembourgish. In *Proceedings of LREC*.
- Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37.
- Zenter fir d’Lëtzebuerg Sprooch, editor. 2019. *D’Lëtzebuerg Orthografie*. Zenter fir d’Lëtzebuerg Sprooch, Stroossen.

Onomasiological Sense Alignment Across Dialect Dictionaries. A Taxonomy-Constrained LLM Classification

Nathalie Mederake, Nico Urbach, Hanna Fischer, Alfred Lameli

Research Center Deutscher Sprachatlas (DSA), Marburg University, Germany

{mederake|urbach|hanna.fischer|lameli}@uni-marburg.de

Abstract

We propose a taxonomy-guided approach to semantic alignment that assigns lexicographic senses to an onomasiological taxonomy derived from the Hallig–Wartburg/Post system. Using an LLM under strict taxonomic constraints, short and heterogeneous meaning descriptions are assigned to a common conceptual space. Evaluation against expert annotation shows that run-to-run model agreement ($\kappa = 0.73$) closely matches human agreement ($\kappa = 0.74$), with robustness at coarse taxonomic levels and predictable degradation at finer granularity. A qualitative network analysis demonstrates the resulting potential for cross-dictionary exploration of dialectal variation in semantics.

1 Introduction

Dialect dictionaries are a central resource for studying language variation and cultural knowledge (Reichmann, 2022), yet they remain notoriously difficult to compare or interlink. The core obstacle is not a lack of data but a lack of semantic interoperability: dialect dictionaries are typically organized semasiologically (i.e., by form), follow heterogeneous editorial conventions despite shared methodological foundations, and encode meaning in non-parallel ways. As a result, identical or closely related concepts are often realized through different expressions and scattered across isolated resources. This fragmentation severely constrains systematic cross-dictionary research on lexical variation.

An onomasiological perspective offers a fundamental alternative. Instead of starting from forms, it starts from meanings and groups words by the concepts they express. In German dialect lexicography, this perspective is well established through semantic taxonomies such as the Hallig-Wartburg system (HW; Hallig and von Wartburg, 1963) and its dialect-lexicographic adaptations. The HW taxonomy provides a conceptual framework that is independent of regional lexical realization and can, in

principle (see Tittel et al., 2020), support concept-based access and comparison across dictionaries. In practice, however, onomasiological access is hard to apply consistently across resources: manual assignment is labor-intensive, semiautomatic methods struggle with sparse glosses, and fully automatic sense alignment techniques typically presuppose structurally homogeneous resources, richer definitions, or parallel sense inventories (Klee, 2024; Bieberstedt et al., 2024).

Large language models (LLMs) provide a qualitatively different option. Their ability to interpret underspecified meaning descriptions, integrate heterogeneous contextual cues, and follow explicit categorical constraints makes them promising candidates for taxonomy-bound semantic classification in dialect lexicography.¹ At the same time, deploying LLMs for this task raises a critical methodological question: can an LLM assign sense-levels to an onomasiological taxonomy reproducibly and lexicographically appropriately across divergent dictionaries, or do stochasticity and ambiguity undermine semantic coherence?

This paper addresses this question by treating taxonomy-bound classification as a practical route to sense-level assignment. We operationalize the HW taxonomy, as adapted for dialect lexicography by Post (1998), and employ it as an explicit conceptual framework. In its extended and partially restructured form (Post, 2026), Post’s system of semantic categories serves as a controlled classificatory backbone, to which individual sense-levels are systematically assigned.

Empirically, we investigate the Hessen-Nassauisches Wörterbuch (HNWB) as a represen-

¹In this paper, **meaning description** denotes meaning-related statements as they occur in dictionary texts. Methodologically, these statements are treated as **sense-levels**, that is, as levels of meaning differentiation within dictionary entries and used as operational units of analysis. **Semantic categories** are conceived as analytical classes used for taxonomic abstraction across individual meaning descriptions.

tative and methodologically challenging dialect dictionary. We implement a reproducible processing pipeline that derives a fine-grained sense-level representation from dictionary encodings and performs LLM-based classification under strict taxonomic constraints. The approach is evaluated through expert-in-the-loop validation and repeated classification runs to quantify both agreement with expert judgments and run-to-run stability.

The paper thus makes three contributions: (i) we propose a taxonomy-guided method for sense-level alignment across dialect dictionaries; (ii) we provide an empirical evaluation of reproducibility via expert annotation and repeated runs; (iii) we illustrate, as a proof of concept, how sense-level assignments derived from the HNWB can be linked to corresponding categories in three other dialect dictionaries from major German dialect regions, laying the foundations for a systematic evaluation of cross-dictionary alignment in future work.

In doing so, we address three interrelated research questions:

RQ1: To what extent can LLM-based classification reliably and reproducibly assign sense-level entries from a single, structurally heterogeneous dialect dictionary to categories of the extended Post taxonomy?

RQ2: To what extent do LLM-generated assignments for the HNWB align with expert judgments under a defined evaluation scheme, and where do systematic deviations occur within the taxonomy?

RQ3: To what extent can a taxonomy-based, sense-level representation of the HNWB be used to link its senses to corresponding categories in other dialect dictionaries?

Beyond dialect lexicography, the proposed approach is, for example, applicable to domain-specific ontologies characterized by sparse and heterogeneous sense descriptions (e.g., historical, terminological, low resources languages).

2 Background

2.1 Semantic Interoperability in Dialect Lexicography

Prior research on sense alignment has proposed a variety of methods operating on both the expression and the content level (Wiegand and Gouws, 2014). Statistical measures used in information retrieval to assess the relevance of terms in documents within a document collection like TF-IDF (Term Frequency-Inverse Document Frequency)

have been applied successfully to longer textual units (Lane and Dyschel, 2025). In lexicography-driven studies, information retrieval is producing particularly robust results for Realia, such as fruit or agricultural products (Burch and Rapp, 2007), but is ill-suited to short or fragmentary definitions. Content-based approaches have explored the comparison of conceptual information extracted from definitions, for example, through the use of hyperlemma lists (Fournier, 2003, 160–167). Sense-alignment work between GermaNet and DWDS has shown that parallel senses provide external evidence for the validity of sense distinctions and allow mutual enrichment of resources, whereas non-parallel cases point to entries in need of reconsideration (Henrich et al., 2014). Distributional methods relying on semantic proximity in vector spaces (Mikolov et al., 2013) offer another perspective, but they presuppose relatively balanced definitions and encounter difficulties when applied to dialectal or historical vocabulary.

More recently, cross-resource sense alignment has been framed as a supervised classification task over sense pairs, achieving improved performance through the combination of handcrafted features and representation learning (Ahmadi and McCrae, 2021). In a comparative case study of monolingual sense alignment, Salgado et al. (2020) highlight substantial challenges arising from inconsistent lexicographic criteria, heterogeneous sense inventories, and divergent wording practices in glosses. While their work demonstrates the practical value of sense alignment for semantic web publication and NLP applications, it also underscores that alignment quality is fundamentally constrained by the degree of semantic explicitness and structural comparability provided by the source dictionaries.

When applied to dialect lexicography, however, the assumptions underlying these approaches quickly reach their limits. Because many of these dictionaries are historical resources with inconsistent or minimal markup, the semantic cues they contain are not machine-readable. Even where technical aggregation infrastructures exist, such as in dictionary networks (e.g., www.woerterbuchnetz.de), the underlying semantic information remains incompatible (Klee, 2024). Interoperability at the level of individual lexicographic meaning descriptions, which is necessary for computational reuse in studies on language variation and change (Garido García, 2021), lexicographic integration, or NLP applications, has therefore not kept pace with

the availability of digitized materials. This gap indicates that sense-level interoperability requires methods that can operate on sparse, heterogeneous definitions while remaining constrained by an explicit conceptual framework.

2.2 Semantic Taxonomy

Semantic taxonomies offer a potential remedy. Although originally developed within Romance studies, HW is a language-independent, onomasiological system designed to structure lexical meaning independently of lexical form. It functions as a controlled, hierarchical ontology and has been widely applied in historical lexicography and lexicological studies of Middle High German as well as medieval and early modern (up to the 16th century) French, Italian, Spanish, Gascon, and Occitan (cf. Tittel et al., 2020). It organizes concepts within a small number of highly abstract domains and a deeply articulated hierarchy whose ordering principle is intrinsic to the system (see Tittel et al., 2020 for a computational account). Concepts are positioned through their relations within this architecture rather than through assignment to mutually exclusive semantic fields.

Take the conceptual field *nature* as an example. In HW, the relevant concepts are embedded, for example, within the domain *A. The universe* and further structured along the path *I. The sky and the atmosphere* → *a) The sky and celestial bodies*, where elements such as *Nature*, *Wind*, and related phenomena appear as coordinated concepts within a system-internal hierarchy (original in French). Post’s (2026) adaptation reorganizes this material for dialect lexicography. While largely preserving the conceptual inventory, Post restructures HW’s hierarchical pathways into explicitly named and numerically indexed domains, for example, by differentiating *nature* into categories such as *Inanimate Nature* (Figure 1).

This restructuring transforms the HW system into an taxonomy designed for consistent lexicographic assignment. Post’s (1998) modification is widely used in dialect dictionaries (e.g., Bickel, 2013; Breuer and Stöckle, 2023; Schwarz, 2022).²

By contrast, WordNet-style taxonomies are primarily synset-based and lexeme-driven, having been developed for contemporary English. Although they exhibit a hierarchical structure, their

²The version used in this paper follows the augmented taxonomy employed in the Fränkisches Wörterbuch (<https://wbfbadw.de/wbfbadw-digital/zur-dokumentation.html>).

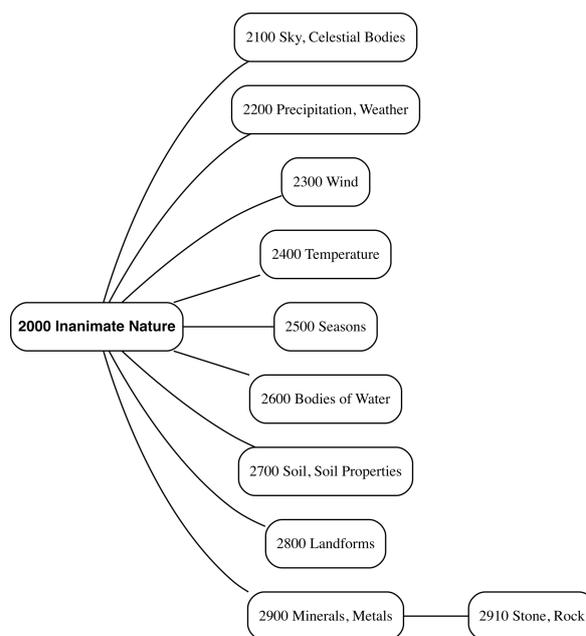


Figure 1: Extract from Post’s (2026) modification of the HW taxonomy

organization is induced from the lexical inventory of a specific synchronic language stage. As a result, they are less suitable for lexical-semantic mapping in historical lexica or terminological resources, where conceptual structures may not align neatly with modern lexical units or where semantic relations diverge diachronically (see Trotter et al., 2016). The practical challenges of applying WordNet to historical lexical resources have received only limited attention in the literature (cf. Tittel, 2025). This structural limitation points to a fundamental difference: as a concept-driven and language-independent ontology, HW does not presuppose a particular synchronic lexicalization and therefore provides a more stable framework for diachronic and cross-linguistic mapping.

However, its computational application has remained experimental, for example, asking to what extent LLMs could effectively support the editorial process in a dictionary project (Raaf and Röhler, 25.09.2025). Whether such a taxonomy can support automated large-scale alignment across dictionaries remains an open question.

3 Material and Methods

3.1 Dialect Data

This study takes the Hessen–Nassauisches Wörterbuch (HNWb) as its primary empirical object. While large-scale German dialect dictionaries are broadly comparable in their micro- and macrostruc-

tural design (Lameli, 2021), they differ substantially from one another in editorial practice, temporal depth, and lexical coverage (Moulin, 2010; Lenz and Stöckle, 2020), limiting systematic cross-dictionary comparison. Focusing on the HNWB as a single case allows us to investigate taxonomy-based sense-level assignment under realistic lexicographic conditions. However, in order to illustrate the extensibility of the resulting sense-level representation, three additional dialect dictionaries are considered in a secondary role: the Mecklenburgisches Wörterbuch (MWb), the Pfälzisches Wörterbuch (PfWb), and the Schweizerisches Idiotikon (SchwId). Together, these resources cover major dialect regions from Low German (MWb) over Central German (HNWB, PfWb) to Upper German (SchwId) (Table 2).

The empirical analysis is restricted to the alphabetic range N , which is uniformly accessible across all four resources. While this restriction is motivated by data availability, it also offers methodological advantages: N spans a wide range of semantic domains (e.g., natural phenomena, temporal expressions, abstract operators), comprises both high-frequency function words and low-frequency content words, and exhibits substantial editorial heterogeneity across dictionaries. Furthermore, unlike other initial consonants, it does not exhibit onset variation (e.g., <f> vs. <v>, <t> vs <d>) and therefore provides a stable baseline for comparison. These properties make it a suitable test case for taxonomy-bound sense-level assignment.

Within the N range, the dictionaries differ considerably in size and microstructural granularity. The HNWB comprises 981 articles with 1,163 sense-levels. For the three additional dictionaries, the corresponding numbers are: MWb: 1492/1735; PfWb: 1502/1887; SchwId 3591/2982. Dictionary entries or single sense-levels lacking usable glosses—predominantly in SchwId—were excluded from the assignment.

3.2 Workflow

The methodological framework follows an onomasiological approach that aligns dictionary entries via the concepts they express. The HW taxonomy as extended by Post (2026) serves as a controlled conceptual foundation. Each sense-level is mapped to a four-digit taxonomy code, positioning it within a hierarchical domain and thereby enabling nested evaluation across different levels of conceptual granularity (e.g., code 2000 “Inanimate Nature”,

extended by 2200 “Precipitation and Weather”; see Figure 1). The definition of sense-levels is based on the Standard German definitions provided in the dictionaries, rather than on dialectal keywords. Examples of dialectal usage, which typically take the form of multiword expressions, were explicitly excluded from the analysis because they often refer to multiple lexical items.

The LLM classification follows a detailed workflow (Figure 2) outlined in the following sections.

Preprocessing The HNWB entries were brought to a uniform level of annotation with regard to meaning specification using lightweight markers inspired by TUSTEP conventions (e.g., `#marker + ... #marker-`). The resulting information was stored in a JSON schema that explicitly separates (a) lexical form, (b) grammatical metadata, (c) sense-levels, and (d) auxiliary gloss material (e.g., Latin equivalents). This schema accommodates heterogeneous metadata while enabling systematic extraction of sense-levels, batch processing, and reproducible interaction with the LLM-based classification workflow. Structural inconsistencies and reference entries were resolved manually according to pre-defined annotation rules. The same preprocessing framework was subsequently applied to the additional dialect dictionaries used for illustrative cross-dictionary linking, ensuring structural compatibility at the sense-level.

Prompt Engineering Semantic classification was performed using the OpenAI model GPT-5, accessed via the KISSKI interface.³ We employed an in-context learning prompt (Figure 10) comprising five carefully selected few-shot examples. These examples were designed to enforce strict taxonomic constraints while minimizing over-interpretation of dialectal form variation (e.g., phonological and morphological variation). Dictionary entries were processed in batches of approximately 8,000–9,000 tokens. Using this procedure, all entries in the alphabetic range N of the HNWB were distributed across 14 subsets, each containing all information required for classification.

³AI Service Centre for Sensitive and Critical Infrastructure; see <https://kisski.gwdg.de> for details. The portal provides a simplified API interface. Any associated costs were billed directly through the University of Marburg, which was essential given the absence of dedicated project funding. Model outputs were generated using default sampling parameters (temperature = 0.5, top-p = 0.5, no frequency or presence penalty). This configuration corresponds to the KISSKI default and is reported to OpenAI.

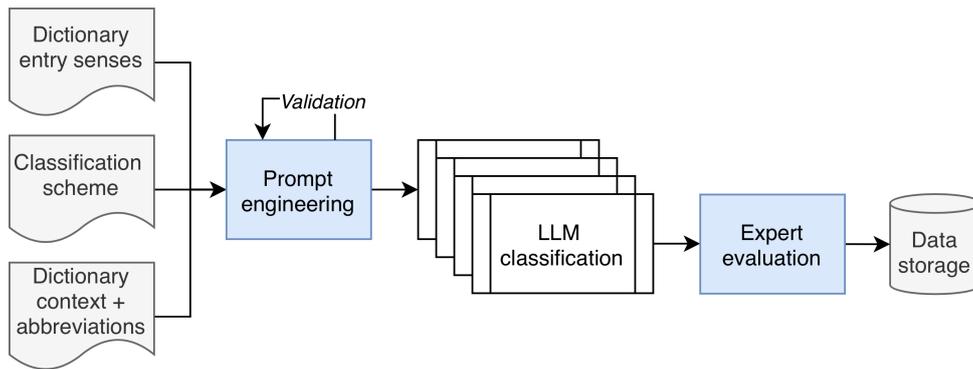


Figure 2: Workflow for LLM sense-level alignment.

Entries that did not contain semantic information and merely referred to other entries—either within the same batch or outside the N range—were labeled as “out of index” and excluded from further evaluation. Since such entries point to other lemmas via explicit cross-references, assigning their meanings constitutes a straightforward rule-based task. In addition, almost all categorizations of these reference entries produced by the LLM were correct.

To ensure consistent and context-aware classification, each batch—comprising the lemma entry, variants, grammatical information, and sense-levels—was paired with a comprehensive context file. This file included abbreviation lists, editorial metadata, dialect background information, and the full extended Post taxonomy (approximately 10,700 tokens⁴).

Classification The automated semantic classification was conducted in three stages.

Stage 1 (primary assignment): The LLM-based classifier was first applied to the HNWB in order to generate taxonomy-based sense-level assignments under controlled prompt settings (1st run). The classification was carried out in multiple iterative calls, with a practical upper limit of approx. 20,000 tokens per call.

Stage 2 (reproducibility assessment). To assess run-to-run reliability, the HNWB was classified a second time under identical prompt settings and experimental conditions (2nd run). This 2nd run served to quantify the stability and reproducibility of the automated sense-level assignment. The results of the two runs were evaluated on the basis of expert-validated sense-level assignments.

Stage 3 (illustrative extension). In a final step (3rd run), the same classification procedure was ap-

⁴Tokenization as determined by the tiktoken 0.12.0 Python library.

plied to two additional dialect dictionaries (MWb, SchwId). These assignments were not subjected to the same reproducibility or expert-based evaluation but served to demonstrate how sense-level representations derived for the HNWB can be linked to, a broader dialect-lexicographic context. Additionally, we used the available sense-level assignments from the PfWb (see Post, 1998).

Expert Evaluation To assess the quality and reliability of the LLM-based semantic classification, the model-generated sense-level assignments of the HNWB were evaluated by two expert lexicographers at two distinct points in the workflow.

Evaluation 1 (quality check for prompt design): A preliminary quality check was conducted on a random sample of approx. 100–120 sense-level assignments per dictionary (Validation during prompt engineering in Figure 2). The purpose of this check was to verify whether the prompt design and model behavior met the requirements for large-scale application. Each assignment was manually inspected and judged dichotomously (correct vs. incorrect) by the expert lexicographers working independently. An acceptance threshold of 80% correct assignments was defined ex ante as the minimum level required to justify full deployment. This value was chosen conservatively based on a pilot evaluation of the random sample and serves as an operational decision point within the pretesting phase. If the threshold was not met, the prompt was revised and the evaluation repeated.

Evaluation 2 (quality check after classification): A second expert evaluation was carried out after completion of the 1st run, focusing exclusively on the 997 assigned sense-level entries of range N in the HNWB. All assignments were independently corrected by the same two lexicographers. This led to two individual classifications which were

statistically evaluated in terms of reliability. To this end, a three-level assessment scheme was applied: (1) *Correct*, indicating an exact match between the model-generated and expert-assigned category; (2) *Almost correct*, indicating assignment to a semantically adjacent node within the correct higher-level category; and (3) *False*, covering all remaining cases. The individual classifications were then compared and any disagreements were resolved to arrive at a consensual final assignment. The finalized expert assignments derived from this evaluation subsequently served as the baseline for the automated evaluation of the 2nd run.

4 Results

4.1 Sense-Level Assignment

Human Baseline To establish a reference point for the level of agreement achievable under Post’s semantic taxonomy, inter-rater agreement between the two expert lexicographers was calculated for all entries in the HNWB *N* range. Agreement was quantified using unweighted Cohen’s κ , yielding a value of $\kappa = 0.743$. This level reflects the intrinsic complexity of the classification task. Post’s taxonomy is hierarchically structured, and category boundaries are not always sharply defined. Consequently, variability in expert judgments should be interpreted as an effect of semantic granularity and ambiguity rather than as annotator error.

Run-to-run Reliability To assess the internal consistency of the LLM-based classification, the two independent classification runs on the HNWB were compared using identical prompts and experimental conditions. The expert-validated sense-level assignments served as the reference standard. Agreement between the 1st and 2nd run was substantial ($\kappa = 0.733$), closely approximating the human baseline. In absolute terms, the proportion of strictly correct assignments decreased from 86.9% (875 cases) in the 1st run to 81.0% (812 cases) in the 2nd run. This difference is statistically significant ($\chi^2(1) = 14.42, p < .001$), although the associated effect size was small ($\phi = .08$).

Stability Beyond aggregate agreement, the item-level overlap between the two runs was examined (Table 1). Of the assignments classified as strictly correct in the 1st run, 89.40% (776 cases) retained this status in the 2nd run. By contrast, assignments classified as almost correct showed lower retention (60.34%, 35 cases). Incorrect classifications were

retained in 74.64% (53 cases). Across all items, roughly 25% of sense-level assignments changed their evaluation level (correct / almost correct / false) between runs (Figure 3).

Evaluation	1st run	2nd run	Overlap
Correct	868	811	776 (89.40%)
Almost correct	58	68	35 (60.34%)
False	71	111	53 (74.64%)

Table 1: Overlap of sense-level alignments between 1st and 2nd run.

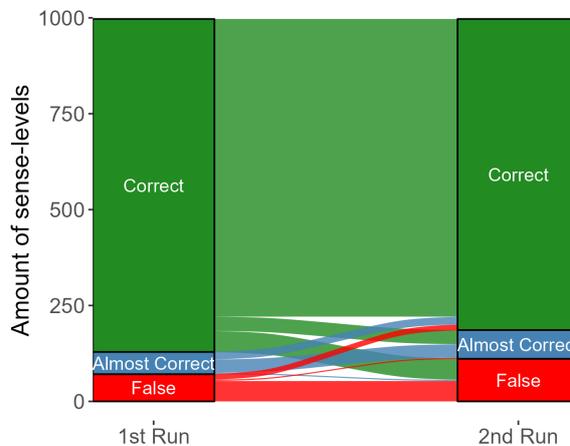


Figure 3: Item overlap of sense-level alignments between 1st and 2nd run.

Taxonomic Granularity Classification accuracy varies systematically with taxonomic depth. When evaluated at higher, more abstract levels of Post’s taxonomy, assignments show higher accuracy and greater stability. Accuracy decreases as categories become more fine-grained (Figure 4), with increased variability across repeated runs. Error bars indicate 95% Wilson confidence intervals for binomial proportions.

4.2 Semantic Inspection

Taxonomy-Sensitive Stability Beyond aggregate agreement scores, run-to-run stability varies systematically across semantic regions of the taxonomy. As shown in Figure 6 and Figure 7, sense-level entries of the HNWB differ markedly in their propensity to change categories between the two runs. Highly referential domains (e.g., 3000 “Plants and fruits”) show consistently low

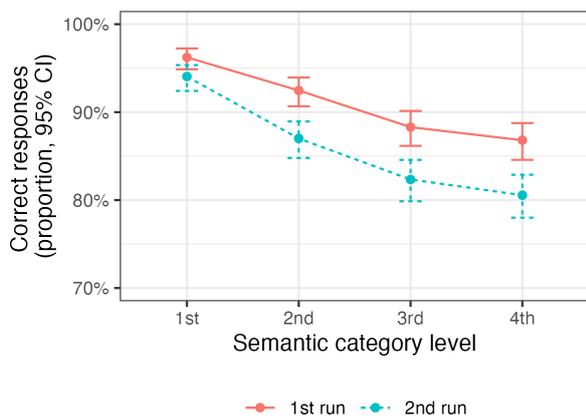


Figure 4: Classification accuracy across hierarchical semantic levels; error bars indicate 95% Wilson confidence intervals for binomial proportions.

change rates and high proportions of strictly correct assignments (see Figure 8), whereas, for example, abstract or evaluative categories exhibit substantially higher variability. There are also collective categories that have no semantic relevance and therefore lead to misclassifications, such as the high-level category 1000 “Special Words” that, among others, contain temporal and spatial adverbs which—semantically—could be attributed to 9500 “Space” or 9600 “Time”.

A further, qualitatively distinct source of misclassification arises in normatively charged areas of the taxonomy. Meanings referring to human behavior or personal traits are repeatedly assigned to evaluative categories such as 6116 “improper behaviour” or 6510 “moral judgment (adjectival)” (Figure 8), even when no moral judgment is encoded in the meaning description. Importantly, these missassignments are systematic rather than random and suggest that the model tends to activate evaluative supercategories when processing human-related semantics. The phenomenon points to an interaction between the taxonomy’s normative structuring and the model’s tendency to infer implicit evaluation, rather than to isolated classification errors.

Semantic Network We further illustrate how taxonomy-based, sense-level assignment derived from the HNWB can restructure access to dialect dictionary data beyond traditional alphabetical lookup. While the distributional analysis establishes robustness at the level of aggregated categories, it does not show how individual sense-level entries relate to one another within the shared conceptual space. To make this relational structure visible, we include sense-level assignments from

three additional dialect dictionaries obtained in the 3rd run and project all classified senses into a common taxonomic space for exploratory purposes. This space is represented as a semantic network of cross-dictionary relations. For illustration, we focus on sense-level entries across all subcategories under node 2000 “Inanimate Nature” (Figure 9), a region of the taxonomy that shows comparatively low volatility across runs (Figure 6). Within this network, the categories of Post’s taxonomy delineate conceptual regions, within which individual sense-level entries are positioned as nodes.

Mixed colors within a node indicate categories with strong cross-dictionary representation. Take category 2100 “Sky, Celestial Bodies” as an example, where the concept ‘new moon’ forms a tightly connected semantic neighborhood (Figure 5). Across dictionaries, formally divergent but etymologically related designations converge within this category. Entries such as *Neumond* (hnwb_N00568), *Neu-mond* (pfbw_PN01033), *Niman* (mwb_N01074), and *Nü(w)män* (schwi_159172) share a common etymon and are thus connected as cognate realizations of the same basic concept. In addition, synonymous expressions based on alternative lexical motivations, such as *Neulicht* (hnwb_N00565), *Nü(w)* (schwi_159114), and *Neu-schein* (pfbw_PN01081), are integrated into the same semantic region.

Beyond naming variants, the network also captures systematically related concepts that extend the semantic field of *new moon*. These include oppositional or contrastive relations, as in *Nacht-helling* ‘moonlight’ (pfbw_PN00239), as well as derivational extensions such as the adjective *nü(w)mänlich* ‘taking place at the time of the new moon’ (schwi_159173). By bringing together cognate forms, synonyms, and semantically related derivations from multiple dictionaries within a single taxonomic category, the network reveals conceptual relationships that would be difficult to detect systematically through alphabetical access.

The resulting network shows that sense-level entries from the HNWB consistently co-occur with semantically corresponding entries from the additional dictionaries within the same taxonomic regions. Formally divergent lexicalizations, including cognate forms, synonymous expressions, and derivationally related senses, cluster in coherent conceptual neighborhoods, indicating that the taxonomy-based sense-level assignments support meaningful cross-dictionary linking.

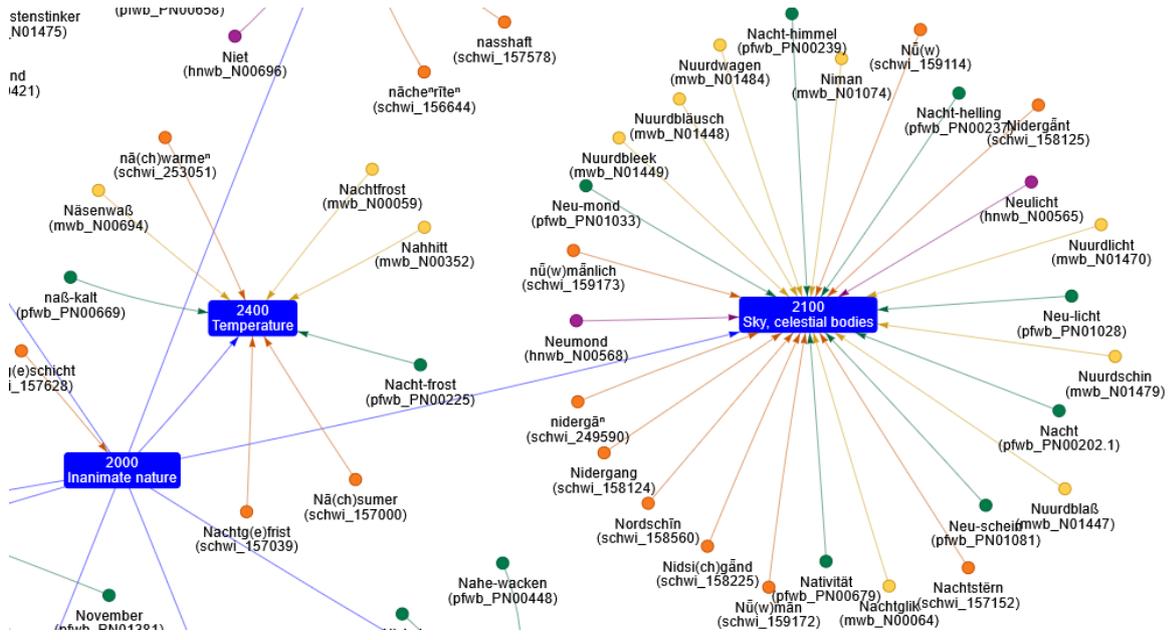


Figure 5: Semantic network of sense-levels for node 2000 “Inanimate Nature” of the taxonomy following Post (2026) (extraction of Figure 9); colors according to dictionaries, labels with lemma and ID of sense-levels.

5 Discussion

Reliability Relative to Human Agreement Addressing RQ1 and RQ2, the results show that taxonomy-bound LLM classification achieves a level of reliability comparable to expert annotation. Human inter-rater agreement ($\kappa = 0.743$) and run-to-run model agreement ($\kappa = 0.733$) fall into a similar range, indicating that a substantial share of classification variability reflects genuine semantic ambiguity rather than model instability (see Artstein and Poesio, 2008; Passonneau and Carpenter, 2014). Given the hierarchical structure of Post’s taxonomy and the often underspecified or context-dependent nature of meaning descriptions in dialect dictionaries, perfect agreement cannot be expected. Convergence between human and model agreement therefore suggests that the LLM operates within the bounds of lexicographic interpretative practice, rather than introducing qualitatively different error patterns. From this perspective, volatility becomes an informative signal: it highlights regions where the taxonomy enforces discrete choices on meanings that are inherently perspectival or multi-dimensional, and thus marks the practical limits of fine-grained, deterministic classification.

Sources of Instability and Semantic Ambiguity

Item-level analysis reveals that instability is concentrated in borderline cases. Assignments classified as almost correct show lower retention across runs

than strictly correct or incorrect assignments, indicating competition between semantically adjacent taxonomic categories. Incorrect classifications, by contrast, tend to remain stable once a decision falls outside the correct taxonomic region. This asymmetry suggests that variability is driven by semantic underspecification rather than random noise.

Overall, approx. 25% of sense-level assignments change their evaluation-status across runs. Rather than undermining reliability, this volatility provides an empirical estimate of the proportion of cases for which fine-grained taxonomic assignment is inherently uncertain. This is not solely due to the classification, but also to the taxonomy, which in some cases contains vague categories or collective categories (e.g., 9214 “other”). At the same time, the comparison of the 1st and the 2nd run reveals a theoretical upper bound for semantic assignment accuracy. Although the classifications of the runs overlap substantially, they differ in 35 strictly correct assignments (see Table 1). Aggregating all correct assignments in either run would raise accuracy to 90.57%. This constitutes a best-case estimate conditioned on the two runs. Given the probabilistic nature of the classification and the observed volatility, deterministic aggregation is not feasible in practice. However, the complementarity across runs suggests that controlled resampling may increase coverage, an issue left for future work.

Furthermore, the strong dependence of accuracy

on taxonomic depth highlights a central trade-off: coarse-grained categories (1st or 2nd level in the taxonomy) are assigned robustly and reproducibly, whereas fine-grained distinctions (3rd or 4th level) show increased variability. This pattern mirrors human annotation behavior under hierarchical taxonomies and indicates that taxonomy-based alignment is suited for applications that operate at intermediate or higher levels of abstraction, while finer distinctions should be interpreted probabilistically.

Qualitative Sense-Level Assignment and Cross-Dictionary Linking Addressing RQ3, qualitative inspection of taxonomy-based semantic networks demonstrates how a sense-level representation derived from the HNWB can be extended beyond alphabetical access and linked to other dialect dictionaries (see 190 Klosa and Müller-Spitzer, 2016; Moulin, 2010). Projecting sense-level entries from multiple resources into a shared taxonomic space, reveals coherent cross-dictionary clustering within individual categories. Formally divergent lexicalizations (Niebaum, 1986)—including cognate forms, synonymous expressions, and derivational extensions—consistently co-occur within the same conceptual regions, indicating successful taxonomy-based linking at the sense-level.

Crucially, the graph-based representation renders explicit conceptual relations that would not be readily observable through linear, alphabetically organized dictionary access. For example, lexical items such as *Neulicht* ‘new moon’ (hnwb_N00565; cf. Figure 5), corresponding to Standard German *Neumond*—with *Neulicht* literally meaning ‘new light’—and *Nachthimmel* ‘night sky’ (pfbw_PN00239) would not ordinarily be associated on the basis of alphabetical ordering or surface-level semantic similarity. Within the taxonomic network, however, their relation becomes apparent through shared higher-level conceptual nodes, thereby exposing latent semantic connections and yielding interpretive added value.

Beyond facilitating cross-dictionary exploration, this perspective also highlights a structural property of parts of dialect lexicography: some dialect dictionaries document the differential lexicon relative to the standard language more systematically than basic conceptual inventories. Apparent gaps in basic concepts therefore often reflect editorial practice rather than conceptual or lexical absence. Taxonomy-based alignment makes such asymmetries explicit and enables comparative anal-

ysis of how semantic domains are lexically populated across regions. At the same time, recurrent misclassifications and systematic category overextensions observed in the LLM-based assignments provide empirical feedback for refining the taxonomy itself, identifying nodes that are overly abstract, normatively overloaded, or insufficiently discriminative. In this way, taxonomy-guided sense-level assignment not only supports linking but also serves as a diagnostic lens on both lexicographic practice and the taxonomy.

Scope and Limits of Taxonomy-Based Alignment Taxonomy-guided sense-level assignment does not, by itself, resolve the full problem of semantic mapping. Assigning multiple sense-levels to the same Post category (e.g., 2100 “Sky, Celestial Bodies”) situates them within a shared conceptual space, but it does not entail semantic equivalence at finer levels of differentiation. Within this framework, the Post taxonomy functions as a controlled conceptual backbone that constrains the semantic search space and enables first-order alignment across heterogeneous sense descriptions. Further differentiation within a category would require additional analytical layers, such as relation typing or sub-clustering at the node level. Taxonomy-based classification should therefore be understood as a foundational step toward sense-level interoperability: it provides a stable, hierarchically organized structure that conditions—but does not substitute for—more fine-grained semantic analysis. At the same time, grouping semantically related concepts supports resource-efficient workflows, particularly important for LLM-assisted approaches to identifying semantic relations between dictionary entries.

6 Conclusion

This paper introduced a reproducible workflow for taxonomy-based, sense-level assignment using the HNWB as a case study. Combining a semantic taxonomy with LLM-supported classification and expert evaluation, we show that meaning descriptions are reliably assigned to common semantic categories, with robust reproducibility across sense-levels. The resulting representation supports structured semantic access and enables linking to other dialect dictionaries, establishing taxonomy-guided LLM classification as a foundation for sense-level interoperability in dialect lexicography and related low-resource NLP settings.

Limitations

Several limitations of the present study should be acknowledged. First, the empirical evaluation is restricted to the alphabetic range *N*, which is the only segment uniformly available across all four dictionaries. Although this range spans multiple semantic domains and provides a meaningful stress test for interoperability, it does not support claims regarding full-dictionary coverage or domain-complete semantic analysis. Extending the approach to additional alphabetic ranges remains a task for future work.

Second, the proposed approach depends on the availability and quality of explicit meaning descriptions. Dictionary entries and single sense-levels lacking usable glosses were excluded from classification, and deviations from Standard German in the metalanguage (e.g., mixed dialect–standard formulations or Latin quotations) required additional preprocessing. Dictionaries characterized by minimal, implicit, or highly heterogeneous meaning descriptions may therefore necessitate further normalization efforts.

Third, classification quality is influenced by prompt configuration, contextual information, and the specific LLM version and interface employed. Although the workflow is reproducible, alternative model settings or future model updates may require re-tuning of prompts and contextual parameters. Moreover, while expert-in-the-loop validation ensures high-quality assignments, it introduces manual effort that may limit scalability in fully automated scenarios.

Finally, the present study focuses on establishing methodological feasibility and semantic plausibility rather than on evaluating downstream applications. Tasks such as systematic variation of batch size and order, controlled resampling of sense-level alignments, and large-scale clustering or similarity measurement constitute important directions for future research. While the study covers a wide range of structurally diverse dialects, its findings are necessarily limited to variation within German. Nonetheless, the proposed methodology is not language-specific and can, in principle, be extended to other languages and dialect continua.

Acknowledgments

This work is funded by the Academy of Sciences and Literature Mainz (Grant REDE 0404). We would like to thank three anonymous reviewers

for their valuable comments and discussion. We would also like to thank Rudolf Post for sharing the sense-level assignments of the PfWb.

References

- Sina Ahmadi and John P. McCrae. 2021. [Monolingual word sense alignment as a classification problem](#). In *Proceedings of the 11th Global Wordnet Conference*, pages 73–80, University of South Africa (UNISA). Global Wordnet Association.
- Ron Artstein and Massimo Poesio. 2008. [Survey article: Inter-coder agreement for computational linguistics](#). *Computational Linguistics*, 34(4):555–596.
- Hans Bickel. 2013. Fortschreitende Digitalisierung. Neue Zugriffe auf das Idiotikon. In *150 Jahre Schweizerisches Idiotikon: Beiträge zum Jubiläumskolloquium in Bern, 15. Juni 2012*, pages 121–134, Bern. Schweizerische Akademie der Geistes- und Sozialwissenschaften.
- Andreas Bieberstedt, Nico Förster, Petra Himstedt-Vaid, and Christoph Schmitt. 2024. Vom gedruckten Wörterbuch zur virtuellen Forschungsumgebung: Digitale Vernetzungsszenarien dialektaler Großwörterbücher am Beispiel des Mecklenburgischen Wörterbuchs. In Antje Dammel and Markus Denkler, editors, *Grosslandschaftliche Dialektwörterbücher zwischen Linguistik und Landeskunde*, pages 17–55. Böhlau Verlag, Wien, Köln.
- Ludwig M. Breuer and Philipp Stöckle. 2023. [Informationssystem Österreich - Zugriff und Vernetzungsmöglichkeiten: Version 2 \(22.06.2023, 16:23\)](#).
- Thomas Burch and Andrea Rapp. 2007. Das Wörterbuchnetz: Verfahren – Methoden – Perspektiven. In Daniel Burckhardt, Rüdiger Hohls, and Claudia Prinz, editors, *Geschichte im Netz: Praxis Chancen Visionen. Beiträge der Tagung .hist 2006*, pages 607–627. Humboldt Universität zu Berlin, Berlin.
- Johannes Fournier. 2003. Vorüberlegungen zum Aufbau eines Verbundes von Dialektwörterbüchern. *Zeitschrift für Dialektologie und Linguistik*, (70):155–176.
- C. Garrido García. 2021. [A structural analysis of dictionaries as semantic networks](#). Ph.D. thesis, Universidad de Chile.
- Rudolf Hallig and Walther von Wartburg. 1963. *Be-griffssystem als Grundlage für die Lexikographie: Versuch eines Ordnungsschemas*. Veröffentlichungen des Instituts für Romanische Sprachwissenschaft. Akademie-Verlag, Berlin.
- Verena Henrich, Erhard W. Hinrichs, and Reinhild Barkey. 2014. [Aligning word senses in germanet and the dwds dictionary of the german language](#). In *Global WordNet Conference*.

- Anne Klee. 2024. Vernetzungsstrategien zwischen Dialektwörterbüchern - am Beispiel des Trierer Wörterbuchnetzes. In Antje Dammel and Markus Denkler, editors, *Grosslandschaftliche Dialektwörterbücher zwischen Linguistik und Landeskunde*, pages 113–131. Böhlau Verlag, Wien, Köln.
- Annette Klosa and Carolin Müller-Spitzer, editors. 2016. *Internetlexikografie*. De Gruyter, Berlin, Boston.
- Alfred Lameli. 2021. Dialektwörterbücher zwischen Web 0.0 und Web 3.0. In *Das Sudetendeutsche Wörterbuch: Bilanzen und Perspektiven*, pages 45–70. Frank & Timme.
- Hobson Lane and Maria Dyshel. 2025. *Natural language processing in action*, second edition edition. Manning Publications, Shelter Island, NY.
- Alexandra N. Lenz and Philipp Stöckle. 2020. *Germanistische Dialektlexikographie zu Beginn des 21. Jahrhunderts*. Steiner, Stuttgart.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Preprint*, arXiv:1301.3781.
- Claudine Moulin. 2010. 33. Dialect dictionaries - traditional and modern. In Peter Auer and Jürgen Erich Schmidt, editors, *Language and Space. An International Handbook of Linguistic Variation. Volume 1 Theories and Methods*, pages 592–612. Walter de Gruyter, Berlin, New York.
- Hermann Niebaum. 1986. Lemma und Interpretament: Zur Problematik der Artikelgestaltung in Dialektwörterbüchern. In Hans Friebertshäuser and Heinrich J. Dingeldein, editors, *Lexikographie der Dialekte*, pages 125–144. De Gruyter, Berlin.
- Rebecca J. Passonneau and Bob Carpenter. 2014. The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics*, 2:311–326.
- Rudolf Post. 1998. Möglichkeiten der elektronischen Strukturierung, Vernetzung und Verfügbarmachung von lexikographischen Daten bei der Arbeit am Pfälzischen Wörterbuch. In Rudolf Grosse, editor, *Bedeutungserfassung und Bedeutungsbeschreibung in historischen und dialektologischen Wörterbüchern: Beiträge zu einer Arbeitstagung der deutschsprachigen Wörterbücher, Projekte an Akademien und Universitäten vom 7. bis 9. März 1996 anlässlich des 150jährigen Jubiläums der Sächsischen Akademie der Wissenschaften zu Leipzig*, pages 211–220. Hirzel, Stuttgart.
- Rudolf Post. 2026. Postsche Sachkategorien - CSV. *LinguRep*.
- Manuel Raaf and Ines Röhrer. 25.09.2025. Semantische Klassifikation lexikographischer Inhalte mithilfe künstlicher Intelligenz neu denken? Ergebnisse einer Studie zur Erweiterung des Bayerischen Wörterbuchs um Ontologie mithilfe von LLMs. FORGE 2025 - Daten neu denken (FORGE), Rostock.
- Oskar Reichmann. 2022. Dimensionen der Wortbedeutung und historische Lexikographie. In Gerhard Diehl and Volker Harm, editors, *Historische Lexikographie des Deutschen. Perspektiven eines Forschungsfeldes im digitalen Zeitalter*, pages 229–254. De Gruyter, Berlin, Boston.
- Ana Salgado, Sina Ahmadi, Alberto Simões, John P. McCrae, and Rute Costa. 2020. Challenges of word sense alignment: Portuguese language resources. In *Workshop on Linked Data in Linguistics*.
- Brigitte Schwarz. 2022. *Das dialektologische Informationssystem von Bayerisch-Schwaben: Dokumentation und mögliche Präsentation von Sprachdaten mit Multimedia im Internet*. Ph.D. thesis, Universität Augsburg, Stuttgart.
- Sabine Tittel. 2025. Historisierte Ontologien für Linguistic Linked Open Data-Ressourcen des Mittelalters. In *Das Mittelalter. Perspektiven mediävistischer Forschung*. Heidelberg University Publishing.
- Sabine Tittel, Frances Gillis-Webber, and Alessandro A. Nannini. 2020. Towards an ontology based on Hallig-Wartburg's Begriffssystem for historical linguistic linked data. In *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)*, pages 1–10, Marseille, France. European Language Resources Association.
- David Trotter, Andrea Bozzi, and Cédric Fairon. 2016. Un'ontologia per il ditmao (dictionnaire des termes médico-botaniques de l'ancien occitan). In *Actes du XXVIIe Congrès international de linguistique et de philologie romanes (Nancy, 15-20 juillet 2013). Section 16: Projets en cours; ressources et outils nouveaux*. ATILF, Nancy.
- Herbert Ernst Wiegand and Rufus H. Gouws. 2014. Access structures in printed dictionaries. In Rufus H. Gouws, Ulrich Heid, Wolfgang Schweickard, and Herbert Ernst Wiegand, editors, *Dictionaries: An International Encyclopedia of Lexicography*, volume 5.4 of *Handbooks of Linguistics and Communication Science*, pages 110–148. De Gruyter Mouton, Wien, Köln.

A Appendix

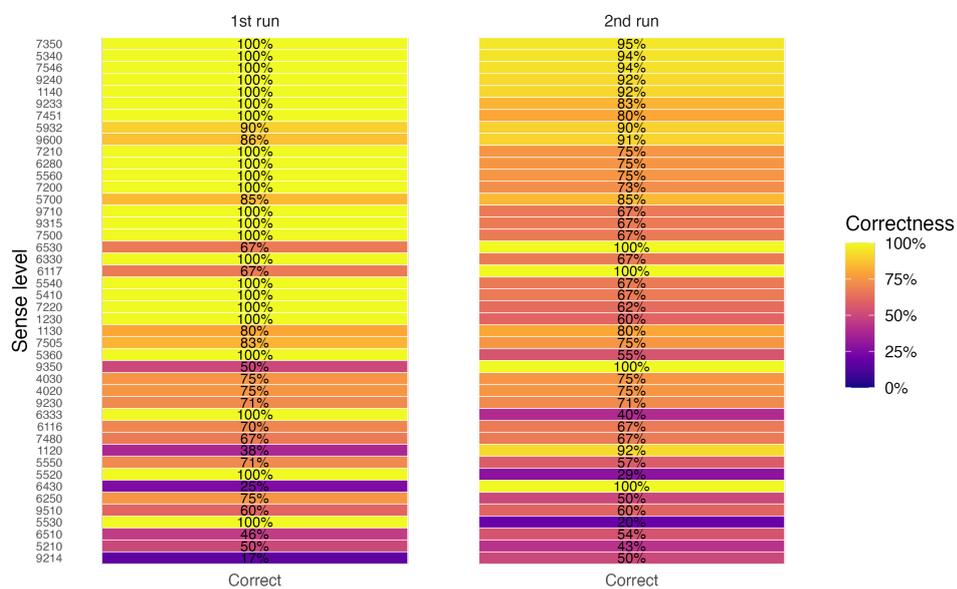


Figure 7: Correctness of sense-levels at the evaluation level *correct* across the 1st and the 2nd run; shown are only those sense-levels that exhibit volatility either within a run or between runs and that are evaluated as *correct* in at least one of the two runs. Sense-level labels are detailed in Figure 8.

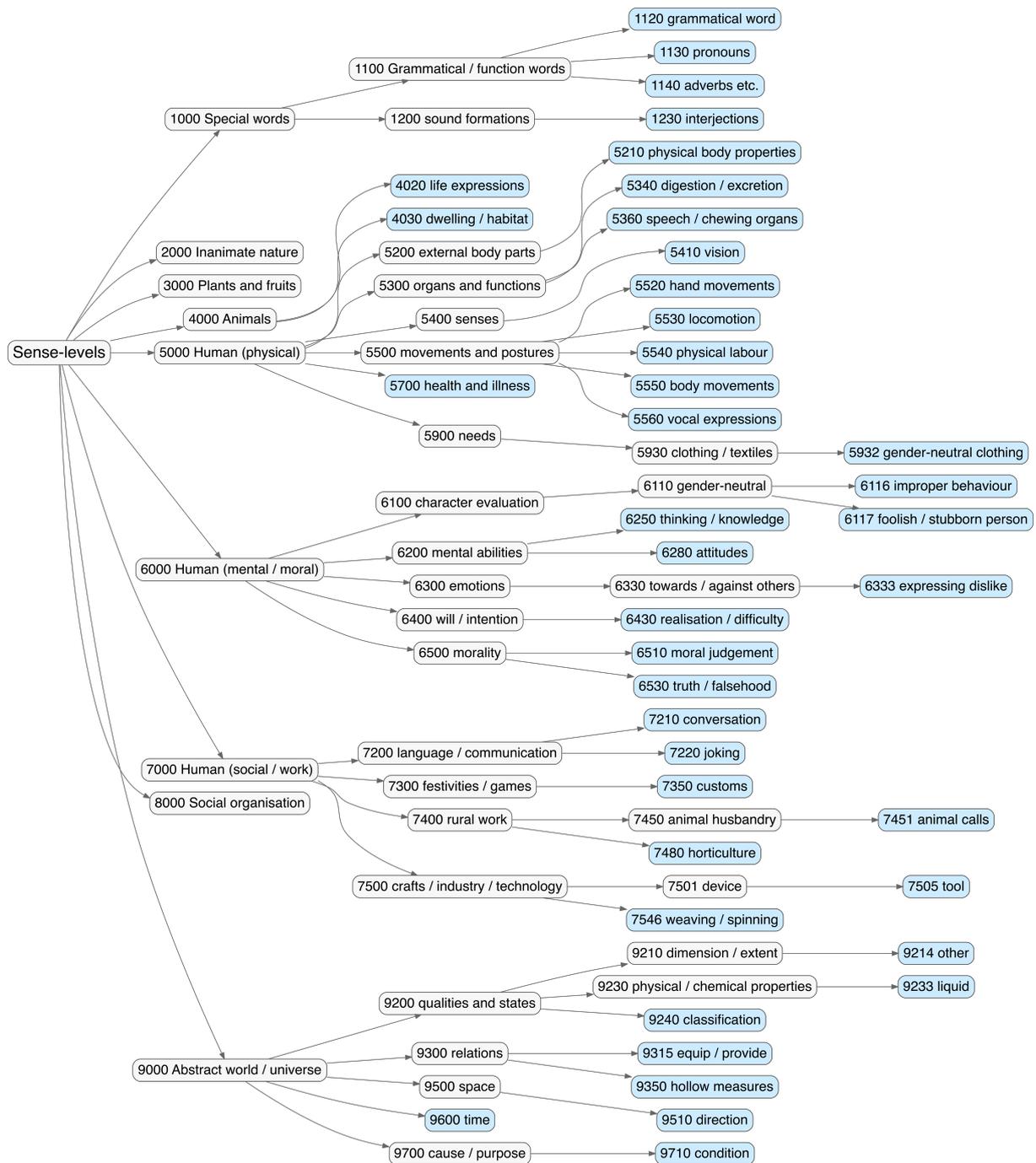


Figure 8: Taxonomic positions of sense-levels exhibiting volatility at the evaluation level *correct*. The same set of sense-levels as in Figure 7 is shown, positioned within the semantic taxonomy following Post (2026). Blue nodes mark the volatile endpoints of the taxonomic paths.

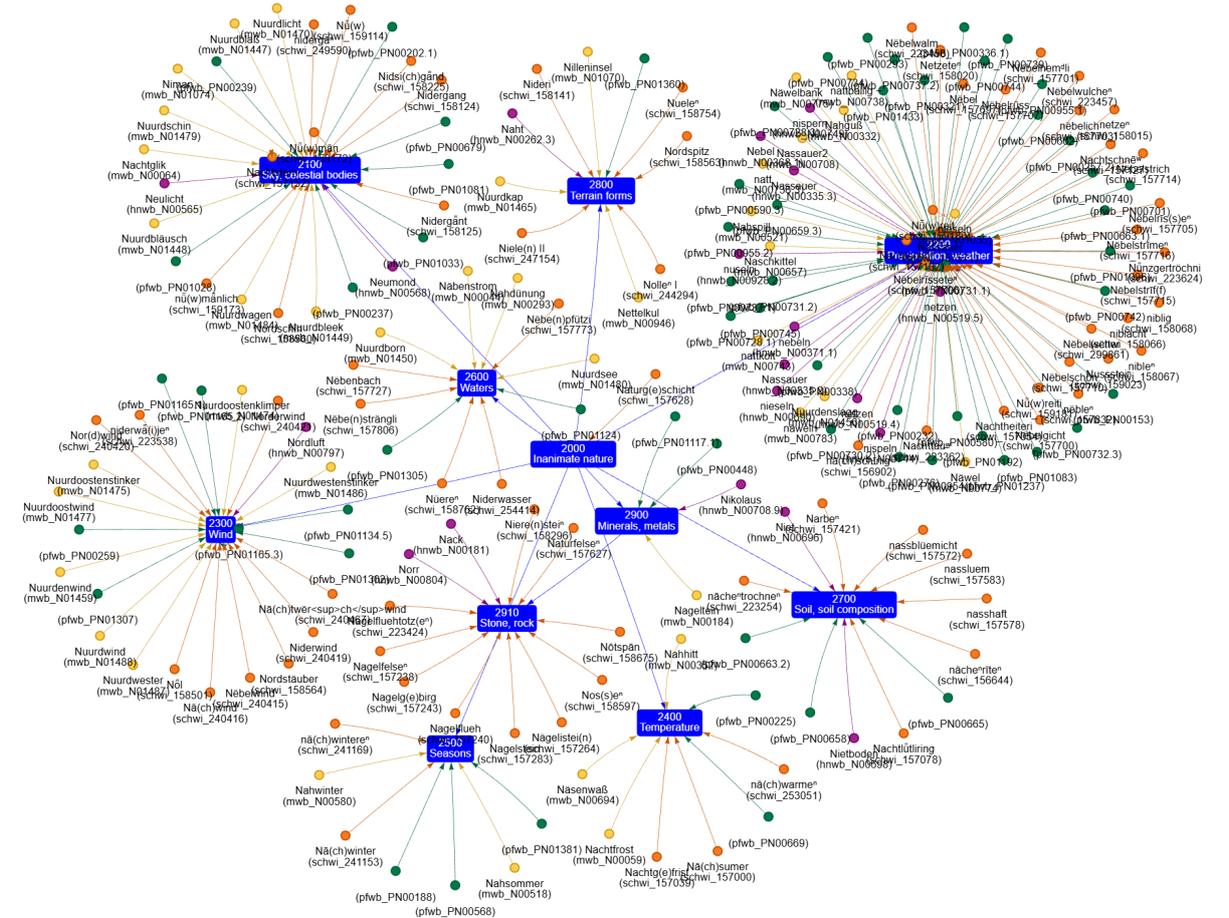


Figure 9: Semantic network based on the semantic taxonomy following (Post, 2026), showing clustering of sense-level entries across all subcategories under node 2000 “Inanimate Nature”, with assignments from the 1st, 2nd, and 3rd run merged into a shared conceptual space.

Role and Goal

- You are a linguistic classification model for historical lexicography.
- Task: Assign exactly one Hallig-Wartburg category (four-digit taxonomy ID) to each input line (an entry/sense with unique article_id).
- Think silently and return exclusively the required output format.

Context and Resources

- Binding taxonomy: Four-digit ID → Short designation. The taxonomy is authoritative and overrides external world knowledge. Use only IDs from this list; do not invent IDs.

[...]

Block 1 – Understanding and Preprocessing

- Parse the JSON fields. Use meaning (meaning specification) and translation_german (High German translation) as primary semantic source.
- Use lemma/lemma variants/grammatik information/ translation_latin(Latin, scientific designation) as secondary cues; draw morphological conclusions only if the lemma is unambiguously identified as Standard German. [...]

Block 2 – Feature Extraction → Candidates → Evaluation (incl. Consistency and Ambiguity)

[...]

- Derive suitable taxonomy candidates from the signals: first broad domain, then descend to the specifically matching leaf node (four-digit ID). Generate 3–5 candidates internally. Prefer more specific categories over coarser categories.

[...]

Block 3 – Decision and Output

- Select the best-matching specific leaf node (exactly one four-digit ID).
- Output exactly one output line for each entry in the same order. No use of conversation history.

Strict Output Format

- Format per Line: 'article_id': 'taxonomie_id'
- taxonomie_id: exactly four digits (0–9), preserve leading zeros.
- No additional text, no explanations, no blank lines, no sorting, no "unknown".

[...]

Figure 10: Prompt extract used for sense-level alignment.

On the Intelligibility of Romance Language Varieties: Spanish and Portuguese in Europe and America

Liviu P. Dinu^{♣,♥} Ana Sabina Uban^{♣,♥}

Teodor-George Marchitan^{♣,♥} Ioan-Bogdan Iordache^{♣,♥} Simona Georgescu^{♣,♥}

University of Bucharest, [♣] Faculty of Mathematics and Computer Science,

[♣] Faculty of Foreign Languages and Literatures, [♥] HLT Research Center

{ldinu, auban}@fmi.unibuc.ro,

{teodor.marchitan, ioan.iordache}@es.unibuc.ro,

simona.georgescu@lls.unibuc.ro

Abstract

Mutual intelligibility within language families presents a significant challenge for multilingual NLP, particularly due to the prevalence of dialectal variation and asymmetric comprehension. In this paper, we present a corpus-based computational analysis to quantify linguistic proximity across Romance language variants, with a focus on major Spanish (Argentine, Chilean and European) and Portuguese (Brazilian and European) varieties and the other main Romance languages (Italian, French, Romanian). We apply a computational metric of lexical intelligibility based on surface and semantic similarity of related words to measure mutual intelligibility for the five main Romance languages in relation to the Spanish and Portuguese varieties studied.

1 Introduction

The study of mutual intelligibility (the degree to which speakers of different, yet related, languages can understand one another, cf. [Gooskens and Van Heuven \(2021\)](#)) is fundamental to both historical linguistics and practical applications in Natural Language Processing (NLP). This phenomenon is particularly pronounced within the Romance language family, a group united by common Latin ancestry but distinguished by centuries of independent evolution.

In this work, we present a systematic, corpus-based investigation into mutual intelligibility across major Romance language variants, aiming to quantify their linguistic proximity and understand the factors driving comprehension success or failure, from a lexical perspective. Specifically, we focus on different Spanish (Argentine, Chilean and European) and Portuguese (Brazilian and European) varieties in relation to the other languages in the main Romance language family (Spanish, Italian, French, Romanian and Portuguese).

We rely on a complete database of related words in the Romance languages ([Dinu et al., 2023](#)), as well as on a collection of corpora for each of these languages and varieties. By analyzing shared lexical features based on their vocabularies and corpus usage, we attempt to answer the following research questions:

1. Can computational metrics accurately model the known asymmetry of intelligibility (e.g. Portuguese speakers generally understanding Spanish better than the reverse)?
2. How do regional variations in Spanish and Portuguese impact intelligibility with non-Iberian Romance languages compared to their European variants?

We propose to explore the issue of language mutual intelligibility, taking as our basis the Romance lexicon present in the five main Romance languages in terms of number of speakers (Spanish, Portuguese, French, Italian, Romanian). We focus on vocabulary because it represents the surface of language, the first layer we come into contact with when we approach a new language ([Gooskens and Van Heuven, 2021](#)). Later, when we hear or read a language that is genetically close to our mother tongue or another language we know, before we comprehend the morphology and syntax, we understand the vocabulary. Several studies have indicated that, overall, morpho-syntactic differences contribute less to intelligibility than lexical and phonological differences ([Hilton et al., 2013](#); [Gooskens and Schneider, 2016](#)). Words are the interface that either makes the language seem familiar and therefore intelligible, or raises a wall that is impossible to cross without prior training ([Gooskens, 2024](#)).

By addressing the issue of mutual intelligibility among Romance languages, we implicitly explore the issue of language similarity. Language

similarity is one of the most challenging research problems addressed in historical and comparative linguistics, because of the difficulty of finding a unitary scientific method to quantify the degree of closeness between languages. In most cases, the similarity of natural languages is treated as a fairly vague notion (McMahon and McMahon, 1995), precisely because it is used by both academic scholars and the general public, giving rise to more or less justified intuitions about which languages are more similar to which others. In many situations, these intuitions are simply based on the very subjective opinions of either linguists or non-linguists¹.

When it comes to the degree of relatedness between languages, we believe that the best unit of measurement is the ability of a native speaker of one language to understand other languages in the same family, without having any prior knowledge of them (see also Gooskens (2013)). That is why our approach is based on measuring the degree of intelligibility between any two pairs of Romance languages.

We chose to address intelligibility at the level of linguistic varieties, not just standard languages, in order to track any asymmetry in the ability of speakers of other Romance languages to understand varieties in relation to the standard language. Personal experience has often allowed us to observe a higher level of understanding of South American Spanish by Italian or Romanian speakers, but a lower level when it comes to French or Portuguese speakers. Therefore, the study of the Chilean and Argentine varieties of Spanish tests this empirical observation using a scientific method. In addition, we add the Brazilian variety of Portuguese to provide a basis for comparison and to be able to draw more general conclusions about the relationship between the level of intelligibility within and beyond the borders of Europe.

Therefore, the findings of this paper offer a significant contribution to the field of computational sociolinguistics. The resulting quantitative models of cross-lingual and intra-lingual intelligibility can be directly applied to enhance several multilingual NLP tasks, including low-resource cross-lingual transfer learning, dialect-aware machine translation, and the design of more robust language pro-

¹An anecdote related by McMahon (1994) shows how subjective the perception on similarity is: when asked how close Hungarian and Finnish are, a connoisseur of the two languages answered "oh, very", but when asked whether "like Italian and Spanish", the answer was "not that close, more like English and Persian".

cessing tools for global Romance-speaking communities.

2 Methodology

In order to obtain intelligibility scores across the Romance languages and the varieties studied, we use a metric of lexical intelligibility based on usage of words with common etymology (cognates and loanwords), as they occur in language use, considering that intelligibility of a language for a non-speaker relies on related words which are similar in form and meaning.

2.1 Dictionaries and Corpora

Since we start from the premise that intercomprehension is based on the ability of the average speaker to identify words in the flow of speech, we ground our analysis on authentic acts of speech, not simply lexicographic works. Therefore, we analyze two public corpora for the five Romance languages, RomCro - containing literary texts in different languages, translated in Romance languages and Croatian (Mikelenić et al., 2024), and EuroParl - focusing on proceedings of the European Parliament (Koehn, 2005). We extract related word frequencies in corpora used in our metrics based on the two parallel corpora. We employ a highly diverse collection of five publicly available corpora, spanning both the Spanish and Portuguese languages across distinct regional varieties and modalities. The Spanish-language component consists of three corpora. We leveraged two written reference corpora from Latin America (Argentina Corpus (Marcos-Marín and Zumárraga, 1992) and Chile Corpus (Marcos-Marín and Evans Espiñeira, 1991) each with around 2,000,000 words) to cover a broad range of formal written registers. These include substantial representations of Journalistic (Periodísticos), Humanistic, Scientific, Technical, Literary, Juridical and Scholastic (Escolares) texts, with detailed frequency breakdowns guiding the selection criteria. For Spanish oral data, we utilized the CORLEC (Corpus Oral de Referencia de la Lengua Española Contemporánea) (Marcos-Marín, 2000s), which provides transcribed spontaneous and formal speech across categories such as Conversation, Political, Educational, Scientific and various journalistic formats (e.g. debate, news, interviews). The Portuguese-language component includes both Brazilian and European varieties. The C-ORAL Brasil 2 (Raso et al., 2012–2015)

is a corpus of spontaneous Brazilian Portuguese speech, structured into three principal subcorpora: Formal in Natural Context (e.g., business, teaching, political speech), Media (e.g., interviews, talk shows), and Telephonic (public and private interactions). Lastly, the Corpus Português Fundamental (CLUL (Centro de Linguística da Universidade de Lisboa), 1984) provides data for European Portuguese, featuring a frequency corpus of spontaneous spoken language (approx. 700,000 words) and an availability corpus of thematic vocabulary (approx. 481,800 words) derived from domain-specific areas. This broad selection ensures robust coverage of stylistic, geographic and domain variation necessary for comprehensive analysis.

We perform our analysis on related word pairs extracted from the most comprehensive database of related words in cognate languages up to date, sourced from etymological dictionaries and manually curated, RoBoCoP (Dinu et al., 2023). As a source of cognate word pairs, we use the freely available subset ProtoRom (Dinu et al., 2024a), a database of cognate tuples and etymons in the five Romance languages, with 19,222 entries (tuples with at least 2 cognates). We extract borrowings from the original RoBoCoP database, totaling 46,490 borrowing pairs across Romance language pairs (Dinu et al., 2024b).

In order to identify occurrences in text for our pairs of related words (i.e. cognates and borrowings) for a given language pair, we process parallel sentence examples from the employed datasets as follows: we tokenize the sentences using spaCy (Honnibal et al., 2020), remove stop-words, and match each token with a corresponding instance from RoBoCoP, if possible. In order to account for inflections of the dictionary form from RoBoCoP perform normalization including accent removal and stemming, using a Snowball stemmer (Porter, 2001). For each sentence pair we count how many of the words from one sentence are words related to the other language. We also count, for a given related words pair, how many times that pair appears in aligned examples, in other words: how many times the related words pair corresponds to a proper translation. Table 1 shows these counts.

2.2 Surface and Semantic Similarity

The intelligibility computation is twofold: we combine the orthographic similarity of word pairs across languages with their semantic similarity.

For the former, we measure string similarity us-

ing the normalized Levenshtein distance (Levenshtein, 1966) on the orthographic (after removing accents) representation. Thus we obtain scores in the interval $[0, 1]$, where 1 means identical representations.

The latter (i.e. semantic similarity) is computed based on word representations trained on large corpora to represent meaning in context.

A metric of semantic similarity is computed using cosine similarity based on FastText word embedding vectors (Bojanowski et al., 2016). Since we are assessing similarity across languages, we need our embedding spaces to be aligned. For that reason, we use pretrained prealigned static embeddings, based on a large multilingual corpus (Wikipedia) (Joulin et al., 2018). Since even on a large corpus, such as the ones used to pretrain these embeddings, some of our related words extracted from RoBoCoP may not be represented, but an inflected form may be available, we find for each unrepresented word its closest match in form (via stemming and edit distance) that is present in the pretrained embeddings and we use that vector as our canonical representation.

2.3 Lexical Intelligibility Index

We rely on the D_{LI} metric for computing cross-lingual intelligibility automatically, introduced in Dinu et al. (2026). It is computed via the following formula:

$$D_{LI} = \frac{S_s S_L (2 - S_s - S_L)}{1 - S_s S_L} \quad (1)$$

where S_L is the formal lexical similarity between two words (i.e. 1 minus the normalized Levenshtein distance, computed on the orthographic representations), and S_s is the semantic similarity between two words (i.e. 1 minus the cosine distance between static word embeddings or the average cosine distance computed on contextual embedding clusters).

We further obtain aggregate intelligibility scores at language pair level. For each language pair (A, B) , where A is the speaker language and B is the listener language, given one of our employed corpora, we pick each sentence in language A and compute an intelligibility score with respect to B as the weighted average of the intelligibility indices of all the words in the sentence that are related to language B (i.e. the sum of these indices divided by the total number of words in the sentence, excluding stop-words). We compute a corpus-level

Table 1: Corpus Statistics: total sentences, words, unique words.

Dataset	Sentences	Words	Unique	Unique (no stop)
C-ORAL Brasil 2	9,920	182,457	13,646	13,493
CORLEC Spanish (European)	83,512	1,005,158	36,881	36,590
CORLEC Spanish Chile	44,471	920,129	47,935	47,681
CORLEC Spanish Argentina	115,957	2,239,468	61,711	61,510
Português Fundamental	8,890	97,659	8,603	8,432
TOTAL	262,750	4,444,871	168,776	167,706

Table 2: Results for Spanish variants

Corpus	es-it	es-fr	es-pt	es-ro
Argentina	0.193	0.184	0.227	0.186
Chile	0.224	0.216	0.261	0.215
Spain (oral)	0.229	0.209	0.291	0.181

Table 3: Results for Portuguese variants

Corpus	pt-it	pt-es	pt-fr	pt-ro
Brasil (oral)	0.262	0.252	0.136	0.101
Portugal	0.280	0.297	0.170	0.121

aggregate scores by combining all of the sentence-level scores (equivalent to combining all of the sentences into a singular text). In this way, the final overall intelligibility score between two languages will be affected both by the usage of related words in corpora in context as well as by their mutual individual intelligibility.

3 Results Analysis

The results in Table 2 only partially confirm our initial assumption: the highest level of intelligibility for a Romanian listening to the three varieties of Spanish will be in relation to the Chilean variety, followed by the Argentine variety, with European Spanish coming in third place. In contrast, an Italian will understand peninsular Spanish better, followed closely by the Chilean variety and, lastly, the Argentine variety. For French speakers, the order will be the same as for Romanian speakers, only with a higher level of comprehension of peninsular Spanish. For Portuguese speakers, peninsular Spanish will be by far the most easily understood, which should come as no surprise given the geographical proximity and constant linguistic contact over the centuries (see also Gooskens (2024)). If we look at the results for Portuguese vs. Brazilian (table

3), we see in all language pairs the prevalence of peninsular Portuguese over the Brazilian variety.

There are two factors that can explain these differences in intelligibility. First, there is an inequality in the type of data: the corpus for peninsular Spanish is oral, while for the Chilean and Argentine varieties it consists of written texts. It is true that today diaphasic and diamesic differences are not as significant as they were a century ago, when access to formal education was limited; however, in oral communication there is a tendency to use familiar variants, with colloquial nuances, which are not found with the same frequency in writing. Therefore, a Romanian speaker who has no direct contact with Spanish and its evolution in everyday language will have greater difficulty understanding current speech. In addition, in a written text the style is more refined and the lexicon is not marked by sociolinguistic features, which makes it easier to achieve a much higher degree of intelligibility.

Secondly, one must take into account the specific characteristics of South American varieties, which are characterized by two trends that differ from those observed in European Spanish. On the one hand, the lexicon of the main corpus is somewhat more conservative than that of the peninsular variety, which is understandable from a historical perspective: the Spanish brought by colonists to South America is that of the 16th and 17th centuries, which, even though it has evolved, like any living language, retains features of classical Spanish that the language in Spain has lost; therefore, there is a greater chance of finding more Latinisms and more old meanings, which are also shared by the Romance language at the eastern end of Europe, Romanian. To give an example, in South American Spanish, the verb most commonly used for "to catch" is *prender*, a cognate of Romanian *prinde*, used with the same meaning (from Latin *prehendere* 'to catch'). In European Spanish, however,

a different verb, *pillar*, is increasingly preferred for this sense. It was borrowed from Italian and has no cognates in Romanian, making it potentially incomprehensible to a Romanian speaker.

The second trend is the much greater permeability to Anglicisms of South American varieties (also as a result of historical realities) compared to Peninsular Spanish, a permeability that makes them easier to understand for speakers of a language that adopts a large amount of vocabulary from English and who learn English from an early age, such as Romanians. For example, in South American varieties, speakers use the word *rancho* 'ranch' borrowed from En. *ranch*, which is no doubt more comprehensible for Romance speakers with a basic knowledge of English than *finca* or *hacienda*, its synonyms preferred in Peninsular Spanish.

In future work, including a similar analysis based on phonetic representations and taking into account regional pronunciations might be useful to confirm the differences in oral communication versus written text for intelligibility. It has been previously observed that identifying orthographic correspondences between languages can increase the intelligibility (Fischer et al., 2016), while pronunciation may pose a barrier to cross-linguistic intelligibility. We offer a concrete example: in standard European Spanish, the phonetic sequences /ce/ and /ci/, as well as /z/ in any position, are pronounced [θ], whereas in South America they are pronounced [s]. The [θ] sound is absent in the other Romance languages under consideration, which makes the South American pronunciation more comprehensible for speakers of these languages than European Spanish: for example, the word *zebra* pronounced [sébra] would be more intelligible for the average Romance language speaker than [θébra]. However, we can also encounter the opposite situation: a Romanian, Italian, French or Portuguese listener who hears the Spanish word *hacer* pronounced in a South American variety [asér] may interpret it as the sequence *a ser* (preposition *a* "to" + infinitive *ser* "be"), through a false analysis — provided they have at least minimal knowledge of Spanish. This South American pronunciation can be particularly misleading if the listener has previously been exposed to European Spanish, where the verb would be pronounced [aθér]. In the absence of any prior exposure, it is most likely that the word will not be understood when encountered in the flow of speech. However, when presented in written form, comprehension is more likely, even without prior knowl-

edge. This may be partly due to speakers' ability to reconstruct meaning from context, even when elements are unintelligible in isolation — a capacity well documented in cloze tests (cf. Gooskens and van Heuven (2017); Gooskens (2024)). It may also result from the inherent orthographic similarity to the speaker's native-language verb, e.g. Ro. *face* / *facere*, whose Levenshtein distance from *hacer* is relatively small.

4 Conclusion

We proposed a study of the mutual intelligibility of Romance languages, with a focus on different varieties of Spanish and Portuguese. We apply a metric of lexical intelligibility based on surface and semantic similarity of related words to compute intelligibility between each of the five main Romance languages and European and American varieties of Spanish and Portuguese, based on a selection of corpora of Argentine, Chilean and European Spanish, and Brazilian and European Portuguese, respectively. We thus provide a corpus-grounded quantification of linguistic distance across these Romance variants, and open the way for further research into the study of language varieties from the perspective of lexical intelligibility.

Acknowledgements

This research was supported by the Ministry of Education and Research, CNCS-UEFISCDI, project SIROLA, number PN-IV-P1- PCE-2023-1701, within PNCDI IV.

References

- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606*.
- CLUL (Centro de Linguística da Universidade de Lisboa). 1984. Corpus Português Fundamental. Linguistic Resources, Centro de Linguística da Universidade de Lisboa.
- Liviu P Dinu, Ana Uban, Alina Cristea, Ioan-Bogdan Iordache, Teodor-George Marchitan, Simona Georgescu, and Laurentiu Zoicas. 2024a. Verba volant, scripta volant? Don't worry! There are computational solutions for protoword reconstruction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6314–6326.
- Liviu P Dinu, Ana Uban, Anca Dinu, Ioan-Bogdan Iordache, Simona Georgescu, and Laurentiu Zoicas.

- 2024b. It takes two to borrow: a donor and a recipient. Who's who? In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6023–6035.
- Liviu P Dinu, Ana Sabina Uban, Alina Maria Cristea, Anca Daniela Dinu, Ioan-Bogdan Iordache, Simona Georgescu, and Laurentiu Zoicas. 2023. Robocop: A comprehensive romance borrowing cognate package and benchmark for multilingual cognate identification. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.
- Liviu P Dinu, Ana Sabina Uban, Bogdan Iordache, Anca Dinu, and Simona Georgescu. 2026. [Measuring cross-language intelligibility between romance languages with computational tools](#). Preprint, arXiv:2602.07447.
- Andrea Fischer, Klára Jágrová, Irina Stenger, Tania Avustinova, Dietrich Klakow, and Roland Marti. 2016. Orthographic and morphological correspondences between related slavic languages as a base for modeling of mutual intelligibility. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4202–4209.
- Charlotte Gooskens. 2013. Experimental methods for measuring intelligibility of closely related language varieties.
- Charlotte Gooskens. 2024. *Mutual intelligibility between closely related languages*, volume 30. Walter de Gruyter GmbH & Co KG.
- Charlotte Gooskens and Cindy Schneider. 2016. Testing mutual intelligibility between closely related languages in an oral society.
- Charlotte Gooskens and Vincent J van Heuven. 2017. Measuring cross-linguistic intelligibility in the germanic, romance and slavic language groups. *Speech communication*, 89:25–36.
- Charlotte Gooskens and Vincent J Van Heuven. 2021. Mutual intelligibility. *Similar languages, varieties, and dialects: A computational perspective*, pages 51–95.
- Nanna Haug Hilton, Charlotte Gooskens, and Anja Schüppert. 2013. The influence of non-native morphosyntax on the intelligibility of a closely related language. *Lingua*, 137:1–18.
- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, and 1 others. 2020. spacy: Industrial-strength natural language processing in python.
- Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.
- Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit x: papers*, pages 79–86.
- VI Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics-doklady*, volume 10.
- Francisco A. Marcos-Marín. 2000s. Corpus Oral de Referencia de la Lengua Española Contemporánea (CORLEC). Web Archive of the Laboratorio de Lingüística Informática, Universidad Autónoma de Madrid.
- Francisco A. Marcos-Marín and Ernesto Evans Espiñeira. 1991. Corpus Lingüístico de Referencia de la Lengua Española en Chile. Web Archive of the Laboratorio de Lingüística Informática, Universidad Autónoma de Madrid.
- Francisco A. Marcos-Marín and Verónica Zumárraga. 1992. Corpus Lingüístico de Referencia de la Lengua Española en Argentina. Web Archive of the Laboratorio de Lingüística Informática, Universidad Autónoma de Madrid.
- April MS McMahon. 1994. *Understanding language change*. Cambridge university press.
- April MS McMahon and Robert McMahon. 1995. Linguistics, genetics and archaeology: internal and external evidence in the amerind controversy. *Transactions of the Philological Society*, 93(2):125–225.
- Bojana Mikelenić, Antoni Oliver, and Marko Tadić. 2024. Expansion of the romcro corpus with texts in catalan. In *CLARIN Annual Conference Proceedings 2024*, pages 135–139. Barcelona: CLARIN.
- Martin F Porter. 2001. Snowball: A language for stemming algorithms.
- Tommaso Raso, Heliana Mello, Maryualê Malvessi Mittmann, and Alessandro Panunzi. 2012–2015. C-ORAL Brasil 2. Corpus website of the Laboratory of Psycholinguistics, Federal University of Minas Gerais.

Dialect Matters: Cross-Lingual ASR Transfer for Low-Resource Indic Language Varieties

Akriti Dhasmana and Aarohi Srivastava and David Chiang

Computer Science and Engineering

University of Notre Dame

Notre Dame, IN, USA

{adhasman, asrivas2, dchiang} @nd.edu

Abstract

We conduct an empirical study of cross-lingual transfer using spontaneous, noisy, and code-mixed speech across a wide range of Indic dialects and language varieties. Our results indicate that although ASR performance is generally improved with reduced phylogenetic distance between languages, this factor alone does not fully explain performance in dialectal settings. Often, fine-tuning on smaller amounts of dialectal data yields performance comparable to fine-tuning on larger amounts of phylogenetically-related, high-resource standardized languages. We also present a case study on Garhwali, a low-resource Pahari language variety, and evaluate multiple contemporary ASR models. Finally, we analyze transcription errors to examine bias toward pre-training languages, providing additional insight into challenges faced by ASR systems on dialectal and non-standardized speech.

1 Introduction

Automatic speech recognition (ASR) systems for Indic languages have made significant progress through large-scale multilingual pre-training on high-resource, standardized languages. However, these advances have largely overlooked the linguistic reality of the region, where speech is characterized by extensive dialectal variation, spontaneous and noisy recording conditions, and frequent code-switching. As a result, state-of-the-art Indic ASR models often perform poorly when applied to low-resource dialects and language varieties, even those that are closely related to languages seen during pre-training.

Such efforts for Indic ASR typically rely on pre-training on a small set of mainstream languages such as Hindi, Marathi, and Bengali, under the assumption that phylogenetic similarity will facilitate transfer to related varieties. While this as-

sumption provides a useful baseline in many cross-lingual settings, its sufficiency for dialectal speech remains underexplored. In practice, dialects and non-standard varieties often exhibit distinct phonological, lexical, and orthographic properties that are not well-captured by standard language data, raising questions about the effectiveness of relying solely on high-resource languages for transfer in dialectal ASR.

In this work, we examine cross-lingual transfer for Devanagari-script Indic varieties using spontaneous, noisy, and code-mixed speech. Rather than focusing solely on phylogenetic similarity, we investigate how different choices of fine-tuning data affect ASR performance on dialectal speech. Our results suggest that incorporating even limited amounts of dialectal speech during fine-tuning can be as effective as, or more effective than, relying on larger amounts of standardized language data, highlighting the role of dialectal variation in shaping transfer behavior.

To complement our cross-lingual analysis, we present a case study on Garhwali, a low-resource Pahari language variety that has been largely absent from prior ASR research. We evaluate several self-supervised speech models on Garhwali and perform a detailed error analysis to characterize the challenges posed by dialectal variation, code-mixing, and model bias toward pre-training languages. This case study provides concrete insights into the limitations of current Indic ASR models when applied to low-resource language varieties.

Contributions In this paper, we make three contributions:

1. We provide a comprehensive empirical analysis of cross-lingual transfer for low-resource Devanagari-script Indic dialects and show that fine-tuning on other dialects is often more effective than relying on closely-related high-resource standardized languages.

2. We present the first detailed ASR study for Garhwali, including model evaluation and qualitative and quantitative error analysis.
3. We introduce a diagnostic approach for quantifying bias toward pre-training languages in dialectal ASR, enabling systematic analysis of model behavior on non-standardized and code-mixed speech.

2 Related Work

Recent work has highlighted systematic performance gaps between standard language varieties and regional or minority dialects across NLP and speech technologies. [Kantharuban et al. \(2023\)](#) provide a large-scale evaluation of state-of-the-art models for machine translation and automatic speech recognition across regional dialects of multiple high- and low-resource languages, showing that dialectal performance disparities are widespread and variably correlated with linguistic, social, and data-related factors. Complementary to this line of work, [Blaschke et al. \(2025\)](#) study standard-to-dialect transfer in spoken and written settings for German, demonstrating that speech-based models are more robust to dialectal variation than text-based or cascaded approaches, particularly when orthographic normalization is involved.

In the Indic context, prior work has investigated dialectal variation primarily through multilingual or language-specific ASR systems. For example, [\(Kumar et al., 2025\)](#) evaluate a multilingual dialect identification and ASR pipeline across 33 dialects of eight Indic languages using read speech, while other efforts have focused on building comprehensive ASR toolkits for standardized Indic languages ([Singh Chadha et al., 2022](#)). However, cross-dialectal and cross-lingual transfer for Indic ASR especially in zero-shot settings and on spontaneous, non-standardized speech has not yet been systematically examined. Our work addresses this gap by analyzing cross-lingual transfer behavior and dialectal bias in low-resource Indic language varieties.

The effects of orthographic irregularity and transcription variability on ASR performance have been examined in other language contexts, including cross-family settings ([Taguchi and Chiang, 2024](#)) and Swiss German dialectal ASR ([Nigmatulina et al., 2020](#)). We build on these findings to analyze how similar forms of orthographic variation impact ASR performance for Indic dialects.

Prior work on speech technology for Garhwali has primarily focused on language identification ([Gusain et al., 2023](#)) and the creation of domain-specific datasets, such as for agriculture-related applications ([Riyal et al., 2016](#)). To the best of our knowledge, no prior work has involved training or evaluating an ASR model for Garhwali.

The widespread presence of multiple language varieties in India, shaped by historical and sociolinguistic factors, has led to extensive code-mixing in everyday speech, posing additional challenges for ASR systems. While code-mixed speech has been studied in specific settings such as Hindi-Marathi ASR ([Palivela et al., 2025](#)), we provide a systematic way to quantify transcription errors arising from code-mixing in dialectal ASR.

3 Languages and Data

The Indic linguistic landscape is characterized by a high degree of diversity, encompassing hundreds of languages and dialects with varying levels of standardization and resource availability. Many widely spoken languages coexist with numerous regional dialects that differ substantially in phonology, morphology, and lexicon, despite phylogenetic relatedness ([Massica, 1993](#)). In everyday use, speakers frequently engage in code-mixing, particularly with English, and in spontaneous and acoustically noisy environments. These factors pose major challenges for ASR systems trained primarily on clean, standardized, and monolingual speech ([Diwan et al., 2021](#)). We present a phylogenetic tree of the languages included in our experiments in Figure 1.

Most existing ASR resources for Indic languages focus on a small subset of high-resource, standardized languages, leaving dialectal varieties underrepresented or entirely absent. To address this gap, we fine-tune and evaluate our models on the VAANI dataset ([VAANI, 2025](#)).¹ VAANI captures the rich linguistic diversity present across speakers of Indic languages, including different accents, grammatical variation, loan words, and code-switching patterns, all of which are vital to include in the context of Indic ASR.

VAANI consists of spontaneous speech collected by prompting participants to describe images in their local dialect. The dataset contains over 150,000 hours of speech, of which approx-

¹<https://huggingface.co/datasets/ARTPARK-IISc/VAANI>

imately 10% is transcribed, and covers 156,534 speakers from 773 districts across India. As Indic languages are written in multiple scripts, we focus our analysis on language varieties written in the Devanagari script. Due to the broad coverage of dialects and language varieties, the amount of available data varies substantially across languages in the dataset; we account for this variability as much as possible in our experimental design.

Beyond conducting cross-lingual experiments across Indic language varieties, we also conduct a detailed analysis of ASR for one nonstandard language variety. We select *Garhwali* for this case study.² *Garhwali* is a language that belongs to the Pahari *subgroup*³ spoken in the Himalayan region, particularly in the state of Uttarakhand (Gusain et al., 2023).

4 Experiments and Results

In our experiments, we aim to answer the following questions to analyze broader patterns in Indic ASR:

- RQ1. How does a strong Indic ASR baseline perform across a diverse set of low-resource language varieties, particularly those not seen during model pre-training?
- RQ2. To what extent does orthographic variability correlate with ASR performance in Indic language varieties?
- RQ3. How does cross-lingual transfer for Indic ASR vary with phylogenetic distance, and does this relationship hold for dialectal speech?

In addition, we conduct a focused case study on *Garhwali* to characterize systematic error patterns in dialectal ASR. We analyze how systemic bias towards pre-training languages (Hindi) manifest in transcriptions generated by ASR models fine-tuned on a dialect (*Garhwali*).

Metrics We employ two standard metrics, word error rate (WER) and character error rate (CER), to evaluate the generated transcriptions. Both metrics are calculated based on Levenshtein distance, representing the minimum number of insertions, deletions, and substitutions required to align the

²According to the Post-1971 Census of India, these languages were recognized as “mother tongues” and were designated as dialects of Hindi (Khubchandani, 1991); however, some linguists argue that *Garhwali* and other Pahari languages are distinct languages (Gusain et al., 2023).

³We refer to the lowest non-leaf nodes in a phylogenetic tree as *subgroups*.

hypothesis with the reference text. We report WER in our main results and include CER in the appendix.

4.1 RQ1: Baseline Assessment for Indic ASR

Setup We begin by assessing the performance of a state-of-the-art Indic ASR model, *IndicWav2Vec* (Kumar et al., 2022), on language varieties present in the VAANI dataset. *IndicWav2Vec* is based on the *Wav2Vec 2.0* architecture (Baevski et al., 2020), and is pre-trained on 17,000 hours of unlabeled *clean* speech data from YouTube, as well as *Newsonair* data curated from radio channels, across 40 Indic languages. In our experiments, we employ *IndicWav2Vec-Hindi*⁴ (*IndicWav2Vec* fine-tuned on Hindi), since pre-trained ASR models largely learn language-agnostic acoustic representations during pre-training and require language-specific fine-tuning to map these representations to text transcriptions.

Results We evaluate *IndicWav2Vec-Hindi* on test samples for all 30 Devanagari-script varieties in VAANI (up to 1 hour each) and report these results as part of a phylogenetic tree in Figure 1. We also divide the results into two tables depending on whether the language was used to pre-train *IndicWav2Vec* (Table 7) or it was not explicitly seen by the model (Table 8). These results show that, despite being fine-tuned on Hindi, *IndicWav2Vec-Hindi* achieves a best-case word error rate of 50.4% on VAANI Hindi test speech (one hour). Similarly high error rates are observed for several languages seen during pre-training, indicating that pre-training alone does not ensure robust performance on spontaneous and noisy speech, even for languages included in the pre-training corpus. We also observe substantial variation in performance across dialects and closely-related language varieties. These observations suggest that other factors, such as geographical proximity or phonological feature similarity, could have an impact on performance, motivating our subsequent analyses.

4.2 RQ2: Orthographic Consistency

Setup Orthographic consistency is an important factor in achieving high ASR performance, as inconsistent spellings introduce additional variability

⁴<https://huggingface.co/ai4bharat/IndicWav2Vec-Hindindi>

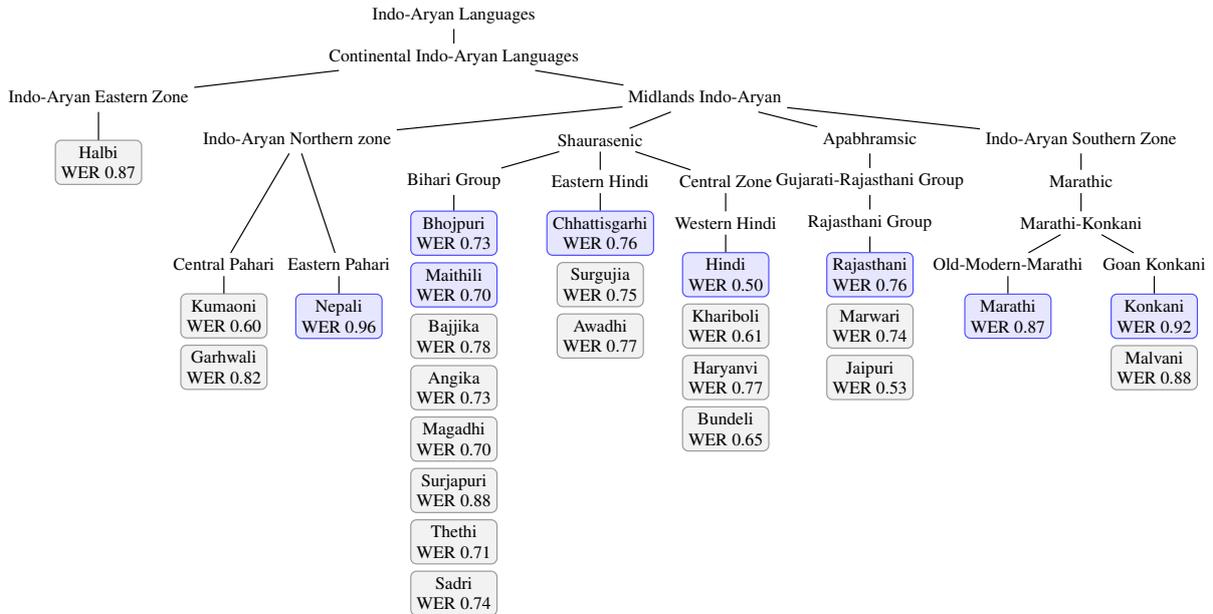


Figure 1: Subset of the Indo-European language family tree showing Devanagari-script Indic languages in the VAANI dataset, based on Glottolog (Hammarström et al., 2024). Languages are annotated with WER for IndicWav2Vec-Hindi. Blue highlights indicate languages used during pre-training.

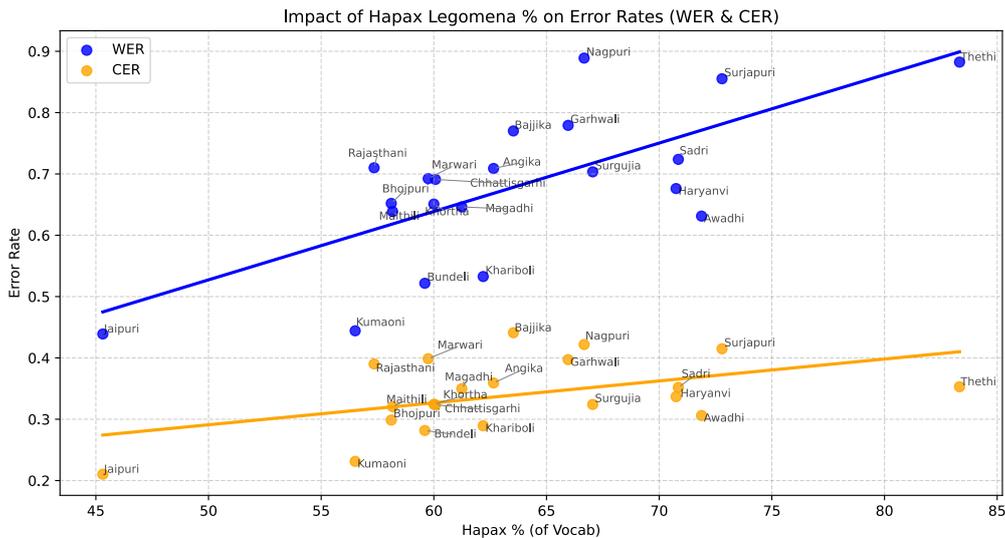


Figure 2: Hapax legomena (once-seen word types) % per language in the test split of VAANI plotted against WER (Pearson’s $\rho = 0.705$, $p = 4 \times 10^{-4}$) and CER (not significant) using IndicWav2Vec-Hindi.

ity that can increase modeling and decoding errors (Nigmatulina et al., 2020). This challenge is particularly pronounced for Indic languages that encompass multiple dialects where orthographic conventions are not fixed. To quantify orthographic irregularity in the data, we measure the frequency of each word type in the transcripts and analyze the proportion of tokens that occur only once, along with the type-to-token ratio. We then examine how these measures correlate with ASR performance.

Results We observe a clear trend between orthographic variability, as measured by the proportion of *hapax legomena* (unique words) in the test set, and ASR performance. Figure 2 shows that languages with a higher percentage of unique words exhibit higher WERs when evaluated with IndicWav2Vec-Hindi (Pearson’s $\rho = 0.705$, $p = 4 \times 10^{-4}$), indicating increased difficulty for languages with less consistent orthographic conventions (e.g., Thethi, Surjapuri). A similar but

weaker trend is observed for CER.

Additional evidence is provided by the type-token statistics reported in Table 9. Several languages (e.g., Gondi, Thethi) exhibit a high number of distinct word types relative to the total number of tokens, resulting in elevated type-to-token ratios. These measures reflect greater orthographic and lexical variability and are associated with higher ASR error rates. Taken together, these results suggest that orthographic inconsistency is an important contributing factor to reduced ASR performance in nonstandard Indic varieties.

4.3 RQ3: Cross-Lingual Transfer

We wanted to examine whether there are any performance gains from cross-dialectal fine-tuning in zero-shot settings. For each of these experiments, we fine-tune w2v-bert-2.0⁵ (selected through preliminary comparison of models) on 1 to 7 hours of speech, depending on the availability. Each model was then evaluated on the test set of all Devanagari script languages available from VAANI. We compute correlations between the error rate and phylogenetic distance under two conditions: across all evaluation languages and restricted to nonstandard test varieties. We additionally measure the association between the error rate and whether the fine-tuning language represents dialectal or standard speech. We consider the following languages from the dataset to be dialects or nonstandard varieties: Angika, Awadhi, Bajjika, Bhili, Chhattisgarhi, Garhwali, Halbi, Haryanvi, Jaipuri, Khariboli, Khortha, Kumaoni, Magadhi, Malvani, Marwari, Nagpuri, Surjapuri, and Thethi.⁶

Results Figure 3 summarizes cross-lingual transfer performance for w2vBERT models fine-tuned on 1 to 7 hours of data per language and evaluated in a zero-shot setting across a wide range of Indic language varieties, providing a holistic view of how ASR performance varies as a function of both the fine-tuning and the evaluation languages.

Across all evaluation languages, we observe a clear association between phylogenetic distance and ASR performance: larger distances between the fine-tuning and evaluation languages generally correspond to higher word error rates (Spear-

man’s $\rho = 0.333$, $p = 1.11 \times 10^{-7}$). This result aligns with prior findings in cross-lingual ASR and confirms that phylogenetic relatedness provides a useful baseline for transfer across Indic languages. However, this aggregate trend does not fully characterize model behavior when evaluation is restricted to dialectal speech.

4.3.1 Dialectal Transfer Effects

When focusing specifically on dialects and non-standard test varieties, the heatmap reveals additional structure that is not explained by phylogenetic distance alone. Although the association between phylogenetic distance and word error rate remains statistically significant in this setting (Spearman’s $\rho = 0.274$, $p = 3.604 \times 10^{-4}$), several consistent deviations from this trend are observed. In particular, models fine-tuned on non-standard varieties frequently outperform models fine-tuned on phylogenetically closer but standard, higher-resource languages.

One representative example from Figure 3 is the strong transfer from Marwari (a dialect of Rajasthani) to Kumaoni (a Pahari language variety), even though the two belong to distinct phylogenetic subgroups and are geographical distant. Another example is the consistently competitive performance of the models fine-tuned on Marwari and Magadhi across multiple evaluation language varieties, often outperforming models fine-tuned on higher-resource standardized languages such as Hindi, Marathi, and Rajasthani. Notably, this trend holds despite the fact that the Magadhi model is trained on less data (5 vs. 7 hours). These cases illustrate that while fine-tuning on a closely related high-resource standardized language may be the most natural strategy, it is not necessarily the most effective choice for transfer to dialects and related language varieties.

To further examine this pattern, we analyze whether the dialectal status of the fine-tuning language itself is associated with performance on unseen dialects. We find a statistically significant trend indicating that fine-tuning on dialectal speech is associated with lower WER on dialectal evaluation sets (point-biserial correlation $r_{pb} = -0.196$, $p = 5.79 \times 10^{-3}$). This trend holds even when the available dialectal training data is smaller than that of the corresponding standardized languages, suggesting that dialectal fine-tuning captures information that is not adequately represented by mainstream language data alone.

⁵<https://huggingface.co/facebook/w2v-bert-2.0>

⁶None of these languages belong to the Eighth Schedule to the Indian Constitution listing the officially recognized languages (https://en.wikipedia.org/wiki/Eighth_Schedule_to_the_Constitution_of_India).

Cross-Lingual Performance Heatmap

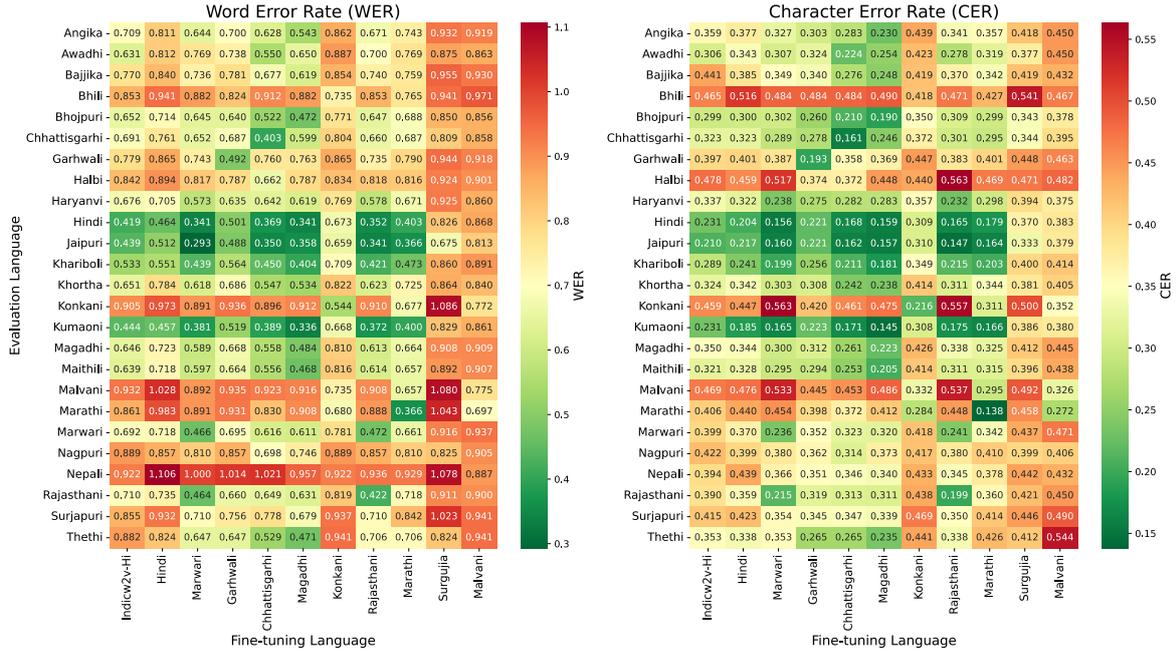


Figure 3: Cross-Lingual Performance of w2vBERT models fine-tuned on 1 to 7 hours of data per language vs. off-the-shelf IndicWav2Vec-Hindi.

Fine-Tuning Language	Hours	Jaipuri WER	CER
Rajasthani + Marwari	16.8	0.309	0.151
Marwari (<i>dialect only</i>)	7.2	0.301	0.162

Table 1: Effect of fine-tuning language choice on Jaipuri ASR (Rajasthani subgroup). Mainstream language is in bold.

Controlled In-Group Transfer Experiments

To isolate the effect of dialectal vs. standard fine-tuning data from broader cross-lingual trends, we conduct controlled transfer experiments within individual phylogenetic subgroups. These experiments compare models fine-tuned on dialectal varieties against models fine-tuned on higher-resource standardized languages within the same subgroup. Across the Rajasthani, Western Hindi, and Bihari subgroups, we consistently observe that fine-tuning on small amounts of dialectal speech yields performance comparable to or better than fine-tuning on substantially larger amounts of standardized language data. For example, within the Rajasthani subgroup (Table 1), a model fine-tuned solely on a dialectal variety (Marwari) achieves comparable performance on an unseen dialect (Jaipuri) to a model trained on a combination of dialectal and macro-language data. Similar trends are observed in the Western Hindi and Bihari sub-

Fine-Tuning Language	Hours	Thethi WER	CER
Angika, Bajjika, Bhojpuri, Khortha, Magadhi, Maithili, Sadri	50.6	0.412	0.132
Angika, Bajjika, Khortha, Magadhi, Sadri (<i>dialects only</i>)	15.6	0.353	0.147

Table 2: Effect of mainstream language inclusion on Thethi ASR (Bihari subgroup). Mainstream language is in bold.

Fine-Tuning Language	Hours	Haryvani WER	CER
Hindi + Bundeli + Khariboli	101.7	0.930	1.452
Bundeli + Khariboli (<i>dialects only</i>)	1.7	0.513	0.267

Table 3: Impact of fine-tuning (FT) languages on Haryanvi ASR (Western Hindi subgroup). Mainstream language is in bold.

groups (Tables 3 and 2), where excluding higher-resource mainstream languages does not degrade, and in some cases improves, performance on dialectal evaluation sets.

Taken together, the heatmap analysis and controlled in-group experiments demonstrate that while phylogenetic similarity provides a useful starting point for cross-lingual transfer, effective

Model	Metric Optimized	WER	CER
XLS-R	WER	0.650	0.270
wav2vec2-BERT	CER	0.493	0.193
Whisper-small	CER	0.629	0.650
HuBERT	CER	0.515	0.199

Table 4: Comparison of speech models fine-tuned on Garhwali. Training details are in Table 10.

ASR for dialectal speech depends critically on the inclusion of dialectal training data itself.

5 Garhwali ASR Case Study

5.1 Baseline Selection

Setup In order to determine the ideal model architecture for Garhwali, we compared several widely-used ASR models:

- Wav2Vec2 (Baevski et al., 2020)
- HuBERT (Hsu et al., 2021)
- XLS-R (Babu et al., 2022)
- Whisper (Radford et al., 2023)
- w2vBERT (Chung et al., 2021)

We fine-tune each of these models on the Garhwali subset of VAANI, using an 87%–6%–7% train–test–validation ratio.

We fine-tuned and evaluated several self-supervised speech models on the Garhwali subset of the VAANI dataset. Table 4 reports performance across these models.

Results Among the evaluated configurations, the w2vBERT-based model achieves the lowest error rates and is therefore selected for subsequent analysis. We compare this model against IndicWav2Vec-Hindi (used in our cross-lingual experiments). Our fine-tuned model performs better than the IndicWav2Vec-Hindi model; however, the resulting error rates remain high, indicating that Garhwali ASR remains challenging even with dialect-specific fine-tuning. We therefore focus our error analysis on this best-performing configuration to better understand the remaining sources of error.

5.2 Error Analysis

We conduct an extensive error analysis of the best fine-tuned model on Garhwali to identify the following:

Inconsistent English Transliteration English text is marked in VAANI’s annotations; we assess ASR output specifically on these segments.

Model	# Non-Hi	Correct	To Hi	To Wrong
wav2vec2-BERT	1873	34.8%	23.3%	38.8%
Indicw2v2	2109	4.2%	37.3%	42.9%

Table 5: Comparison of w2vBERT model fine-tuned on Garhwali against IndicWav2Vec-Hindi on Garhwali non-Hindi word handling. Key: Hi = Hindi.

A prominent source of error arises from inconsistent transliteration of English words into the Devanagari script. Due to frequent code-switching in spontaneous Indic speech, English words commonly appear in the ground-truth transcripts. However, the absence of standardized conventions for English-to-Devanagari transliteration, combined with imperfect phoneme-to-grapheme mappings, results in substantial variation in the labels. For example, the training data contains multiple spellings for the English word *photo*, such as फोटु and फोटो, reflecting pronunciation differences.

This variability introduces ambiguity during training and evaluation, complicating the model’s ability to learn consistent lexical representations. Whether standardizing transliterated forms would improve dialectal ASR performance or instead remove linguistically meaningful variation remains an open question.

Bias Towards Hindi from Pre-training We identify non-Hindi words in the ground-truth transcripts using Hindi-HunSpell,⁷ a Hindi spell-checker. These includes Garhwali words, as well as transliterated English words; we separate the English words (see below). We then assess which non-Hindi words are preserved in the generated transcription vs. which are transcribed to a Hindi word. We analyze commonly omitted, added, and substituted characters in the generated transcriptions, qualitatively identifying systematic errors and highlighting areas for future improvement. We also observe systematic bias towards Hindi arising from model pre-training. Many errors involve Garhwali words or transliterated English terms being incorrectly normalized to valid Hindi words.

To quantify this effect, we identify non-Hindi words in the ground-truth transcripts, including Garhwali-specific vocabulary and transliterated English terms, using a Hindi spell checker. We then track three outcomes: non-Hindi words cor-

⁷<https://github.com/Shreeshrri/hindi-hunspell>

Actual Deva IPA	Predicted Deva IPA	#	Actual Deva IPA	Predicted Deva IPA	#
ε	ε	438	ε	ॐ	/n/ 68
ε	ॐ	289	ॐ	ॐ	/u/ 66
ॐ	/a:/	175	ॐ	ॐ	/i:/ 55
ॐ	ε	170	ॐ	ε	/e:/ 51
ॐ	ॐ	163	ॐ	ε	/u/ 51
ॐ	/n/	145	ॐ	ε	/ɪ/ 48
ॐ	ॐ	126	ह	ε	/f/ 46
ॐ	/u:/	94	ε	ॐ	/e:/ 45
ॐ	/i/	90	ॐ	ॐ	/e:/ 44
ॐ	य	81	ε	क	/k/ 44
ॐ	/i:/	76	ब	भ	/b ⁶ / 42
य	/j/	76	र	ε	/r/ 42

Table 6: Most frequent transcription errors for Garhwali ASR. Key: ॐ = space; ε = empty string; Deva = Devanagari script.

rectly preserved (Correct), non-Hindi words converted into valid Hindi words (No-Hi→Hi), and non-Hindi words converted into incorrect forms (No-Hi→Wrong). Note that the total number of non-Hindi words in the ground truth transcripts differs slightly between the two models due to preprocessing and tokenization; however, the observed trends are robust to these discrepancies.

As shown in Table 5, the w2vBERT model fine-tuned on Garhwali converts approximately 22.9% of non-Hindi terms into Hindi words, while preserving roughly one-third of such terms. In contrast, IndicWav2Vec-Hindi preserves fewer than 5% of non-Hindi words, converting or distorting the majority. These results indicate that ASR pre-training on mainstream Indic languages can substantially hamper a model’s ability to retain dialect-specific information.

Character-Level Error Patterns We further analyze character-level errors produced by w2vBERT fine-tuned on Garhwali. Table 6 summarizes the most frequent substitutions, insertions, and deletions, which can be grouped into three broad categories: (1) word-boundary errors (e.g., space versus deletion), (2) vowel and consonant length confusions (e.g., ॐ vs. ॐ), and (3) aspiration-related errors (e.g., ब vs. भ).

Word-boundary errors are the most frequent, with 438 instances of omitted spaces and 289 instances of spurious space insertions, contributing to the observed WER despite relatively lower CER. The next most common errors involve vowel length and halant usage. In Garhwali, consonant-final words and consonant clipping, often marked using the halant ॐ, are common, whereas Hindi

typically enforces an inherent vowel at word endings. Models pre-trained on Hindi orthographic conventions therefore tend to regularize Garhwali forms toward Hindi norms, resulting in systematic deletion or insertion of the halant. These error patterns are consistent with phonological and orthographic differences between Hindi and Garhwali.

6 Conclusion

In this work, we present a comprehensive study of dialectal speech recognition across a diverse set of Devanagari-script Indic language varieties, with a particular focus on understanding how dialectal variation interacts with cross-lingual transfer in low-resource ASR. Through a large-scale empirical evaluation, we find that while ASR performance is generally associated with phylogenetic distance across languages, this factor alone does not explain performance in the dialectal setting. In particular, when evaluating on dialects in the zero-shot setting, we observe lower word error rates when the fine-tuning language is a dialect or nonstandard variety. In many cases, fine-tuning on small amounts of dialectal speech yields performance comparable to or better than fine-tuning on larger amounts of phylogenetically closer, high-resource standardized languages.

Across multiple phylogenetic subgroups, our results consistently demonstrate that including higher-resource mainstream languages during fine-tuning does not reliably improve zero-shot ASR performance on dialectal evaluation sets. Instead, whether the fine-tuning data itself reflects dialectal speech emerges as a more informative predictor of performance than phylogenetic proximity alone. These findings highlight the importance of treating dialects as distinct acoustic and linguistic entities rather than as minor variants of standardized languages when designing ASR systems.

We further present the first detailed ASR analysis for Garhwali, a nonstandard Pahari language variety, and show that a w2vBERT-based model fine-tuned on Garhwali achieves the best performance among the evaluated architectures. Although the resulting word error rate of 49.3% remains insufficient for fully automated transcription, this case study illustrates both the challenges of dialectal ASR and the benefits of dialect-specific modeling. Our quantitative error analysis further reveals substantial bias toward Hindi in both multilingual and Hindi-fine-tuned models, manifest-

ing in systematic normalization of dialectal and code-mixed forms, and underscoring the need for dialect-aware data selection and modeling strategies in future ASR systems.

Overall, our findings suggest that effective ASR for low-resource dialects requires moving beyond default assumptions of phylogenetic similarity and toward evaluation and modeling practices that explicitly account for dialectal variation.

Limitations

Our study is based on the VAANI dataset, which contains varying amounts of data across language varieties. Although this diversity allows us to evaluate ASR performance across a wide range of realistic dialectal settings, differences in dataset size may influence performance comparisons across languages. We mitigate this effect where possible by controlling the amount of fine-tuning data used across languages, though some variability remains inherent to the dataset.

In addition, VAANI consists of spontaneous and naturally-occurring speech collected across diverse regions. While this enables evaluation under realistic acoustic and conversational conditions, the presence of background noise, disfluencies, and region-specific recording environments may introduce additional variability in model performance.

Our analysis is limited to Indic dialects and language varieties written in the Devanagari script. Although this choice allows for controlled comparisons within a shared orthographic system, it excludes Indic languages written in other scripts, and our findings may not directly generalize beyond the Devanagari-script subset.

Finally, our experiments focus on a specific set of self-supervised ASR architectures and fine-tuning strategies. While these models are representative of widely used contemporary approaches, different architectures or training objectives may exhibit different transfer behaviors.

References

- Arun Babu, Changan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. [XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale](#). In *Proc. Interspeech*, pages 2278–2282.
- Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. [wav2vec 2.0: A framework for self-supervised learning of speech representations](#). In *Proc. NeurIPS*, volume 33, pages 12449–12460.
- Verena Blaschke, Miriam Winkler, and Barbara Plank. 2025. [Standard-to-dialect transfer trends differ across text and speech: A case study on intent and topic classification in German dialects](#). *Preprint*, arXiv:2510.07890.
- Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. [w2v-BERT: Combining contrastive learning and masked language modeling for self-supervised speech pre-training](#). In *Proc. IEEE ASRU*, pages 244–250.
- Anuj Diwan, Rakesh Vaideeswaran, Sanket Shah, Ankita Singh, Srinivasa Raghavan, Shreya Khare, Vinit Unni, Saurabh Vyas, Akash Rajpuria, Chiranjeevi Yarra, Ashish Mittal, Prasanta Kumar Ghosh, Preethi Jyothi, Kalika Bali, Vivek Seshadri, Sunayana Sitaram, Samarth Bharadwaj, Jai Nanavati, Raoul Nanavati, and Karthik Sankaranarayanan. 2021. [MUCS 2021: Multilingual and code-switching asr challenges for low resource Indian languages](#). In *Proc. Interspeech*, pages 2446–2450.
- Rachana Gusain, Satya Ranjan Dash, Shantipriya Parida, and Girish Nath Jha. 2023. [Automatic language identification: a case study of Pahari languages](#). *Language Resources and Evaluation*, 57:1361–1387.
- Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2024. [Indo-aryan](#). *Glottolog* 5.0.
- Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. [HuBERT: Self-supervised speech representation learning by masked prediction of hidden units](#). *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:34513460.
- Anjali Kantharuban, Ivan Vulić, and Anna Korhonen. 2023. [Quantifying the dialect gap and its correlates across languages](#). In *The 2023 Conference on Empirical Methods in Natural Language Processing*.
- Lachman M. Khubchandani. 1991. [India as a sociolinguistic area](#). *Language Sciences*, 13(2):265–288.
- Gokul Karthik Kumar, V PraveenS., Pratyush Kumar, Mitesh M. Khapra, and Karthik Nandakumar. 2022. [Towards building text-to-speech systems for the next billion users](#). In *Proc. ICASSP*, pages 1–5.
- Saurabh Kumar, Amartyaveer, and Prasanta Kumar Ghosh. 2025. [Jointly Improving Dialect Identification and ASR in Indian Languages using Multimodal Feature Fusion](#). In *Proc. Interspeech*, pages 2770–2774.

- Colin Massica. 1993. *The Indo-Aryan Languages*. Cambridge University Press.
- Iuliia Nigmatulina, Tannon Kew, and Tanja Samardzic. 2020. ASR for non-standardised languages with dialectal variation: the case of Swiss German. In *Proc. VarDial*, pages 15–24.
- Hemant Palivela, Meera Narvekar, David Asirvatham, Shashi Bhushan, Vinay Rishiwal, and Udit Agarwal. 2025. Code-switching ASR for Low-Resource Indic Languages: A Hindi-Marathi case study. *IEEE Access*, 13:9171–9198.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proc. ICML*.
- Mk Riyal, Vinod Khanduri, Nikhil Rajput, and Nagma Irfan. 2016. Creation and analysis of (agriculturally) speech database for Uttarakhand. *Indian Journal of Industrial and Applied Mathematics*, 7:181.
- Harveen Singh Chadha, Anirudh Gupta, Priyanshi Shah, Neeraj Chhimwal, Ankur Dhuriya, Rishabh Gaur, and Vivek Raghavan. 2022. Vakyansh: ASR Toolkit for Low Resource Indic languages. *arXiv e-prints*, arXiv:2203.16512.
- Chihiro Taguchi and David Chiang. 2024. Language complexity and speech recognition accuracy: Orthographic complexity hurts, phonological complexity doesn't. In *Proc. ACL*, pages 15493–15503, Bangkok, Thailand.
- VAANI. 2025. VAANI: Capturing the language landscape for an inclusive digital India (phase 1). <https://vaani.iisc.ac.in/>.

A Appendix

A.1 Baseline Evaluation

We used the IndicWav2Vec-Hindi model to establish a baseline for the performance of off-the-shelf models on the languages in the VAANI dataset. We further divided the results into two tables based on whether the Evaluation Language was seen by the IndicWav2Vec-Hindi model during pretraining or not.

Language	CER	WER
Bhojpuri	0.786	0.732
Chhattisgarhi	0.783	0.757
Hindi	0.752	0.504
Konkani	0.851	0.916
Maithili	0.763	0.696
Marathi	0.821	0.870
Nepali	0.785	0.957
Rajasthani	0.768	0.762

Table 7: IndicWav2Vec-Hindi performance on languages seen during pre-training.

Language	CER	WER	Language	CER	WER
Angika	0.791	0.725	Khortha	0.829	0.743
Awadhi	0.774	0.771	Kumaoni	0.759	0.596
Bajjika	0.784	0.784	Kurukh	0.825	0.875
Bhili	0.860	0.944	Magadhi	0.782	0.700
Bundeli	0.774	0.646	Malvani	0.808	0.881
Garhwali	0.760	0.820	Marwari	0.752	0.740
Gondi	0.718	0.945	Nagpuri	0.797	0.834
Halbi	0.788	0.873	Sadri	0.742	0.738
Haryanvi	0.811	0.766	Surgujia	0.762	0.747
Jaipuri	0.696	0.532	Surjapuri	0.829	0.880
Khariboli	0.784	0.605	Thethi	0.796	0.707

Table 8: IndicWav2Vec-Hindi performance on related but unseen languages.

A.2 Orthographic Analysis

We computed the total number of words (tokens), total number of unique words (types) and the total number of words that only appear once (Hapax) per all the languages in the VAANI dataset. We also compute the total percentage of words that occur only once (Hapax%) and the ratio of unique words to total number of words (TTR) in the table 9.

A.3 Garhwali ASR

We evaluated multiple model architecture varieties on the Garhwali language test split from the VAANI dataset. For each model architecture, we first evaluated the performance without finetuning first and then after finetuning on the Training split of the Garhwali dataset. For each finetuning experiment except for the Whisper model, we generated the vocab of characters present in the training and validation split.

Language	Tokens	Types	Hapax	Hapax %	TTR
Gondi	3,136	1,562	1,168	74.78	0.4981
Thethi	1,556	654	469	71.71	0.4203
Bhili	1,420	632	448	70.89	0.4451
Kurukh	2,022	798	559	70.05	0.3947
Nepali	2,169	776	536	69.07	0.3578
Surjapuri	2,450	860	593	68.95	0.3510
Sadri	6,599	2,096	1,417	67.60	0.3176
Nagpuri	1,571	612	404	66.01	0.3896
Malvani	9,123	2,501	1,636	65.41	0.2741
Konkani	31,668	6,327	4,055	64.09	0.1998
Awadhi	2,346	887	561	63.25	0.3781
Surgujia	5,611	1,524	950	62.34	0.2716
Bundeli	7,734	1,613	969	60.07	0.2086
Khariboli	15,345	2,465	1,476	59.88	0.1606
Angika	34,348	5,420	3,233	59.65	0.1578
Garhwali	77,030	10,431	6,216	59.59	0.1354
Haryanvi	3,747	983	584	59.41	0.2623
Halbi	16,716	3,351	1,979	59.06	0.2005
Marathi	171,424	16,876	9,919	58.78	0.0984
Bajjika	32,049	4,681	2,717	58.04	0.1461
Magadhi	55,891	7,101	4,093	57.64	0.1271
Khortha	38,324	4,499	2,584	57.43	0.1174
Jaipuri	2,074	599	339	56.59	0.2888
Chhattisgarhi	169,861	13,024	7,287	55.95	0.0767
Bhojpuri	209,523	15,243	8,508	55.82	0.0728
Marwari	85,883	8,279	4,582	55.34	0.0964
Kumaoni	22,911	3,190	1,737	54.45	0.1392
Maithili	194,541	14,317	7,779	54.33	0.0736
Hindi	156,389	8,156	4,381	53.72	0.0522
Rajasthani	116,180	7,988	4,273	53.49	0.0688

Table 9: Hapax-Legomena (unique words) per training set.

Model Type	Tr. Lang	FT Lang	Gen Vocab?	Metric	Tr. Error	Val Loss	Tr. Loss	Steps	Test WER	Test CER
wav2vec2CTC	-	Garhwali	Yes	WER	0.769	1.576	0.403	6500	0.769	-
wav2vec2CTC	-	-	Yes	-	-	-	-	-	1.015	2.436
XLS-R	-	Garhwali	Yes	WER	0.735	1.554	0.171	1600	0.650	0.270
XLS-R	-	-	Yes	-	-	-	-	-	1.000	1.292
wav2vec2 BERT	-	Garhwali	Yes	CER	0.197	0.940	0.397	1200	0.493	0.193
wav2vec2 BERT	-	-	Yes	-	-	-	-	-	1.000	1.600
Whisper-small	Hindi	Garhwali	No	CER	0.242	0.974	0.001	4000	0.629	0.650
Whisper-small	Hindi	-	No	-	-	-	-	-	2.518	1.262
HuBERT	-	Garhwali	Yes	CER	-	-	-	-	0.515	0.199

Table 10: Comparison of Speech Models (Garhwali Fine-tuning vs. Baselines)

Ara-HOPE: Human-Centric Post-Editing Evaluation for Dialectal Arabic to Modern Standard Arabic Translation

Abdullah Alabdullah^{*1}, Lifeng Han², Chenghua Lin³,

¹School of Informatics, University of Edinburgh, The United Kingdom

²Leids Universitair Medisch Centrum & LIACS, Universiteit Leiden, NL

³School of Computer Science, University of Manchester, The United Kingdom

Correspondence: a.alabdullah@sms.ed.ac.uk & l.han@lumc.nl, liacs.leidenuniv.nl

Abstract

Dialectal Arabic to Modern Standard Arabic (DA-MSA) translation is a challenging task in Machine Translation (MT) due to significant lexical, syntactic, and semantic divergences between Arabic dialects and MSA. Existing automatic evaluation metrics and general-purpose human evaluation frameworks struggle to capture **dialect-specific MT errors**, hindering progress in translation assessment. This paper introduces *Ara-HOPE*, a human-centric post-editing evaluation framework designed to systematically address these challenges. The framework includes a five-category error taxonomy and a decision-tree annotation protocol. Through comparative evaluation of three MT systems (Arabic-centric Jais, general-purpose GPT-3.5, and baseline NLLB-200), Ara-HOPE effectively highlights systematic performance differences between these systems. Our results show that dialect-specific terminology and semantic preservation remain the most persistent challenges in DA-MSA translation. Ara-HOPE establishes a new framework for evaluating Dialectal Arabic MT quality and provides actionable guidance for improving dialect-aware MT systems. For reproducibility, we make the annotation files and related materials publicly available at <https://github.com/abdullahalabdullah/Ara-HOPE>

1 Introduction

This paper enhances Dialectal Arabic to Modern Standard Arabic (DA-MSA) translation quality assessment through a human-centric evaluation framework, addressing key gaps in current methods for under-resourced language pairs.

Dialectal Arabic (DA) refers to the informal varieties of Arabic used in everyday communication, which vary significantly across regions. In contrast, Modern Standard Arabic (MSA) is the formal

^{*}The previous affiliation for Abdullah Alabdullah (where most of this research was completed) is the University of Manchester, Manchester, M13 9PL, The United Kingdom.

variety used in writing, education, and traditional media (Diab et al., 2010).

Unlike interlingual translation tasks (e.g. translating from English to Arabic), DA-to-MSA translation is a dialect normalization task that introduces additional challenges arising from divergences between Arabic dialects and MSA. Key differences between DA and MSA that affect MT include: (i) **Orthographic differences**: Dialects do not follow standardized spelling rules and the same word may appear in different forms, making automatic text normalization difficult for NLP systems (Alhafni et al., 2024). (ii) **Morphological differences**: Morphology is how words change form to express features such as tense or gender. While MSA has a rich, complex, and standardized morphology, spoken dialects often simplify these systems by omitting rules or using reduced forms (Kirchhoff et al., 2006). (iii) **Lexical differences**: Dialectal vocabulary often includes slang and idiomatic expressions that are typically absent in MSA (Hadj Mohamed et al., 2023). (iv) **Syntactic differences**: Syntax, sentence structure, and word order in many dialects differ from MSA (Bidsy et al., 2009). (v) **Code-switch**: Speakers often mix DA and MSA, and sometimes even foreign words, within a single sentence. This adds more complexity when trying to build systems that automatically translate from DA to MSA (Hamed et al., 2025). These challenges make DA-MSA translation a complex task requiring specialized approaches grounded in a sound understanding of common error types in DA-MSA translation systems.

In this paper we introduce a post-editing human evaluation framework that offers the granularity needed to identify systemic weaknesses in DA-MSA machine translation systems. Unlike general-purpose frameworks, our proposed framework (Ara-HOPE) targets translation errors that result from DA-specific translation challenges.

2 Related Work

2.1 Advancements in Dialectal Arabic Translation

Neural architectures have transformed DA-MSA translation by capturing contextual dependencies and complex syntactic and semantic relationships, effectively modeling dialectal variations and producing more fluent and accurate translations (Baniata et al., 2018). The emergence of large parallel corpora has further advanced Neural Machine Translation by mitigating data scarcity and enabling training on dialectally diverse data. A key example is the MADAR corpus, which contains parallel translations from 25 Arabic city dialects (Bouamor et al., 2018). Recent studies have compared neural architectures, including encoder-decoder models like NLLB-200 and decoder-only models like GPT-3 and GPT-4o (Team et al., 2022; Brown et al., 2020; Alabdullah et al., 2025). Decoder-only models have demonstrated better performance at preserving cultural context (Yakhni and Chehab, 2025).

The emergence of LLMs trained on large and diverse multilingual data, and further optimized through instruction tuning, has enabled models to follow natural-language instructions and generate appropriate outputs for a wide range of tasks, including machine translation, without costly task-specific fine-tuning (Brown et al., 2020). This paradigm is known as In-Context Learning (ICL), where the desired task is specified directly in the prompt, and the model infers the mapping from the provided context. In zero-shot ICL, the model is instructed to translate from a source to a target language without providing any in-prompt examples. Zero-shot prompting has been particularly effective when parallel data is scarce and is widely used in DA-MSA shared tasks such as OSACT (Atwany et al., 2024) and NADI (Abdul-Mageed et al., 2024).

While multilingual LLMs capture general linguistic features through large-scale multilingual pretraining, new Arabic-specialized LLMs like Jais improved the handling of dialectal nuances and cultural references (Sengupta et al., 2023; Mousi et al., 2025), producing more natural and contextually appropriate translations. Despite these advances, neural models continue to underperform on DA translation due to persistent challenges with dialectal nuances and culturally embedded expressions (Mousi et al., 2025; Alabdul-

lah et al., 2025).

2.2 Dialectal Arabic Translation Evaluation

Traditional automatic evaluation metrics such as BLEU (Post, 2018) and METEOR (Banerjee and Lavie, 2005) are limited for DA-MSA translation, as they rely on lexical overlap and perform poorly on morphologically rich languages like Arabic. Bouamor et al. (2014) proposed AL-BLEU, an extension of BLEU that assigns partial credit for stem and morphological matches, yielding better correlation with human judgment than standard metrics. However, AL-BLEU remains a lexical-overlap metric and fails to capture semantic adequacy, particularly an issue for syntactically flexible languages like Arabic. While metrics like BLEU allow evaluation against multiple reference translations, these are challenging to produce. As a result, these metrics favor literal translations with high lexical overlap over contextually appropriate ones that better reflect human judgment, but differ lexically from the reference.

In low-resource settings such as dialectal Arabic, neural evaluation metrics like BERTScore (Zhang et al., 2019) and COMET (Rei et al., 2020) also face challenges. Falcão et al. (2024) showed that COMET’s performance on under-resourced languages is constrained by imbalanced training data. These metrics rely on pretrained models which typically lack sufficient dialectal Arabic training data, leading to lower-quality embeddings.

Human evaluation frameworks such as the Multidimensional Quality Metrics (MQM) assess fine-grained translation errors like omissions and register mismatches (Lommel et al., 2013, 2024). While MQM allows for more precise diagnostic error analysis due to its detailed error taxonomy, this comes at the cost of increased complexity in the annotation framework, requiring extensive annotator training, which can be costly and difficult to achieve for low-resource languages (Kocmi et al., 2024). Moreover, DA translation requires a specialized human evaluation framework that captures the most impactful error types and minimizes subjectivity in assessing translation quality for this specific task, with only minimally trained annotators. To further advance research in this direction, we build our methodology on the HOPE metric (Gladkoff and Han, 2022). HOPE is a task-oriented evaluation framework designed to address the limitations of both automatic evaluation metrics (e.g. BLEU) and highly fine-grained but complex hu-

man evaluation frameworks such as MQM. Ara-HOPE is human-centric because it only incorporates eight evaluation criteria that capture the most critical and recurring errors in translation between the Levantine Arabic dialect and MSA, reflecting translation quality as perceived by native speakers. The assigned error scores correspond to the post-editing effort required to bring the translation to an acceptable quality.

3 Methodology Design

3.1 Framework Development

Ara-HOPE Error Taxonomy Design: Developing a robust human evaluation framework requires clearly defined objectives. For DA-MSA translation, this requires an error taxonomy tailored to the specific challenges of this language pair. While the HOPE framework (Gladkoff and Han, 2022) offers a foundation for general post-editing translation assessment, we adapt the general error types in HOPE to capture the specific challenges of DA-MSA translation. To minimize subjective judgment between annotators, we also reduce the severity scoring range to three levels: from 0 (no errors) to 2 (major error). Our proposed taxonomy is designed to evaluate translation quality, identify system weaknesses, and guide system improvements, while remaining usable by native DA speakers without requiring extensive annotation training. To achieve this, we employ a direct quality estimation approach that evaluates predefined aspects of translation quality (e.g. fluency) at the segment level using discrete severity scores. This design makes our framework easier to train annotators on and faster to apply than MQM.

Design Principles: To ensure a theoretically sound and practical taxonomy, we developed guidelines grounded in best practices in translation quality assessment (Han, 2020; Rivera-Trigueros, 2022) and refined them through iterative feedback from focus groups with native DA speakers. The finalized guidelines comprise six core principles: (1) **Identifiability:** Errors must be detectable by minimally trained native dialect speakers to ensure consistent evaluation. (2) **Distinguishability:** Categories should have minimal conceptual overlap, with clear definitions to avoid confusion. (3) **Actionability:** Error classification must enable targeted system improvements through quantifiable error severity and clear mapping to issues in the translation system. (4) **Comprehensiveness:** The

taxonomy must capture all major errors common in DA-MSA MT, including dialect-specific challenges. (5) **Relevance:** Categories must address issues unique to DA-MSA translation rather than pure general translation errors. (6) **Usability:** The taxonomy should remain manageable in size, balancing high-level categories with sufficient granularity.

Taxonomy Structure: The Ara-HOPE taxonomy categorizes DA-MSA translation errors into three hierarchically structured classes: (1) **Fluency Error (FLU):** Grammatical or linguistic errors in the MSA translation, independent of the DA source sentence. (2) **Meaning Transfer Error:** Failures to accurately preserve source meaning. This category includes Proper Name (PRN) errors, referring to incorrect translations of names of people, places, or organizations; Dialect-Specific Term (TRM) errors, which involve untranslated or mistranslated dialectal expressions that alter meaning; and General Semantic Mistranslation (GSMIS) errors, which includes omissions, additions, or other semantic changes. (3) **Adaptation Error (ADP):** Translations that are unnatural or contextually inappropriate in tone, style, or intent.

GSMIS captures global meaning changes arising from the model’s inability to handle DA’s contextual dependencies and includes all meaning changes not covered by PRN or TRM. This category was introduced after observing that LLMs often produce semantic distortions beyond proper name or dialect-specific term mistranslations.

Decision Tree Implementation While a list format effectively outlined the taxonomy, it was later reformatted as a decision tree to reduce cognitive load for our minimally trained evaluators. The decision tree guides annotators through hierarchical error categories, as shown in Figure 5 in the Appendix. An English translated version of the tree is provided in Figure 6 in the Appendix.

This structure simplifies annotation by providing step-by-step guidance, helping evaluators understand the evaluation process, handle multiple error types systematically, and make consistent judgment at each level. The tree begins with the three primary categories (Fluency, Meaning Transfer, and Adaptation) and expands into more granular subcategories. Each node presents a yes/no question to minimize subjective judgments. Since some error types (e.g. Dialect-Specific Term) can be challenging to distinguish from other categories, we provided annotators with a table of annotation

guidelines with illustrative examples to clarify the distinctions between error types. The guidelines can be found in Figure 7 in the Appendix.

Practical initial testing demonstrated Ara-HOPE’s effectiveness in addressing dialect-specific translation challenges and improving annotation consistency. The decision-tree structure also proved especially effective in guiding evaluators and enhancing usability.

3.2 Dataset Preparation

Our human evaluation experiment utilized 200 tweets from the Levantine development set of the Dial2MSA-Verified dataset (Khered et al., 2025), a high-quality parallel corpus for DA-MSA translation in the social media domain. It extends the original Dial2MSA dataset (Mubarak, 2018) and applies automated corrections and human evaluation by native speakers to produce reliable MSA references.

This tweets dataset was chosen for its comprehensive representation of DA-MSA translation challenges, including: **Lexical variations:** colloquial and multi-word expressions. **Semantic shifts:** cultural references and context-dependent meaning. **Orthographic variations:** common in Levantine social media text.

The tweets in this dataset span a wide range of tones, from casual tweets to heated discourse, ensuring thorough testing of MT systems’ ability to handle DA translation, while preserving sentiment. Moreover, the dataset allows for testing Ara-HOPE’s capacity to capture nuanced dialect-specific translation errors.

Our 200-example subset size exceeds the lower bound below which translation quality estimates become unreliable, addressing concerns that a very small sample may lead to unreliable error analysis (Gladkoff et al., 2022).

3.3 Translation Systems Selection

For this human evaluation experiment, we selected three translation systems using three predefined criteria, which ensure a robust comparative analysis of DA-MSA translation performance: (i) **Reliability and Proven Performance:** Each system has demonstrated strong performance on DA-MSA translation in prior work. (ii) **Architectural Diversity:** The systems represent distinct model families (encoder-decoder vs. decoder-only) and differ in pretraining data sources, encouraging diverse outputs and reducing redundancy. (iii) **Performance**

Scope: We include both advanced and baseline systems to contrast general-purpose and Arabic-specialized approaches.

The Selected Systems Are: **Jais:** A state-of-the-art Arabic-centric model trained with approximately one-third of its training tokens drawn from Arabic data, excelling at capturing DA-MSA nuances (Sengupta et al., 2023). **GPT-3.5:** A general-purpose multilingual LLM. We will use this model to benchmark multilingual systems against an Arabic-specialized model. This model was introduced as an improvement over its predecessor, GPT-3 (Brown et al., 2020). **NLLB-200 3.3B:** A multilingual baseline MT system, providing contrast and revealing errors that more advanced systems may avoid (Team et al., 2022).

Each of the three systems was prompted to translate the same 200 examples. A zero-shot instruction setup was used to fairly evaluate each model’s baseline capabilities without fine-tuning or advanced prompting configuration. Jais and GPT-3.5 received consistent prompts (in Arabic for Jais and English for GPT-3.5) while NLLB-200 required no prompting due to its built-in MT task support.

4 Implementation

4.1 Annotator Selection and Training

Two native speakers of Syrian Levantine Arabic with advanced proficiency in MSA volunteered as annotators. Both hold undergraduate degrees in Arabic Language, enabling them to identify violations of MSA grammatical rules. Recruiting two annotators allowed for inter-annotator agreement (IAA) analysis while keeping the annotators group size manageable.

Annotators received initial training through a 25-minute pre-recorded video and supplementary materials. Training covered: (1) an **Error Taxonomy Guide:** providing detailed definitions of error types, alongside diverse examples of correct and incorrect DA-MSA translations, and (2) a **Decision Tree Workflow:** annotators first read the source and gold translation to establish the intended meaning, then highlighted errors by severity starting with the most critical, and finally used the decision tree (Figure 6 in the Appendix) to classify errors systematically.

Offline support was available throughout the annotation period. Each annotator spent approximately 12 hours on their tasks, supported by multi-

ple feedback sessions to ensure clarity and consistency.

4.2 Pilot Testing and Framework Refinement

Before full-scale evaluation, a pilot study was conducted with one annotator annotating 10 DA-MSA examples. The primary goal was to test the framework’s usability and effectiveness. The feedback received focused on: (1) **Instruction Clarity**: The annotator initially struggled to distinguish between meaning transfer and adaptation errors. The definitions were revised accordingly. (2) **Questionnaire Usability**: The layout was changed to ensure ease of use and reduce cognitive load. (3) **Severity Scale Adjustments**: Confusing mid-range scores were removed, leading to refinement of the scoring scale.

The pilot test confirmed the framework’s overall effectiveness and identified minor improvements needed to enhance clarity and usability. Revisions ensured annotators could complete evaluation efficiently without sacrificing accuracy.

4.3 Questionnaire Design

A structured Excel sheet was created to support annotation and data analysis. Each sheet contained three sets of columns, one per translation engine. Each annotator evaluated 600 translations in total (200 for each translation engine). Figure 1 shows two annotated examples for the Jais system. The first three columns show the DA sentence, the MSA gold translation, and the proposed machine translation. Columns 4-8 represent the five Ara-HOPE error categories, with annotators assigning severity scores from 0 (no error) to 2 (major error). Empty cells indicate a score of 0, and the final column sums the error scores for each sentence.

Annotators assessed adaptation errors only when no meaning transfer errors (PRN, TRM, GSMIS) were present, as style or tone evaluation is irrelevant when meaning is lost. To prevent over-penalizing systems that preserved meaning but erred in adaptation, adaptation scores were weighted at 50% in the segment total error score (SEGS) calculation.

Three physical copies of the Excel sheet were printed (one for each system) as annotators preferred working on paper. Each contained the same DA-MSA pairs, differing only in the translation engine output. Annotators marked scores manually, and data were later transferred to the digital Excel template.

5 Results Analysis

5.1 Inter Annotator Agreement

Inter-Annotator Agreement (IAA) measures consistency between evaluators, with high scores indicating a reliable and reproducible annotation framework. In our study, we employed Quadratic Weighted Cohen’s Kappa (QWK) because it effectively handles ordinal severity ratings by assigning greater penalties to larger disagreements, whereas, the ordinary Cohen’s κ treats all disagreements equally, unweighted κ , e.g. 0 vs 1 = 0 vs 2. In addition, quadratic weights are standard for ordinal ratings (Cohen, 1968). This makes it particularly suitable for error severity judgment on our 0-2 scale, where the magnitude of disagreement matters. The IAA scores are presented in Table 1.

As shown in Figure 6, Meaning Transfer comprises three fine-grained error types (PRN, TRM, and GSMIS). Accordingly, for Meaning Transfer IAA, we compute inter-annotator agreement on an aggregated severity score obtained by summing PRN, TRM, and GSMIS. This reflects annotator agreement on overall meaning-transfer impact rather than on individual sub-error categories. Similarly, SEGS represents the segment-level total error score (produced by summing all five error types’ scores), and the corresponding IAA reflects agreement on overall segment-level error severity.

The IAA scores varied meaningfully across error categories: Fluency showed strong agreement for most MT systems, likely due to annotators’ Arabic language expertise, which minimized inaccurate judgments. Meaning Transfer achieved robust agreement, likely due to the clear criteria regarding preservation of original meaning, as established in the Ara-HOPE taxonomy decision tree (Figure 6). Those definitions were particularly refined before the full evaluation during our pilot testing and feedback sessions. This consistency indicates a shared understanding of meaning preservation requirements among our annotators. Adaptation yielded the lowest agreement (0.122-0.28) due to its inherently subjective nature.

A translation has an adaptation error when it seems unnatural or inappropriate to the context, situation, or intent of the source. Unlike fluency or meaning-transfer errors, which can often be localized to specific parts of the translation, adaptation errors require more holistic, and usually subjective judgment. One example is the dialectal sentence:

	A	B	C	D	E	F	G	H	I
	DA	MSA	Jais	FLU	PRN	TRM	GSMIS	ADP	SEGS
1									
2	كل عام وانت متواضع مثل هيك بابوقوص	كل عام وانت متواضع هكذا يا ابو قوص	كل عام وانت بخير يا ابو قوص.				2.00		2.00
3	تسلم والله قدام عالقل استفيد من خبرتك في سوائف التسجيل و هيك	سلمك الله في المستقبل انشاء الله ساستفيد من خبرتك في مواضيع التسجيل وهكذا	شكرا جزيلا لك لاني ساستفيد بالتاكيد من خبراتك في شؤون التسجيل وما شابه ذلك.					1.00	0.50

Figure 1: Example layout of the annotation questionnaire used for evaluating DA-MSA translations across fluency, meaning transfer, and adaptation error categories. To avoid unfairly penalizing systems that preserved meaning but made adaptation errors, adaptation scores were weighted half in the segment total error score calculation.

عم بحكي من واقع تجربة كانت رح تقضي ع حلم حياتي بس الحمد لله الحب اقوى ("I am speaking from experience. It almost destroyed my life's dream, but thank God, love is stronger"), which the Jais model translated as

أنا أحدث بناءً على تجربتي الشخصية التي كادت تدمر حلمي الأكبر، ولكن بحمد الله، أثبت الحب أنه أقوى بكثير. ("I speak from my personal experience, which almost destroyed my biggest dream, but thank God, love proved to be much stronger.")

While both annotators agreed that the translation is fluent and preserves the source meaning, one judged it to be slightly less natural in MSA (ADP = 1, minor adaptation error), whereas the other annotator found it acceptable and contextually appropriate.

The annotator disagreement likely stems from the rendering of phrases such as بس الحمد لله الحب اقوى ("But thank God, love is stronger") which was translated as ولكن بحمد الله، أثبت الحب أنه أقوى بكثير ("But thank God, love proved to be much stronger"), which may be seen as stylistically over-emphatic or slightly less natural in MSA for this context.

Error Type	Jias	GPT3.5	NLLB200
Fluency	0.507	0.552	0.368
Meaning Transfer	0.529	0.629	0.554
Adaptation	0.171	0.122	0.280
SEGS	0.608	0.629	0.500

Table 1: Quadratic Weighted Kappa (QWK) scores across models and error types. Meaning Transfer IAA is computed on aggregated PRN+TRM+GSMIS severity; SEGS reflects total segment-level error severity.

The IAA scores for the three systems in our study on the segment-level total error score (SEGS) was 0.5 to 0.629, indicating reasonably consistent human evaluation.

Prior work, including (Landis and Koch, 1977), considers kappa scores in the range of 0.41-0.60 to indicate a moderate level of agreement and 0.61-

0.80 to indicate substantial agreement. Nevertheless, it is important to note that the interpretation of standard Kappa and Quadratic Weighted Kappa varies considerably.

Aggregating fine-grained error categories into a composite severity score reduces sparsity and stabilizes the marginal distributions, which is known to improve reliability of composite measures (Fleiss et al., 2003). Moreover, Quadratic Weighted Kappa assigns smaller penalties to near disagreements (Cohen, 1968), so cross-category disagreements often translate into small ordinal differences after aggregation. Consequently, agreement computed on aggregated segment-level scores can exceed that of individual categories.

5.2 Quantitative Error Analysis

The quantitative analysis below presents a comparison of MT system performance, highlighting differences across key evaluation criteria. This assessment identifies each system's strengths and weaknesses, as well as both unique and shared challenges across systems.

Error Severity Analysis This analysis examines translation error severity for 205 sentences produced by the three MT systems using the accumulated sentence total error scores (SEGS). It is essential for understanding how each model handles DA-MSA translation and for identifying systems that generate higher-quality outputs that require minimal post-editing.

SEGS is calculated as the sum of the scores assigned to each of the five error types per sentence, yielding a range of 0-4 per sentence. For comparison, SEGS values were grouped as follows: segments requiring no editing (SEGS = 0), segments with minor errors ($0.5 < \text{SEGS} \leq 1$), and segments with major errors ($\text{SEGS} > 1$).

Figure 2 shows clear differences among the systems. For Jais, 36% of segments required no editing, 34% had minor errors, and 30% had major er-

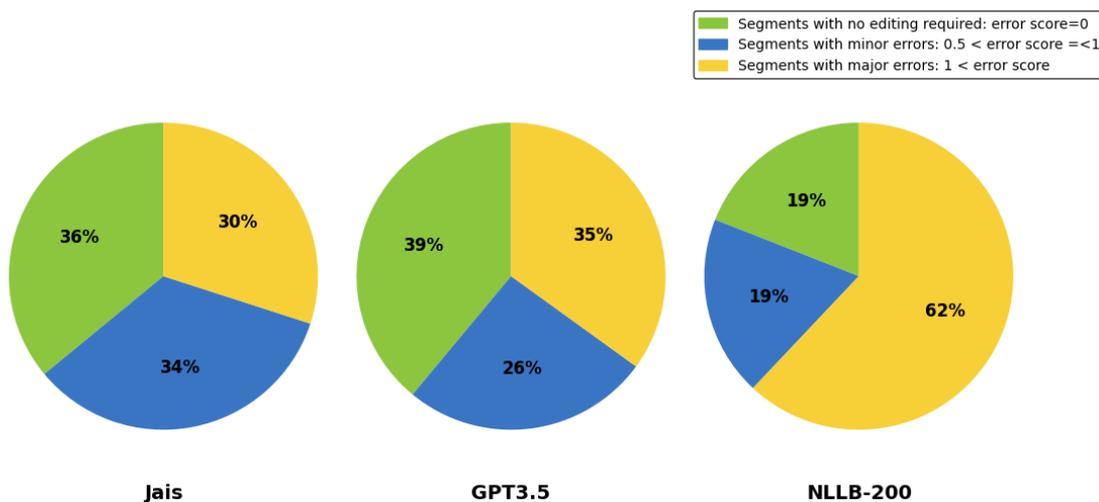


Figure 2: Comparison of error severity distribution among Jais, GPT-3.5, and NLLB-200, highlighting proportions of major, minor, and no-error translations.

rors. GPT-3.5 performed slightly better, with 39% of segments requiring no editing, fewer minor errors (26%), but slightly more major errors (35%). NLLB-200 performed the worst, with only 19% of segments requiring no editing and 62% containing major errors.

These results reflect differences in model architecture and training. Jais, as an Arabic-centric model, handles DA-specific nuances well but still struggles with complex segments. GPT-3.5’s general-purpose design provides balanced performance but lacks Arabic specialization. The strong performance of Jais and GPT-3.5 underscores the relative strength of decoder-only models over encoder-decoder models like NLLB-200.

Error Pattern Examining the error distributions in Figure 3 provides key insights into each system’s approach to DA-MSA translation challenges. NLLB-200 exhibits a high frequency of dialect-specific term (TRM) and semantic preservation (GSMIS) errors, highlighting its encoder-decoder architecture’s difficulty with idiomatic expressions, cultural references, and context-dependent meanings specific to Levantine Arabic. These issues stem from limited domain-specific tuning and insufficient exposure to regional culturally embedded expressions, which are essential for accurate meaning transfer.

Jais, despite its Arabic-specialized pretraining, struggles with proper names (PRN). This might be because the system tends to over-normalize or excessively standardize names transliterated in social media DA content, leading to the loss of original or

intended forms. However, its strong performance with dialect-specific terminology reflects effective handling of DA morphological variations, likely due to its extensive pretraining on Arabic data.

GPT-3.5 shows a balanced error distribution across many categories, suggesting that its large multilingual pretraining and sheer parameter size help offset the lack of explicit Arabic dialectal training. This is particularly evident in challenging cases involving code-mixing (where words from two languages or dialects are used in the same sentence) and pragmatic shifts, which require interpreting meaning based on context, tone, or speaker intent rather than literal words. These cases require advanced comprehension, and the model performs well in them even though it was not specifically trained on Arabic dialects.

The consistently low fluency (FLU) error rates across systems indicate that syntactic reconstruction from DA to MSA is less challenging than lexical-semantic transfer¹. Low adaptation (ADP) error rates across all systems are largely due to this error type being assessed only when meaning transfer errors (PRN, TRM, GSMIS) are absent. When a text fails to convey its intended meaning, evaluating its stylistic or cultural appropriateness becomes less relevant.

A complementary view is provided in Figure 4, which shows the exact error distribution for each system. The total error scores for all sentences

¹Lexical-semantic transfer refers to the mapping of words and their intended meaning from one dialect or language variety to another.

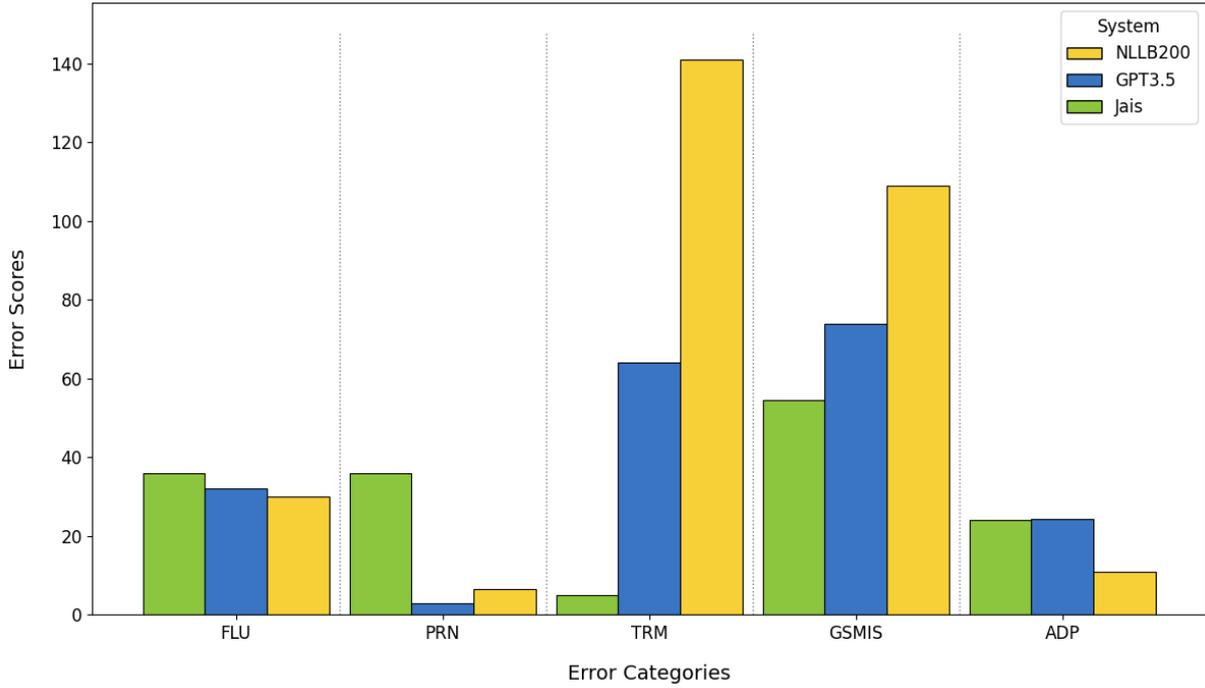


Figure 3: Comparison of Jais, GPT-3.5, and NLLB-200 models’ error scores across error categories

(SEGS). It is clear that NLLB-200 has a significantly higher total error score (297.50) compared to GPT-3.5 (196.25) and Jais (187.50), with Jais producing the least meaning transfer errors. Notably, TRM and GSMIS errors constitute the majority of total errors across all systems, indicating that future MT development for DA-MSA should prioritize improvements in semantic accuracy and dialect-specific terminology handling.

5.3 Qualitative Error Analysis

This qualitative analysis offers concrete examples of two common error types to illustrate how translation quality varies across MT systems, focusing on challenging linguistic phenomena.

5.3.1 Analysis of Dialectal Terminology Errors

Addressing dialect-specific terms (TRM) is a core challenge in DA-MSA translation due to vocabulary mismatches between informal dialects and formal Arabic. Table 2 presents a sentence containing a unique dialect-specific term. The word *بيجن* (“amazing” in Syrian Arabic) lacks a direct MSA equivalent, requiring systems to interpret its meaning contextually instead of relying on word-for-word mapping. Unlike NLLB-200, Jais, with its Arabic-centric pretraining, was better at recognizing this regional idioms. GPT-3.5, though not

Source DA: اليوم الجو ييجن، لازم نطلع نتمشى
Gold MSA: اليوم الطقس رائع، يجب أن نخرج للتنزه “The weather is wonderful today; we should go out for a walk.”
Jais: اليوم الطقس رائع، يجب أن نذهب للتنزه “The weather is wonderful today; we should go for a walk.”
GPT-3.5: اليوم الطقس ممتاز، يجب أن نخرج للتنزه “The weather is excellent today; we should go out for a walk.”
NLLB-200: اليوم الطقس مجنون، يجب أن نخرج “Today’s weather is crazy; we should go out for a walk.”

Table 2: Evaluation of how models handle dialect-specific terms (TRM), highlighting translation challenges in translating informal expressions to MSA.

DA-specialized, leverages its broad language understanding to connect dialect terms to contextually appropriate MSA words (translating *بيجن* as ممتاز “excellent”). These findings demonstrate that success with TRM errors depends heavily on exposure to diverse, dialect-rich data that treats dialectal phrases as meaningful units to be interpreted contextually, rather than treating them as isolated words.

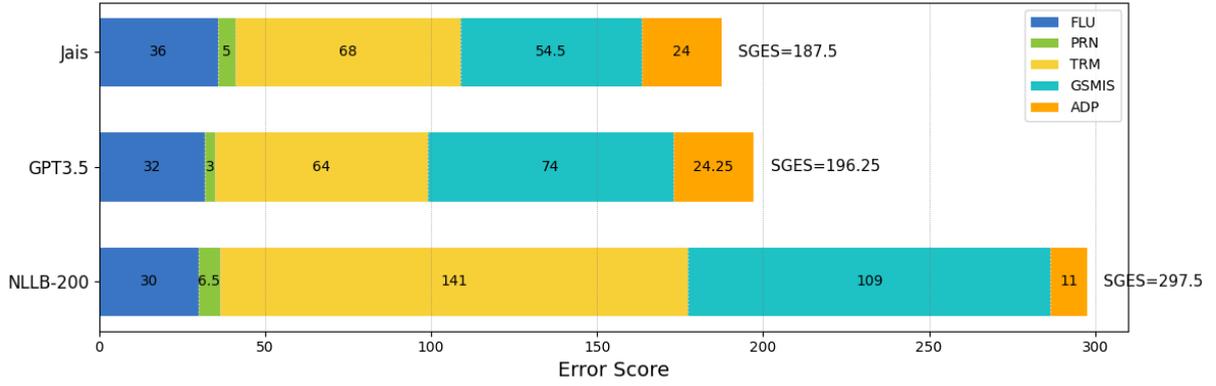


Figure 4: Visualization of accumulated error scores across fluency, meaning transfer, and adaptation error categories for Jais, GPT-3.5, and NLLB-200 in DA-MSA translation.

Source DA: شو قصتك ليش معصب
Gold MSA: ما الأمر؟ لماذا تبدو غاضباً؟ “What’s wrong? Why do you look angry?”
Jais: ما الخطب؟ لماذا تبدو غاضباً “What’s the matter? Why do you look angry?”
GPT-3.5: ماذا حدث؟ لماذا تبدو غاضباً؟ “What happened? Why do you look angry?”
NLLB-200: ما هي قصتك؟ لماذا أنت غاضب؟ “What is your story? Why are you angry?”

Table 3: Comparison of model performance on Adaptation (ADP) errors.

5.3.2 Analysis of Adaptation Errors

Table 3 presents an example illustrating how the different systems preserved the intent and tone of the source when translating to MSA. ADP errors highlight the gap between DA’s context-dependent expressions and MSA’s formality. For instance, translating شو قصتك (casual “What’s wrong?”) requires capturing the speaker’s intent rather than just the literal words. NLLB-200 translates this as ما هي قصتك (“What is your story?”), focusing on a literal interpretation. Jais, benefiting from exposure to both dialect and MSA data, uses the conventional MSA phrase ما الخطب (“What’s the matter?”). GPT-3.5 chooses the phrasing ماذا حدث (“What happened?”). These differences show that reducing ADP errors requires systems to prioritize intent and context over lexical mapping, treating DA as a distinct communication style with its own contextual rules rather than simply a variation of MSA.

Overall, our analyses emphasize that effective DA-MSA translation depends on training strate-

gies that prioritize (1) comprehension of dialect phrases and (2) preservation of speaker intent across registers. The qualitative analysis above shows that low-resource MT is a creative task, involving sub-tasks like sentiment analysis and formality adaptation, which go beyond simple lexical mapping.

6 Conclusion

This paper introduces the *Ara-HOPE* framework as a human-centric approach for evaluating DA-MSA translation, successfully fulfilling its intended objectives through a specialized error taxonomy, an efficient annotation workflow, and a comparative evaluation of different MT systems. The *five-category error classification* system effectively captures translation challenges unique to DA, while the *decision tree protocol* improves annotation consistency. Quantitative findings reveal significant differences in performance among Arabic-centric (Jais), general-purpose (GPT-3.5), and baseline (NLLB-200) systems, with dialect-specific terminology and semantic preservation identified as key challenges. By systematically addressing the complexities of DA-MSA translation assessment through rigorous human evaluation, Ara-HOPE establishes reproducible standards for Arabic MT assessment and provides actionable insights to guide future MT systems development².

²This work is in line with our DA-MSA MT work at (Al-[abdullah et al., 2025](#)) where we examined LLM prompting vs finetuning for Levantine, Egyptian, and Gulf dialects to MSA translation.

7 Limitations

In this work, we only used zero-shot prompting to generate translations. Future research could explore human evaluation of translations produced using alternative prompting strategies, such as few-shot or chain-of-thought prompting. Additionally, the human annotation process was time-consuming. Future work could consider using LLM-as-a-judge approaches to partially or fully automate the annotation process. Our work focused on human evaluation, and we did not investigate the correlation between human judgment and automatic evaluation metrics, lexicon-based and neural-embedding based, like BLEU and BERTScore. We leave that for future work.

Acknowledgment

We would like to thank our volunteer human annotators, Heba Sado Jazea and Rania Mohammad Daher, for their invaluable time and effort throughout the annotation process. Their careful annotations and insightful feedback were essential to establishing our Ara-HOPE framework. We also thank the VarDial reviewers for the valuable comments.

References

- Muhammad Abdul-Mageed, Amr Keleg, AbdelRahim Elmadany, Chiyu Zhang, Injy Hamed, Walid Magdy, Houda Bouamor, and Nizar Habash. 2024. [NADI 2024: The fifth nuanced Arabic dialect identification shared task](#). In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 709–728, Bangkok, Thailand. Association for Computational Linguistics.
- Abdullah Alabdullah, Lifeng Han, and Chenghua Lin. 2025. [Advancing dialectal arabic to modern standard arabic machine translation](#). *Preprint*, arXiv:2507.20301.
- Bashar Alhafni, Sarah Al-Towaity, Ziyad Fawzy, Fatema Nassar, Fadhl Eryani, Houda Bouamor, and Nizar Habash. 2024. [Exploiting dialect identification in automatic dialectal text normalization](#). In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 42–54, Bangkok, Thailand. Association for Computational Linguistics.
- Hanin Atwany, Nour Rabih, Ibrahim Mohammed, Abdul Waheed, and Bhiksha Raj. 2024. [OSACT 2024 task 2: Arabic dialect to MSA translation](#). In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation @ LREC-COLING 2024*, pages 98–103, Torino, Italia. ELRA and ICCL.
- Satanjeev Banerjee and Alon Lavie. 2005. [METEOR: An automatic metric for MT evaluation with improved correlation with human judgments](#). In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Laith H. Baniata, Seyoung Park, and Seong-Bae Park. 2018. [A neural machine translation model for arabic dialects that utilises multitask learning \(mtl\)](#). *Computational Intelligence and Neuroscience*, 2018. Publisher Copyright: © 2018 Laith H. Baniata et al.
- Fadi Biadisy, Julia Hirschberg, and Nizar Habash. 2009. [Spoken arabic dialect identification using phonotactic modeling](#). In *Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages*, Semitic '09, page 53–61, USA. Association for Computational Linguistics.
- Houda Bouamor, Hanan Alshikhabobakr, Behrang Mohit, and Kemal Oflazer. 2014. [A human judgement corpus and a metric for arabic mt evaluation](#). In *Conference on Empirical Methods in Natural Language Processing*.
- Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghrouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. [The madar arabic dialect corpus and lexicon](#). In *International Conference on Language Resources and Evaluation*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Jacob Cohen. 1968. [Weighted kappa: Nominal scale agreement with provision for scaled disagreement](#). *Psychological Bulletin*, 70(4):213–220.
- Mona Diab, Nizar Habash, Owen Rambow, Mohamed Altantawy, and Yassine Benajiba. 2010. [Colaba: Arabic dialect annotation and processing](#). In *Lrec workshop on semitic language processing*, pages 66–74.
- Júlia Falcão, Claudia Borg, Nora Aranberri, and Kurt Abela. 2024. [COMET for low-resource machine translation evaluation: A case study of English-Maltese and Spanish-Basque](#). In *Proceedings of*

- the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 3553–3565, Torino, Italia. ELRA and ICCL.
- Joseph L. Fleiss, Bruce Levin, and Myunghee Cho Paik. 2003. *Statistical Methods for Rates and Proportions*. Wiley.
- Serge Gladkoff and Lifeng Han. 2022. **HOPE: A task-oriented and human-centric evaluation framework using professional post-editing towards more effective MT evaluation**. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 13–21, Marseille, France. European Language Resources Association.
- Serge Gladkoff, Irina Sorokina, Lifeng Han, and Alexandra Alekseeva. 2022. **Measuring uncertainty in translation quality evaluation (TQE)**. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1454–1461, Marseille, France. European Language Resources Association.
- Najet Hadj Mohamed, Malak Rassem, Lifeng Han, and Goran Nenadic. 2023. **AlphaMWE-Arabic: Arabic edition of multilingual parallel corpora with multiword expression annotations**. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 448–457, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.
- Injy Hamed, Caroline Sabty, Slim Abdennadher, Ngoc Thang Vu, Tamar Solorio, and Nizar Habash. 2025. **A survey of code-switched Arabic NLP: Progress, challenges, and future directions**. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4561–4585, Abu Dhabi, UAE. Association for Computational Linguistics.
- Chao Han. 2020. **Translation quality assessment: a critical methodological review**. *The Translator*, 26(3):257–273.
- Abdullah Khered, Youcef Benkhedda, and Riza Batista-Navarro. 2025. **Dial2MSA-verified: A multi-dialect Arabic social media dataset for neural machine translation to Modern Standard Arabic**. In *Proceedings of the 4th Workshop on Arabic Corpus Linguistics (WACL-4)*, pages 50–62, Abu Dhabi, UAE. Association for Computational Linguistics.
- Katrin Kirchhoff, Dimitra Vergyri, Jeff Bilmes, Kevin Duh, and Andreas Stolcke. 2006. **Morphology-based language modeling for conversational arabic speech recognition**. *Computer Speech & Language*, 20(4):589–608.
- Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. **Error span annotation: A balanced approach for human evaluation of machine translation**. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453, Miami, Florida, USA. Association for Computational Linguistics.
- J. Richard Landis and Gary G. Koch. 1977. **The measurement of observer agreement for categorical data**. *Biometrics*, 33(1):159–174.
- Arle Lommel, Serge Gladkoff, Alan Melby, Sue Ellen Wright, Ingemar Strandvik, Katerina Gasova, Angelika Vaasa, Andy Benzo, Romina Marazato Sparano, Monica Foresi, Johani Innis, Lifeng Han, and Goran Nenadic. 2024. **The multi-range theory of translation quality measurement: MQM scoring models and statistical quality control**. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 2: Presentations)*, pages 75–94, Chicago, USA. Association for Machine Translation in the Americas.
- Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2013. **Multidimensional quality metrics: a flexible system for assessing translation quality**. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.
- Basel Mousi, Nadir Durrani, Fatema Ahmad, Md. Arid Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur Absar Chowdhury, and Firoj Alam. 2025. **AraDiCE: Benchmarks for dialectal and cultural capabilities in LLMs**. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4186–4218, Abu Dhabi, UAE. Association for Computational Linguistics.
- Hamdy Mubarak. 2018. **Dial2msa: A tweets corpus for converting dialectal arabic to modern standard arabic**. *OSACT*, 3:49.
- Matt Post. 2018. **A call for clarity in reporting BLEU scores**. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. **COMET: A neural framework for MT evaluation**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.
- Irene Rivera-Trigueros. 2022. **Machine translation systems and quality assessment: a systematic review**. *Lang. Resour. Eval.*, 56(2):593–619.
- Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Xudong Han, Sondos Mahmoud Bsharat, and 13 others. 2023. **Jais and jais-chat: Arabic-centric foundation**

and instruction-tuned open generative large language models. *Preprint*, arXiv:2308.16149.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. [No language left behind: Scaling human-centered machine translation](#). *Preprint*, arXiv:2207.04672.

Silvana Yakhni and Ali Chehab. 2025. [Can LLMs translate cultural nuance in dialects? a case study on Lebanese Arabic](#). In *Proceedings of the 1st Workshop on NLP for Languages Using Arabic Script*, pages 114–135, Abu Dhabi, UAE. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. [BERTScore: Evaluating text generation with BERT](#). *arXiv preprint arXiv:1904.09675*.

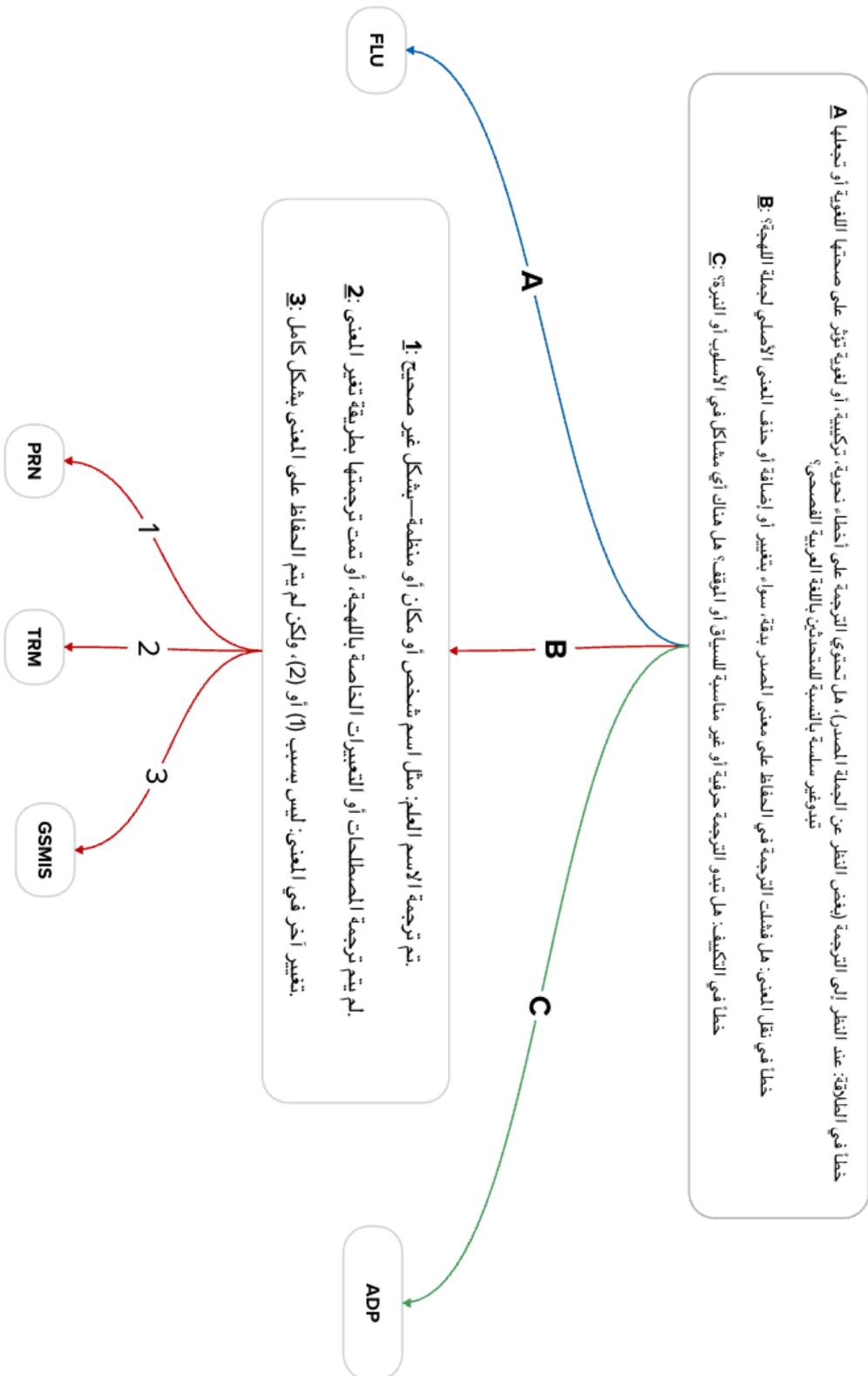


Figure 5: The Arabic version of the Ara-HOPE Annotation Decision Tree. A structured decision tree guiding annotators through error classification for evaluating DA-MSA translations using the Ara-HOPE framework.

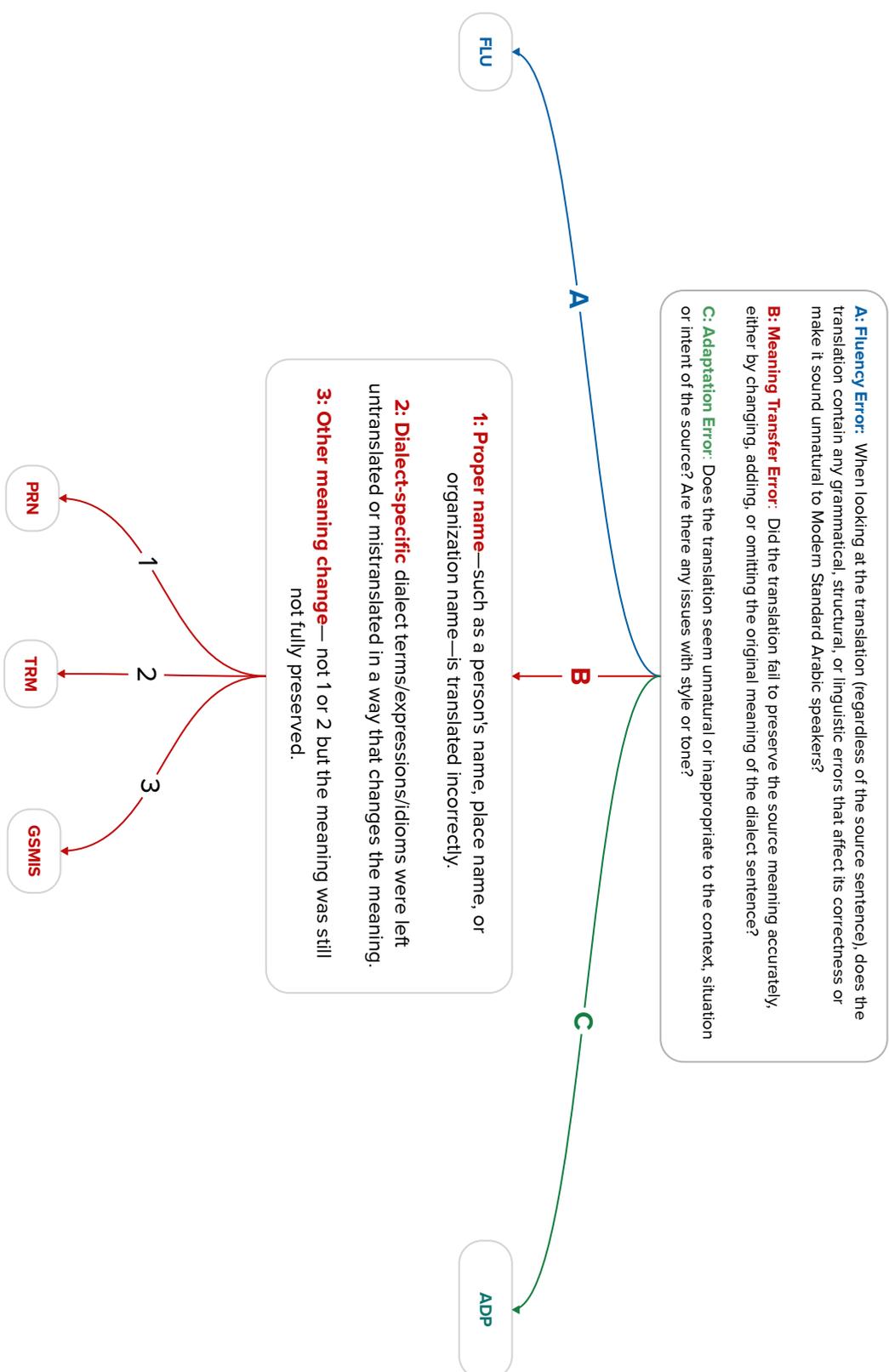


Figure 6: Structured decision tree guiding annotators through error classification for evaluating DA-MSA translations using the Ara-HOPE framework.

الرمز	نوع الخطأ	شرح ومثال
FLU	خطأ في الطلاقة	<p>تشير الطلاقة إلى الجودة اللغوية للترجمة، مع التركيز على الصحة النحوية، والبنية التركيبية السليمة للجمل بعد الترجمة بشكل عام. تكون الترجمة الطليقة: سلسلة وسهلة الفهم للناطقين بالعربية الفصحى.</p> <p>⚠️ ركز فقط على الجملة المترجمة: تجاهل الجملة الأصلية باللهجة.</p> <p>يحدث خطأ في الطلاقة إذا كانت الترجمة:</p> <ul style="list-style-type: none"> تحتوي على أخطاء نحوية أو قواعدية في الأسماء أو الأفعال أو أي أخطاء أخرى. تحتوي على خطأ في تركيبها. تتضمن أخطاء إملائية. تستخدم عبارات غير مألوفة أو تعبيرات لا تبدو سلسة للناطقين بالعربية الفصحى. <p>✗ الخطأ: "ذهب إلى المدرسة مبارحه". ✓ الصحيح: "ذهب إلى المدرسة بالأمس".</p> <p>✗ الخطأ: "الطلاب يدرس كتابهم كل يوم". ✓ الصحيح: "الطلاب يدرسون كتابهم كل يوم".</p> <p>✗ الخطأ: "هذه المسألة في غاية الصعوبة للغاية". ✓ الصحيح: "هذه المسألة في غاية الصعوبة".</p>
PRN	خطأ في ترجمة أسماء العلم	<p>تم ترجمة الاسم العلم: مثل اسم شخص أو مكان أو منظمة—بشكل غير صحيح.</p> <p>✗ الخطأ: "اليوم رح روح ع الشعلان مع رفقاتي. = اليوم سأذهب إلى السيد الشعلان مع أصدقائي." ✓ الصحيح: "اليوم رح روح ع الشعلان مع رفقاتي. = اليوم سأذهب إلى الشعلان مع أصدقائي."</p>
TRM	خطأ في ترجمة المصطلحات الخاصة باللهجة	<p>لم يتم ترجمة المصطلحات أو التعبيرات الخاصة باللهجة، أو تمت ترجمتها بطريقة تغير المعنى.</p> <p>✗ الخطأ: "هالشغلة بدها شوية رواء، لا تستعجل عليها! = هذا الأمر يحتاج إلى بعض الشراب، لا تتعجل به!" ✓ الصحيح: "هالشغلة بدها شوية رواء، لا تستعجل عليها! = هذا الأمر يحتاج إلى بعض الهدوء، لا تتعجل به!"</p> <p>✗ الخطأ: "هالشب طلع شغيل وما بيمل من الشغل! = هذا الشاب طلع عامل، ولا يمل من العمل!" ✓ الصحيح: "هالشب طلع شغيل وما بيمل من الشغل! = هذا الشاب اتضح أنه نشيط جدًا ولا يمل من العمل!"</p> <p>✗ الخطأ: "اليوم الجو بيجنن، لازم نطلع نتمشي! = اليوم الطقس مجنون، يجب أن نخرج للتنزه!" ✓ الصحيح: "اليوم الجو بيجنن، لازم نطلع نتمشي! = اليوم الطقس رائع، يجب أن نخرج للتنزه!"</p>
GSMIS	أخطاء أخرى في الترجمة تؤثر على المعنى	<p>تغيير آخر في المعنى: ليس بسبب (PRN) أو (TRM)، ولكن لم يتم الحفاظ على المعنى بشكل كامل بتغيير أو إضافة أو حذف المعنى الأصلي لجملته باللهجة.</p> <p>✗ الخطأ: "ما بدني فوت بهالسيرة، الموضوع حساس كثير = لا أريد الدخول في هذه القصة، فهي طويلة جدًا." ✓ الصحيح: "ما بدني فوت بهالسيرة، الموضوع حساس كثير = لا أريد التحدث في هذا الموضوع، فهو حساس جدًا."</p>
ADP	خطأ في التكيف	<p>يشير التكيف إلى مدى سلاسة الترجمة وملاءمتها للسياق، والموقف.</p> <p>أمثلة على أخطاء التكيف:</p> <ul style="list-style-type: none"> تبدو الترجمة حرفية جدًا وتحتاج إلى إعادة صياغة لتناسب اللغة الفصحى بشكل أفضل. لا تتناسب نبرة أو أسلوب الترجمة مع الموقف أو السياق (مثل استخدام لغة رسمية جدًا في سياق غير رسمي). <p>✗ الخطأ: "شو قصتك؟ ليش معصّب؟ = ما هي قصتك؟ لماذا أنت غاضب؟" ✓ الصحيح: "شو قصتك؟ ليش معصّب؟ = ما الأمر؟ لماذا تبدو غاضبًا؟"</p> <p>✗ الخطأ: "هاد الحكى ما بيمشي معي! = هذا الكلام لا يمشي معي!" ✓ الصحيح: "هاد الحكى ما بيمشي معي! = هذا الكلام غير مقبول بالنسبة لي!"</p>

تأكيد: إذا وجدت أي من أخطاء PRN أو TRM أو GSMIS في الترجمة فتجاهل الخطأ الأخير ADP

Figure 7: The annotation guidelines provided to human annotators explain each error type with illustrative examples, assisting them in using the Ara-HOPE Annotation Decision Tree.

Indic-TunedLens: Interpreting Multilingual Models in Indian Languages

Mihir Panchal¹, Deeksha Varshney², Mamta .³, Asif Ekbal⁴

¹ Dwarkadas Jivanlal Sanghvi College of Engineering, Mumbai, India

² Indian Institute of Technology Jodhpur, Jodhpur, India

³ King’s College London, London, UK

⁴ Indian Institute of Technology Patna, Patna, India

mihirpanchal5400@gmail.com, deeksha@iitj.ac.in,

mamta.name@kcl.ac.uk, asif@iitp.ac.in

Abstract

Multilingual large language models (LLMs) are increasingly deployed in linguistically diverse regions like India, yet most interpretability tools remain tailored to English. Prior work reveals that LLMs often operate in English centric representation spaces, making cross lingual interpretability a pressing concern. We introduce *Indic-TunedLens*, a novel interpretability framework specifically for Indian languages that learns shared affine transformations. Unlike the standard Logit Lens, which directly decodes intermediate activations, *Indic-TunedLens* adjusts hidden states for each target language, aligning them with the target output distributions to enable more faithful decoding of model representations. We evaluate our framework on 10 Indian languages using the MMLU benchmark and find that it significantly improves over SOTA interpretability methods, especially for morphologically rich, low resource languages. Our results provide crucial insights into the layer-wise semantic encoding of multilingual transformers. Our model is available at <https://huggingface.co/spaces/MihirRajeshPanchal/IndicTunedLens>. Our code is available at <https://github.com/MihirRajeshPanchal/IndicTunedLens>.

1 Introduction

The remarkable advancements of transformer models (Vaswani et al., 2017) across diverse domains, particularly in Natural Language Processing (NLP), underscore the critical need to decipher their internal representations and intricate reasoning processes. This is essential to ensure trust and reliability in the systems. Existing research has explored various methods to extract specific linguistic concepts, such as speech tagging parts or syntactic structure, from hidden states with the help of probing classifiers (Li et al., 2023; Lee et al.,

2025) and LogitLens (nostalgebraist, 2020). Belrose et al. (2023) interpret the layers which offered a significant step forward in extracting meaningful intermediate predictions by training layer specific translators. Their approach builds upon the concept of LogitLens, which directly decodes hidden states into the vocabulary space using the model’s pre-trained unembedding matrix. Most of these approaches fail to generalize effectively to linguistically diverse and low resource languages.

India is a multilingual country and Indian languages exhibit vastly different linguistic structures from English, including rich morphology and more flexible word order (common in Hindi and Punjabi) (Srirampur et al., 2014). These diverse scripts (e.g., Gurmukhi for Punjabi, Devanagari for Hindi and Marathi) and their unique tokenization schemes which often result in distinct subword units and vocabulary distributions, pose another significant challenges in applying English interpretability models to Indian languages (Rahman Khan et al., 2024). Schut et al. (2025) highlights a critical limitation of current multilingual LLMs, even when prompted and generating outputs in non-English languages, the models tend to operate in an English centric latent space. Using logit lens decoding, they show that LLMs first activate English word representations before translating them into the target language, suggesting that key semantic reasoning is biased toward English. However, this analysis is limited to European and East Asian languages, and does not explore morphologically rich and typologically distinct languages such as those found in the Indian subcontinent.

To address these critical limitations, we developed *Indic-TunedLens* framework specifically for multilingual contexts, and demonstrate its efficacy on transformer models for 10 low resource indian languages. Our approach involves training a single affine transformation to learn shared representations across languages, while also enabling align-

ment of intermediate hidden states with the final output distribution for Indic languages. This allows the lens to capture the unique representational nuances of these languages, providing robust and interpretable insights into how these diverse and morphologically rich languages are processed layer by layer within transformer models. We conduct a comprehensive evaluation on the Sarvam-1 model¹ using Bengali, English, Gujarati, Hindi, Kannada, Malayalam, Marathi, Panjabi, Nepali, Tamil and Telugu languages with a multilingual MMLU benchmark (Dac Lai et al., 2023) to demonstrate the effectiveness of *Indic-TunedLens* framework. Our findings not only bridge a significant gap in multilingual transformer interpretability but also offer a crucial tool for understanding and enhancing NLP applications across the rich linguistic tapestry of India. To the best of our knowledge, this is the first work to apply layer wise interpretability to multilingual LLMs for Indian languages. Our work makes the following contributions:

1. We show that interpretability failure in Indian languages is a projection problem, and propose *Indic-TunedLens* as a solution. It aligns intermediate hidden states with output distributions in morphologically rich and syntactically diverse low resource indic languages.
2. We conduct a comprehensive evaluation on the Sarvam-1 model on Bengali, English, Gujarati, Hindi, Kannada, Malayalam, Marathi, Nepali, Tamil and Telugu languages with a multilingual MMLU benchmark.
3. Our results show that *Indic-TunedLens* enables accurate decoding of intermediate representations and reveals distinct layer-wise and token position specific reasoning dynamics across languages.

2 Related Work

2.1 Interpretability in Transformer Models

The growing complexity of transformer based language models has led to increased interest in interpretability techniques that help unpack how internal representations evolve across layers. Early efforts, such as probing classifiers (Li et al., 2023; Lee et al., 2025) attempted to extract linguistic features like syntax and part-of-speech information from hidden states. Methods, like the Logit Lens (nostalgebraist, 2020), directly decode intermediate

activations into vocabulary space using the model’s unembedding matrix. Building on this, the Tuned Lens (Belrose et al., 2023) improves interpretability by learning affine transformations that align hidden states with the final output distribution. Complementary methods, such as causal tracing (Meng et al., 2022) and representation steering (Subramani and Bau, 2022) have also enabled insights into factual recall and controlled generation.

2.2 Multilingual Interpretability

India’s linguistic diversity presents unique challenges for model interpretability. Wen-Yi and Mimno (2023) demonstrate that multilingual models learn input embeddings where translated tokens across languages cluster together, even without an explicit translation objective. Similarly, Gopinath and Rodriguez (2024) observe that self-supervised speech models exhibit diverse attention head behaviors across languages, with diagonal heads playing a crucial role in cross lingual phoneme classification. Schut et al. (2025) demonstrate that multilingual LLMs often process semantically meaningful content in an English centric representation space, regardless of input/output language. This bias becomes more pronounced for lexical tokens, affecting model transparency and fairness. Despite these advances, most interpretability research remains centered on English or non-LLM architectures. Notably, Siddiqui et al. (2024) use LIME to interpret hate speech classifiers in low-resource languages like Urdu and Sindhi, however, this work focuses on traditional classification tasks, not on large language models or multilingual reasoning. Additionally, Saji et al. (2025) show that LLMs often represent non-Roman script languages in Romanized form in intermediate layers, a phenomenon termed Latent Romanization highlighting shared cross-script representations that complement prior findings on English-centric biases.

In contrast to these prior approaches, we propose *Indic-TunedLens*, an interpretability framework tailored for Indian languages. Unlike existing methods that apply English-tuned lens, our approach learns a shared affine transformation adapted to Indic scripts, aligning hidden states with vocabulary distributions for Bengali, English, Gujarati, Hindi, Kannada, Malayalam, Marathi, Panjabi, Nepali, Tamil and Telugu, and enabling fine-grained, language-aware analysis of transformer model representations.

¹<https://www.sarvam.ai/blogs/sarvam-1>

3 Indic-TunedLens

The Tuned Lens was introduced by (nostalgebraist, 2020), is a probing technique designed to interpret intermediate hidden states of transformer models by projecting them into the final output space. It enables inspection of the model’s predictions at each layer without requiring task specific supervision.

We adapt this framework to Indian languages and introduce *Indic-TunedLens*, which applies the same affine transformation approach to better capture the unique challenges posed by Indic scripts and morphologically rich, low-resource settings. Unlike existing applications that evaluate Tuned Lens in English-centric contexts, our method explicitly trains and evaluates on Indian languages, providing new insights into their representation dynamics.

Affine Translation of Representations. For a transformer layer n , let h_n denote the hidden state. The *Indic-TunedLens* maps h_n to a distribution over vocabulary tokens by first applying a learned affine transformation, a *translator*, comprising a matrix $M_n \in \mathbb{R}^{d \times d}$ and a bias vector $b_n \in \mathbb{R}^d$. The transformed representation is then passed through the model’s final layer head to produce token logits:

$$\begin{aligned} \text{Indic-TunedLens}_n(h_n) &= \text{LogitHead}(\tilde{h}_n) \\ \text{where } \tilde{h}_n &= M_n h_n + b_n \end{aligned} \quad (1)$$

where LogitHead refers to the model’s output projection layer (typically a linear map followed by softmax over the vocabulary).

Learning Objective. The *Indic-TunedLens* is trained to reproduce the predictions of the base model’s final layer. For a given input x , the base model produces a next-token probability distribution $p_{\text{final}}(x)$ at its output layer. We treat this distribution as the supervision signal. For each intermediate layer n , the *Indic-TunedLens* produces its own distribution by translating the hidden state h_n and passing it through the model’s output head. We minimize the Kullback-Leibler divergence between these two distributions:

$$\min_{M_n, b_n} \mathbb{E}_{h_n} \left[D_{\text{KL}}(p_{\text{final}}(x) \parallel \text{Indic-TunedLens}_n(h_n)) \right] \quad (2)$$

where $p_{\text{final}}(x) = \text{softmax}(Wh_L)$ denotes the next-token probability distribution produced by the

model’s final layer L on input x , and h_n is the hidden state at intermediate layer n .

In other words, the model’s own final-layer predictions serve as the “labels,” and the *Indic-TunedLens* learns to align intermediate hidden states with this distribution.

4 Experimental Setup and Analysis

Indic-TunedLens is trained using the Sarvam-1 model as the base model. All the implementations are carried out in Pytorch². To train and test *Indic-TunedLens*, we utilize two datasets that provide comprehensive multilingual coverage across Indian languages. We train our model on eleven languages, Bengali, English, Gujarati, Hindi, Kannada, Malayalam, Marathi, Panjabi, Nepali, Tamil and Telugu from the Sangraha dataset (Rahman Khan et al., 2024). For evaluation, we employ a curated subset of the multilingual Massive Multitask Language Understanding (MMLU) dataset (Dac Lai et al., 2023) adapted for Indian languages, comprising Bengali, English, Gujarati, Hindi, Kannada, Malayalam, Marathi, Nepali, Tamil and Telugu samples. The inclusion of Panjabi as a train only language enables assessment of cross lingual transfer capabilities to unseen but related languages. Detailed dataset statistics and implementation details are provided in Appendices A.1 and A.2, respectively.

Entropy Analysis. To analyze the confidence and uncertainty of predictions at each layer, we compute the Shannon entropy of the probability distribution produced by the *Indic-TunedLens*:

$$H_n = - \sum_{i=1}^{|V|} p_i^{(n)} \log p_i^{(n)}, \quad (3)$$

where $p_i^{(n)}$ is the probability assigned to vocabulary token i at layer n , and $|V|$ is the vocabulary size. Lower entropy values indicate confident predictions (probability mass concentrated on fewer tokens), while higher entropy values indicate uncertain predictions (probability mass distributed across many tokens).

Layer-wise Accuracy We quantify alignment between intermediate representations and the final next-token prediction using a layer-wise agreement score. For a sequence of length T , the top-1 token predicted by a lens at layer n for position t is

$$\hat{y}_t^{(n)} = \arg \max_{i \in V} p_i^{(n)}(\cdot | t). \quad (4)$$

²<https://pytorch.org/>

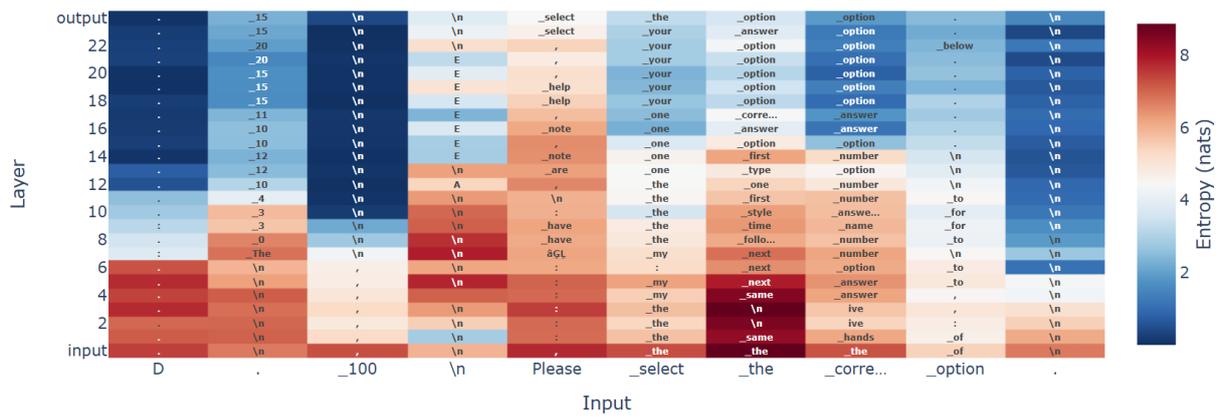


Figure 1: This figure shows the entropy heatmap of the standard Tuned Lens, which was developed for English-centric models. The high and irregular entropy across layers suggests unstable intermediate representations and weak alignment for Indian languages, with predictions biased toward English tokens.

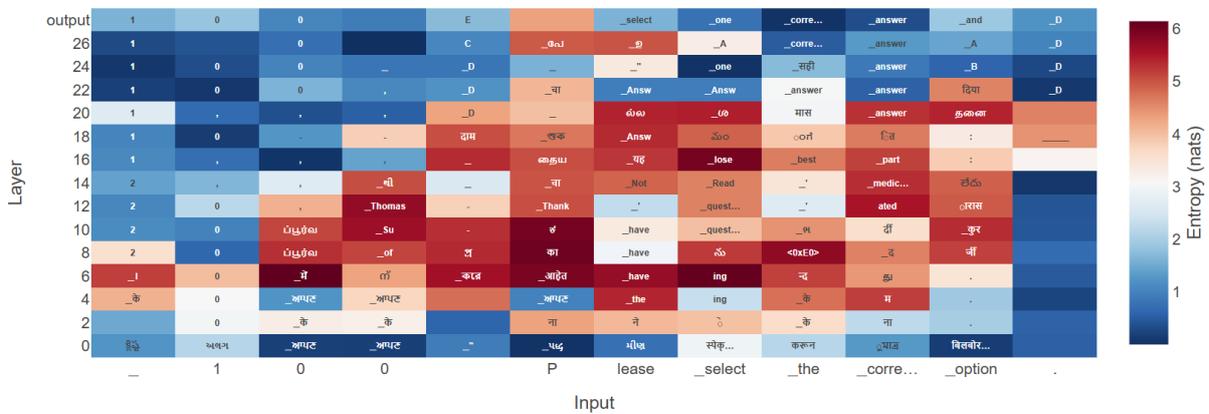


Figure 2: Entropy heatmap for the Indic-TunedLens. Entropy decreases more smoothly across layers, indicating progressive information consolidation and improved semantic alignment, with intermediate predictions increasingly generating meaningful Hindi tokens.

Agreement is measured against the base model’s prediction at the final layer L . This interpretability metric indicates how early sufficient information for the final decision emerges. Higher early-layer agreement reflects better alignment of intermediate states with the vocabulary, enabling faithful intermediate predictions.

5 Results

To assess the effectiveness of Indic-TunedLens for multilingual transformer interpretability, we conduct a comprehensive comparative analysis against the standard Logit Lens approach across 10 Indian languages (detailed in Appendix A).

Layer-wise uncertainty reveals language-specific meaning formation. The standard Tuned Lens approach, originally designed for English centric models, demonstrates significant limitations when applied to Indian language

Question: एक मोनोपोली के लिए ध्यान केंद्रित करने योग्य अनुपात है?
(What is the focus ratio for a monopoly?)

A. 0

B. 5

C. 10

D. 100

Please select the correct option.

Figure 3: Input Question from MMLU Hindi Dataset

processing, as illustrated in Figure 1. We show an example question (Figure 3) about monopoly concentration ratios from the Hindi MMLU dataset³. The entropy (calculated as discussed in Section 4) heatmap which measures uncertainty in the model’s predicted token distribution at each layer reveals inconsistent patterns with high variance across layers, indicating poor alignment between

³https://huggingface.co/datasets/alexandrinst/m_mmlu/viewer/hi

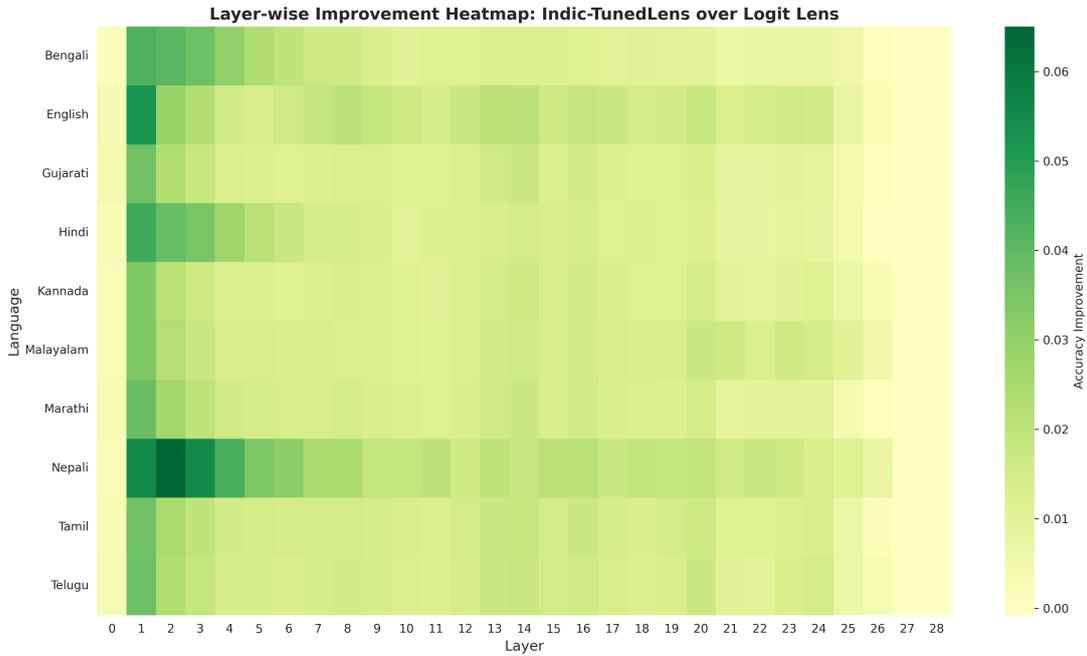


Figure 4: Layer-wise Improvement Patterns

the tuning mechanism and the model’s internal processing pathways for Indian languages. The visualization shows fragmented attention patterns, particularly in the middle layers (8-16), where semantic consolidation should occur, and scattered high entropy regions without clear progressive information refinement. This degraded performance stems from a projection mismatch: while the model internally encodes morphologically rich and semantically meaningful representations for Indian languages, the English-centric Tuned Lens fails to align these representations with the output vocabulary space.

The Indic-Tuned Lens, trained on Indian languages, demonstrates substantially improved performance characteristics as shown in Figure 2. The entropy heatmap exhibits coherent, systematic entropy reduction patterns across layers, indicating proper information consolidation aligned with Indic linguistic structures. Enhanced semantic capture is evident through concentrated attention patterns around key vocabulary elements, while distinct processing phases are observable with early layers managing tokenization and morphological analysis, middle layers handling semantic understanding, and later layers focusing on answer. Without language-aware projection, intermediate representations in Indian languages appear fragmented and difficult to interpret, obscuring how meaning forms across layers. The superior performance of

Indic-TunedLens demonstrates the critical need for developing specialized interpretability frameworks that account for diverse linguistic structures, and semantic processing patterns beyond English centric approaches.

Maximum improvements occur in early layers (1-8) with language specific patterns reflecting morphological complexity. The heatmap in Figure 4 shows that Bengali, Nepali, and Hindi exhibit the strongest improvements in layers 1-4 (0.04-0.06 accuracy gain), corresponding to the morphological analysis phase where rich inflectional systems are processed. English shows concentrated improvements at layer 1 (0.06 gain) and more modest gains throughout middle layers, reflecting its analytic morphology. Gujarati and Kannada show very similar sustained improvements across layers 1-10 (0.02-0.04 gain), while Malayalam exhibits more distributed improvement patterns extending into middle layers (5-15). These layer specific improvement patterns enable linguistic hypotheses about where different linguistic phenomena are computed, suggesting that morphological analysis and early semantic composition occur in early layers, while deeper layers support more complex compositional and syntactic integration. This process happens earlier in languages like Telugu and Tamil, which form words by joining many meaningful parts together, compared to languages with more isolated word structures. The variation in

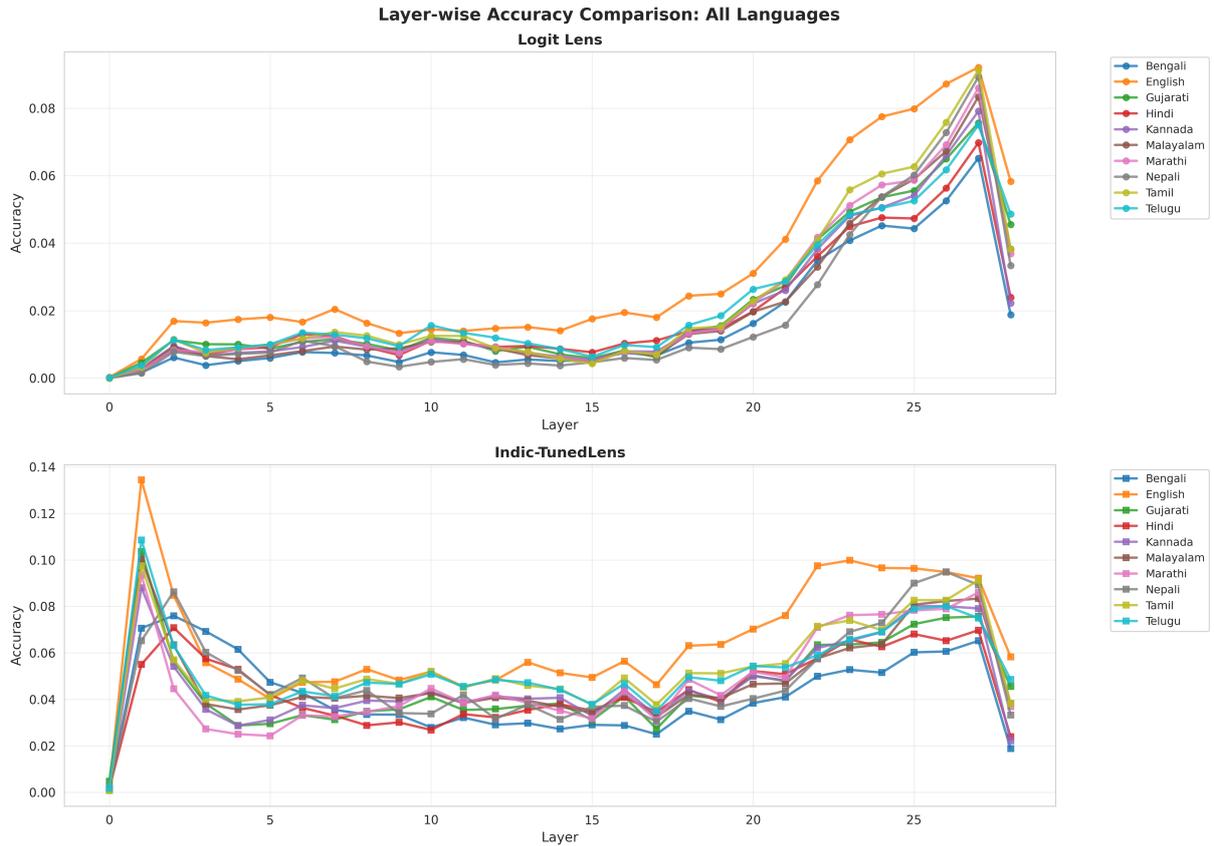


Figure 5: Layer Wise Accuracy Comparison

improvement patterns across languages suggests that the learned transformations adapt to language specific processing dynamics, capturing the distinct computational trajectories by which different Indic languages resolve ambiguity and build compositional meaning. These findings demonstrate that interpretability methods cannot be language-agnostic: effective analysis of multilingual LLMs requires language-family-aware lenses that respect typological differences in morphological and compositional structure.

Indic-TunedLens demonstrates superior accuracy over Logit Lens across all Indian languages, with particularly pronounced improvements in early and middle layers. As shown in Figure 5, the standard Logit Lens exhibits minimal accuracy in early layers (0-15) across all languages, with values remaining below 0.02 for most Indian languages. The accuracy only begins to increase substantially after layer 20, reaching peak performance at layer 27 before dropping sharply at the final layer. In contrast, Indic-TunedLens shows immediate interpretability from layer 1, with English achieving 0.13 accuracy, followed by Telugu (0.11) and Gujarati (0.10). The framework maintains consis-

tent accuracy between 0.03-0.06 throughout middle layers (5-20) for Indian languages, demonstrating robust intermediate representation capture. These findings suggest that language specific affine transformations enable meaningful interpretation of hidden states from the earliest layers of processing, whereas English centric approaches fail to capture the morphological richness and syntactic diversity inherent in Indian languages until much deeper in the network. The superior early layer performance of Indic-TunedLens indicates that specialized training on diverse Indic scripts creates representations better aligned with the model’s internal processing pathways for these languages, enabling more accurate decoding of intermediate semantic states. Individual language specific accuracy curves across all layers are provided in the appendix B.1.

Indic-TunedLens consistently positions correct tokens at higher ranks across all layers, indicating improved confidence in predictions. Figure 6 reveals that Logit Lens maintains average ranks between 4-5 across layers 5-25 for most languages, with Malayalam showing particularly poor performance (ranks 4.5-5.8). The ranks only improve dramatically at layer 27, converging to 3.5-4

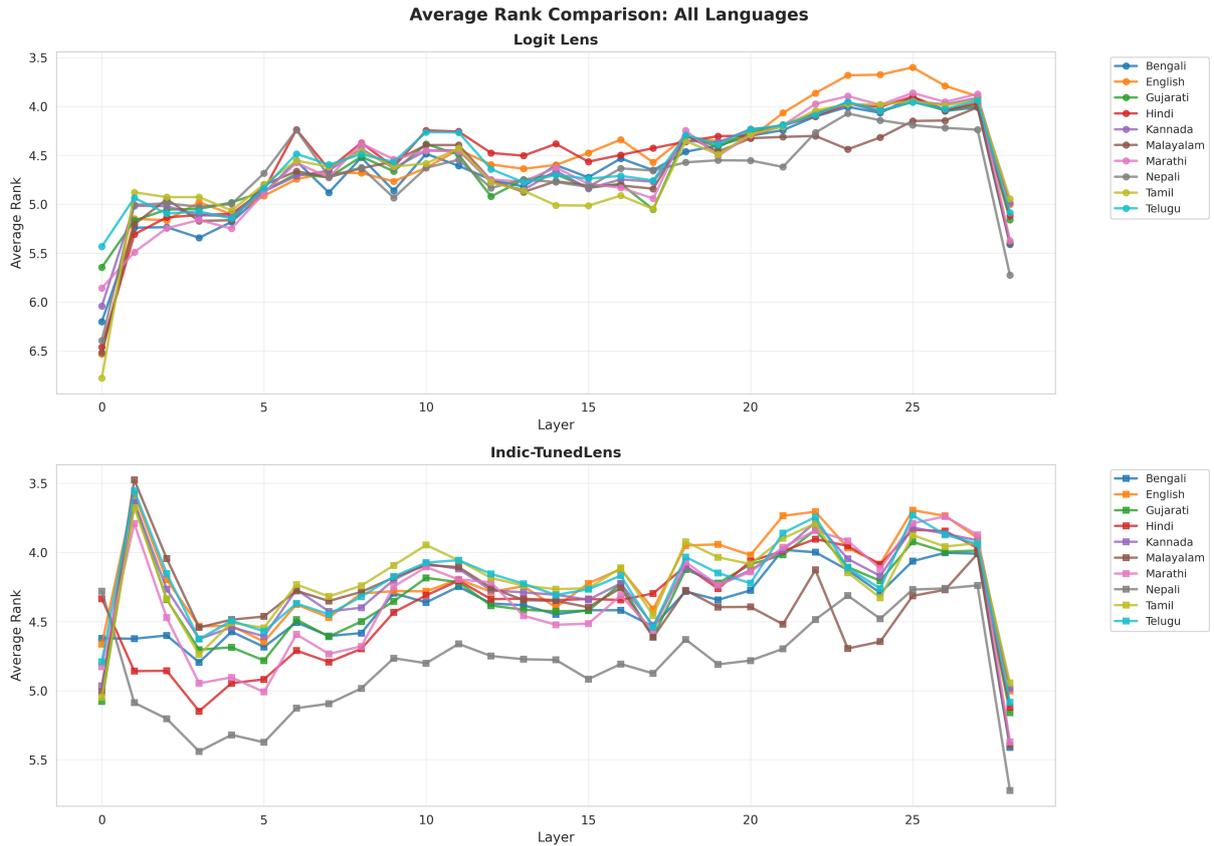


Figure 6: Average Rank of Correct Predictions

across languages. Conversely, Indic-TunedLens exhibits more favorable ranking patterns throughout the network, with most languages maintaining average ranks between 3.9-4.7 across middle layers. Notably, Hindi demonstrates superior rank positioning (4.5-5 in early layers, improving to 3.5-4.0 in later layers), while Malayalam shows the most substantial improvement from its Logit Lens baseline, achieving ranks of 4.2-4.8 in layers 10-25. The consistently better ranking performance of Indic-TunedLens across all languages demonstrates that the learned affine transformations more effectively align intermediate representations with the final vocabulary distribution, placing correct tokens among the top candidates earlier in the decoding process. This improved ranking behavior is particularly critical for morphologically rich languages where token level disambiguation requires integration of multiple linguistic features, and suggests that Indic-TunedLens better captures the hierarchical semantic refinement that occurs across transformer layers for Indian languages. Detailed rank distribution plots for each language are available in the appendix B.2.

Accuracy varies substantially across token posi-

tions, with specific positions showing language dependent spikes corresponding to answer tokens. Figure 7 demonstrates highly variable accuracy patterns across the 600+ token positions in the MMLU dataset samples. Marathi exhibits the most dramatic spike at position 380 (0.95 accuracy), followed by English at position 550 (0.52 accuracy) and Hindi at position 210 (0.58 accuracy). These spikes correspond to answer choice positions in the multiple choice questions, where the model’s predictions align with the correct options. Most languages maintain baseline accuracy of 0.05-0.15 across question and context tokens (positions 0-400), with gradual increases toward answer regions. Bengali, Telugu, and Gujarati show multiple moderate peaks (0.20-0.35) distributed across positions 350-500, suggesting consistent prediction quality across answer choices. The position specific accuracy patterns reveal that Indic-TunedLens successfully captures the model’s reasoning trajectory through question answering tasks, with high accuracy at answer token positions indicating effective alignment between intermediate representations and final decision making processes. The language specific variation in peak positions reflects differ-

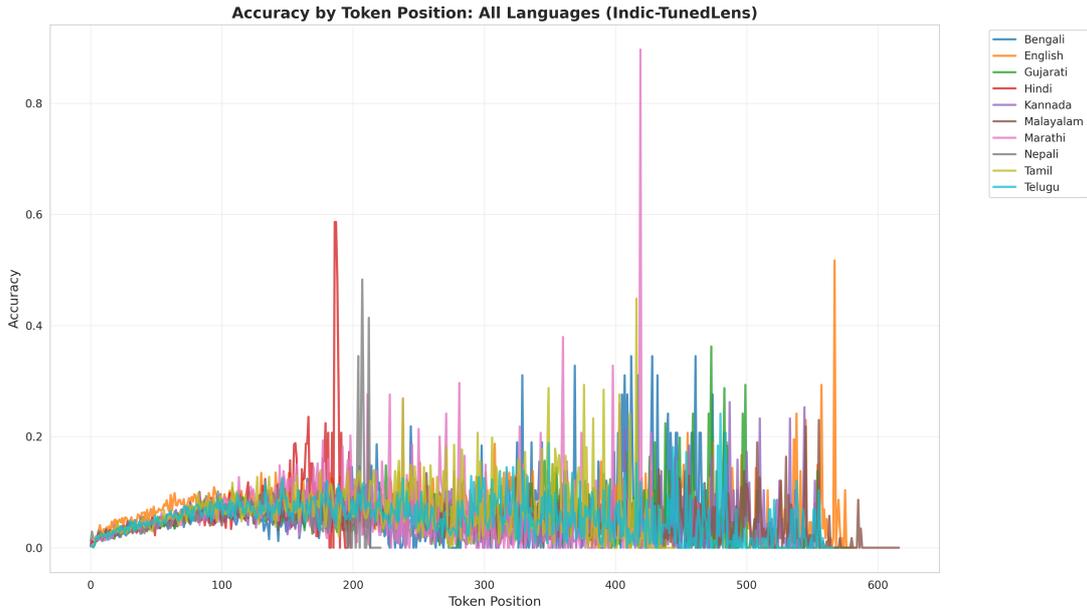


Figure 7: Accuracy by Token Positions

ences in tokenization schemes and morphological complexity, where languages with more complex morphology (Tamil, Telugu) require more tokens to represent equivalent semantic content, shifting the spatial distribution of critical reasoning steps. This position based analysis provides evidence that the interpretability framework maintains fidelity to the model’s internal decision boundaries across diverse linguistic structures and question formats.

6 Discussion

Our work establishes Indic-TunedLens as a specialized interpretability framework for morphologically rich Indian languages through three key findings from our comparative analysis.

As demonstrated in Figure 5, *Indic-TunedLens* enables meaningful interpretation from layer 1 onwards, whereas standard Logit Lens (nostalgebraist, 2020) fails to capture intermediate semantics until layer 20. This 20-layer interpretability gap reveals that English centric methods fundamentally misalign with how multilingual models process morphologically rich languages. The immediate accessibility in early layers suggests that learned affine transformations successfully bridge the representational mismatch between intermediate hidden states and final vocabulary distributions. Recent work shows that language specific neurons concentrate in mid-to-late layers (Tang et al., 2024; Chou et al., 2025), yet our framework exposes semantically meaningful states much earlier, indicating

that the computational bottleneck lies in projection mechanisms rather than internal processing (Schut et al., 2025).

These findings have broader implications for multilingual NLP. First, interpretability cannot be language agnostic. The same technique yields fundamentally different results depending on linguistic structure (Wen-Yi and Mimno, 2023). Second, our framework provides a methodological template for other morphologically rich families. Third, the Indic-TunedLens interpretability tool enables the identification of language specific computational regions and layers, revealing that multilingual competence arises from localized mechanisms that can be selectively analyzed or manipulated for a more controlled, fair, and reliable multilingual behavior (Tang et al., 2024). By introducing the Indic-TunedLens framework, we enable researchers in multilingual NLP to analyze multilingual large language models across the languages on which they are trained. By enabling understanding of how models process diverse languages, *Indic-TunedLens* supports more equitable and responsible AI development in linguistically diverse regions.

7 Conclusion

In this paper, we have introduced Indic-TunedLens, a multilingual interpretability framework that extends the Tuned Lens paradigm to structurally diverse Indian languages. By learning language affine transformations, our approach significantly

improves the layer-wise interpretability of multilingual LLMs like Sarvam-1 on Bengali, English, Gujarati, Hindi, Kannada, Malayalam, Marathi, Nepali, Tamil and Telugu. Extensive experiments demonstrate that Indic-TunedLens outperforms the standard English centric methods in rank based accuracy, layer wise accuracy performance, and entropy alignment. Additionally, we aim to extend interpretability methods to directly improve downstream performance in multilingual tasks.

Limitations

While *Indic-TunedLens* enhances interpretability for Indian languages, it has certain limitations. Due to resource constraints and unavailability of pre-training data, we initially fine tune only the 1B-parameter Sarvam-1 model. Future work can explore multilingual models with higher number of parameters. Second, our evaluation is limited to the MMLU dataset, which restricts the diversity of tasks used to assess large language model behavior. Finally, while we achieve better interpretability, this does not directly translate to improved downstream task performance, which remains an open question for future research.

Ethics Statement

We use publicly accessible datasets for our experiments, strictly for academic purposes and in full accordance with their licensing terms.

Acknowledgments

Mamta gratefully acknowledges the support from the Engineering and Physical Sciences Research Council (EPSRC, grant number EP/X04162X/1).

References

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*.

Cheng-Ting Chou, George Liu, Jessica Sun, Cole Blondin, Kevin Zhu, Vasu Sharma, and Sean O'Brien. 2025. Causal language control in multilingual transformers via sparse feature steering. *arXiv preprint arXiv:2507.13410*.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages

with reinforcement learning from human feedback. *arXiv e-prints*, pages arXiv-2307.

- Sai Gopinath and Joselyn Rodriguez. 2024. Probing self-attention in self-supervised speech models for cross-linguistic differences. *arXiv preprint arXiv:2409.03115*.
- Andrew Lee, Lihao Sun, Chris Wendler, Fernanda Viégas, and Martin Wattenberg. 2025. The geometry of self-verification in a task-specific reasoning model. *arXiv preprint arXiv:2504.14379*.
- Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Emergent world representations: Exploring a sequence model trained on a synthetic task. *ICLR*.
- Kevin Meng, David Bau, McKane Andrus, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- nostalgebraist. 2020. Interpreting gpt: The logit lens. <https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens>. Accessed: 2025-07-28.
- Mohammed Safi Ur Rahman Khan, Priyam Mehta, Ananth Sankar, Umashankar Kumaravelan, Sumanth Doddapaneni, Varun Balan G, Sparsh Jain, Anoop Kunchukuttan, Pratyush Kumar, Raj Dabre, and 1 others. 2024. Indicllmsuite: A blueprint for creating pre-training and fine-tuning datasets for indian languages. *arXiv e-prints*, pages arXiv-2403.
- Alan Saji, Jaavid Aktar Husain, Thanmay Jayakumar, Raj Dabre, Anoop Kunchukuttan, and Ratish Pudupully. 2025. Romanlens: The role of latent romanization in multilinguality in llms. *arXiv preprint arXiv:2502.07424*.
- Lisa Schut, Yarin Gal, and Sebastian Farquhar. 2025. Do multilingual llms think in english? *arXiv preprint arXiv:2502.15603*.
- Jawaid Ahmed Siddiqui, Siti Sophiyati Yuhaniz, Ghulam Mujtaba, Safdar Ali Soomro, and Zafar Ali Mahar. 2024. Fine-grained multilingual hate speech detection using explainable ai and transformers. *IEEE Access*.
- Saikrishna Srirampur, Ravi Chandibhamar, and Radhika Mamidi. 2014. *Statistical morph analyzer (SMA++) for Indian languages*. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 103–109, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.
- Divyansh Subramani and David Bau. 2022. Directional stimulus optimization: Open-ended text generation as activation steering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Wayne Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Andrea W Wen-Yi and David Mimno. 2023. Hyperpolyglot llms: Cross-lingual interpretability in token embeddings. *arXiv preprint arXiv:2311.18034*.

A Experimental Setup

A.1 Dataset

Sangraha Dataset The Sangraha dataset serves as our primary training corpus for developing language specific interpretability lenses that can effectively analyze the internal representations of multilingual models. As detailed in Table A.1, the composition of the dataset is strategically balanced to ensure adequate representation in all languages while maintaining computational feasibility.

Language	Number of Rows
Bengali	11497021
English	17482249
Gujarati	3970097
Hindi	17420932
Kannada	3632345
Malayalam	6370342
Marathi	5865617
Panjabi	1738597
Nepali	91476306
Tamil	7828512
Telugu	7081734

Table 1: Sangraha Dataset for Training Sarvam Tuned Lens

The substantial size of each language subset ensures sufficient statistical power for training robust interpretability mechanisms, while the linguistic diversity across multiple Indian languages provides the necessary foundation for developing generalizable interpretability methods.

MMLU Dataset Our evaluation framework employs a curated subset of the multilingual Massive Multitask Language Understanding (MMLU) dataset, adapted for Indian languages [Dac Lai et al.](#)

(2023). As shown in Table 2, the testing dataset includes samples across ten languages, with relatively balanced distribution ranging from 215 to 277 samples per language. This balanced distribution enables fair comparative analysis while testing the generalization capabilities of our approach across the diverse linguistic landscape.

Language	Number of Rows
Bengali	216
English	277
Gujarati	243
Hindi	235
Kannada	261
Malayalam	265
Marathi	221
Nepali	215
Tamil	251
Telugu	272

Table 2: MMLU Dataset for Testing Tuned Lens

A.2 Implementation Details

Our experimental framework is built around the Sarvam-1 model, which has been specifically optimized for 10 Indian languages: Bengali (bn), English (en) Gujarati (gu), Hindi (hi), Kannada (kn), Malayalam (ml), Marathi (mr), Nepali(np), Punjabi (pa), Tamil (ta), and Telugu (te). This multilingual optimization makes it an ideal candidate for studying interpretability across diverse Indian languages, as it inherently possesses cross lingual capabilities developed through extensive multilingual pre training.

For training the tuned lens, we utilize the tuned-lens package, which provides a robust framework for training and evaluating lenses on transformer models like Sarvam-1 ([Belrose et al., 2023](#)). The package enables us to probe intermediate layers of the model to analyze how multilingual tokens influence prediction patterns.

The training configuration detailed in Table 3 utilizes a distributed setup with careful resource allocation to ensure stable training while maintaining computational efficiency. We employ Fully Sharded Data Parallel (FSDP) training to handle the large scale nature of the Sarvam-1 model effectively. The single node configuration with 5 processes per node represents an optimal balance between computational efficiency and resource availability. The per GPU batch size of 1 is chosen to

Hyperparameter	Value
Model Name	sarvamai/sarvam-1
# of Nodes	1
Processes per Node	5
Per-GPU Batch Size	1
FSDP	Enabled
Launch Mode	Standalone

Table 3: Training Configuration and Hyperparameters for Tuned Lens with Sarvam-1

maximize memory utilization while preventing out of memory errors during training of the tuned lens components, given the substantial memory requirements of the Sarvam-1 model architecture.

B Detailed Analysis

B.1 Language Wise Layer-wise Accuracy Comparison between Indic-TunedLens and LogitLens

This section presents a comprehensive language specific analysis of layer wise accuracy patterns, comparing the performance of Indic-TunedLens against the standard Logit Lens across all evaluated Indian languages. Each plot illustrates the accuracy across all 28 layers of Sarvam-1 model, revealing distinct processing characteristics for individual languages. The layer-wise accuracy comparison is shown for Bengali in Figure 8, English in Figure 9, Gujarati in Figure 10, Hindi in Figure 11, Kannada in Figure 12, Malayalam in Figure 13, Marathi in Figure 14, Nepali in Figure 15, Tamil in Figure 16, and Telugu in Figure 17.

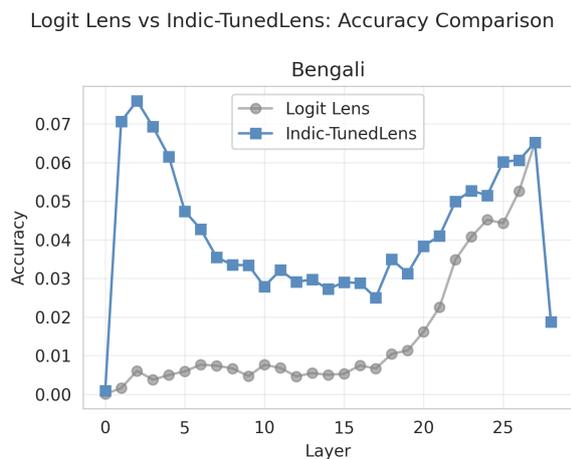


Figure 8: Layer-wise Accuracy Comparison for Bengali

Logit Lens vs Indic-TunedLens: Accuracy Comparison

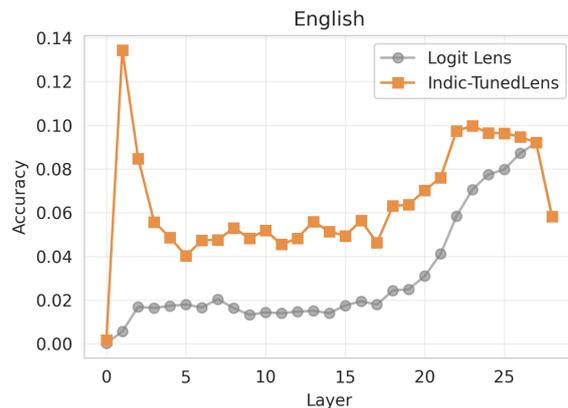


Figure 9: Layer-wise Accuracy Comparison for English

Logit Lens vs Indic-TunedLens: Accuracy Comparison

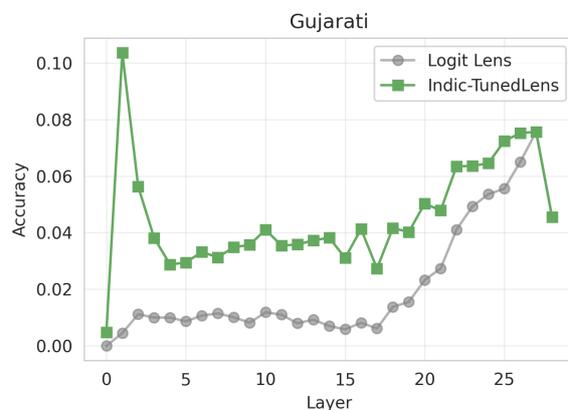


Figure 10: Layer-wise Accuracy Comparison for Gujarati

Logit Lens vs Indic-TunedLens: Accuracy Comparison

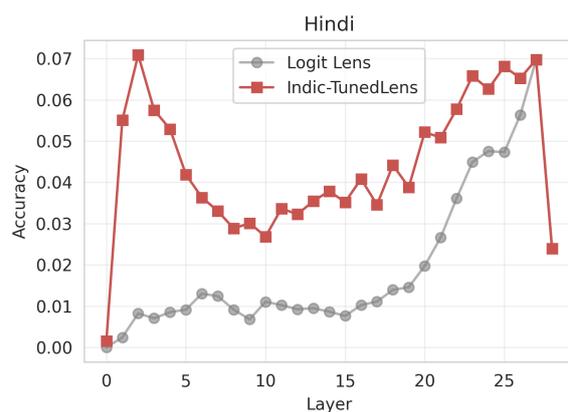


Figure 11: Layer-wise Accuracy Comparison for Hindi

Logit Lens vs Indic-TunedLens: Accuracy Comparison

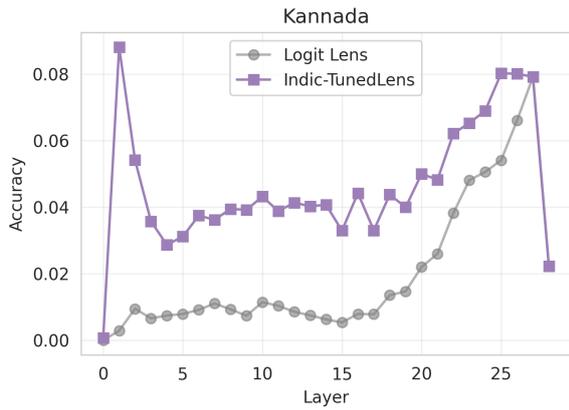


Figure 12: Layer-wise Accuracy Comparison for Kannada

Logit Lens vs Indic-TunedLens: Accuracy Comparison

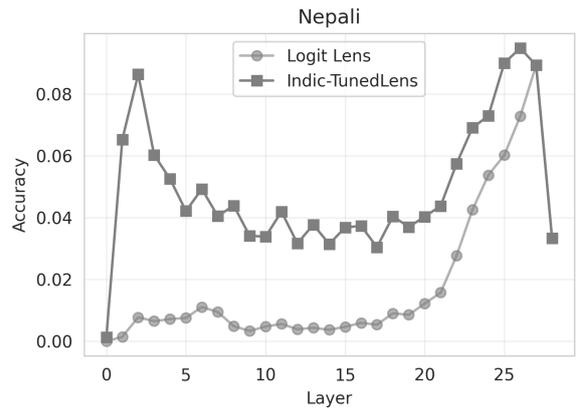


Figure 15: Layer-wise Accuracy Comparison for Nepali

Logit Lens vs Indic-TunedLens: Accuracy Comparison

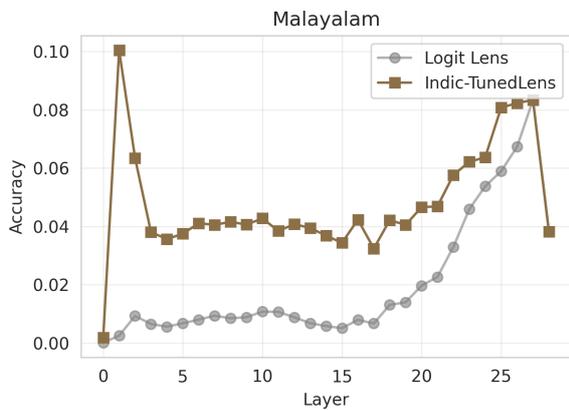


Figure 13: Layer-wise Accuracy Comparison for Malayalam

Logit Lens vs Indic-TunedLens: Accuracy Comparison

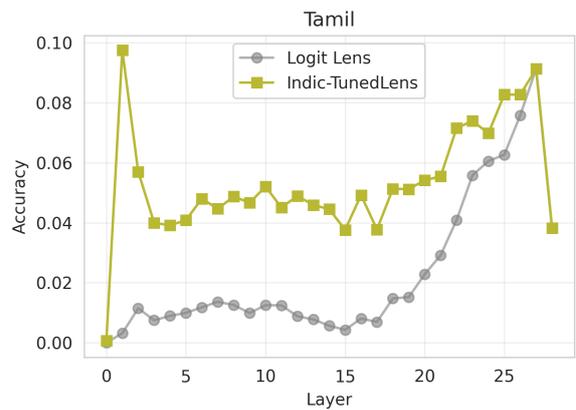


Figure 16: Layer-wise Accuracy Comparison for Tamil

Logit Lens vs Indic-TunedLens: Accuracy Comparison

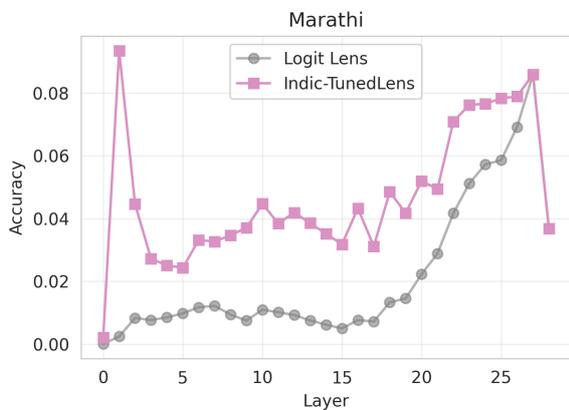


Figure 14: Layer-wise Accuracy Comparison for Marathi

Logit Lens vs Indic-TunedLens: Accuracy Comparison

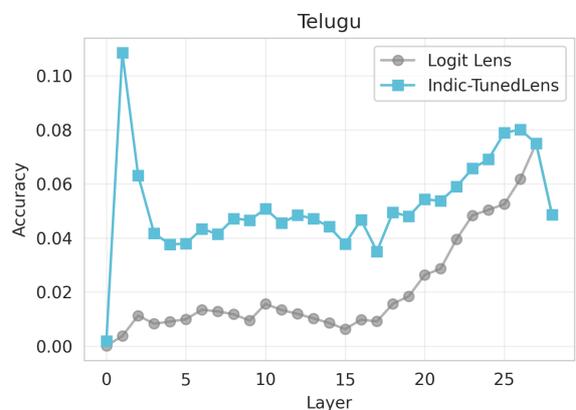


Figure 17: Layer-wise Accuracy Comparison for Telugu

B.2 Language Wise Average Rank of Correct Predictions between Indic-TunedLens and LogitLens

This section provides a detailed examination of the average rank positions assigned to correct tokens across all layers of each language. The average rank metric quantifies how confidently the model places the correct prediction among its top candidates at each layer, with lower ranks indicating higher confidence and better alignment between intermediate representations and final outputs. The average rank of correct predictions is shown for Bengali in Figure 18, English in Figure 19, Gujarati in Figure 20, Hindi in Figure 21, Kannada in Figure 22, Malayalam in Figure 23, Marathi in Figure 24, Nepali in Figure 25, Tamil in Figure 26, and Telugu in Figure 27.

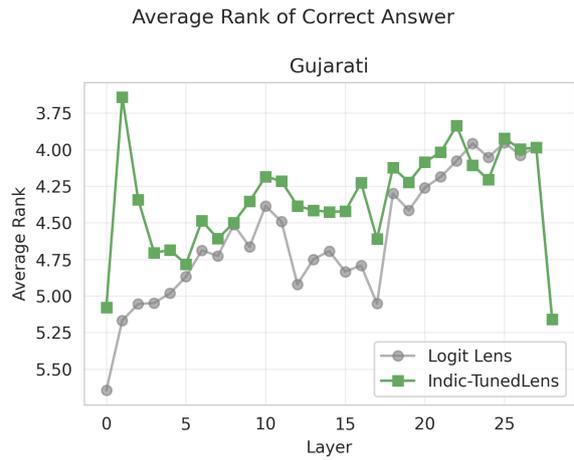


Figure 20: Average Rank of Correct Predictions for Gujarati

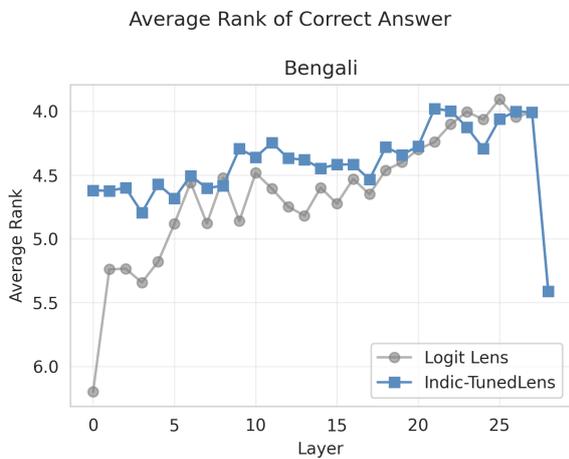


Figure 18: Average Rank of Correct Predictions for Bengali

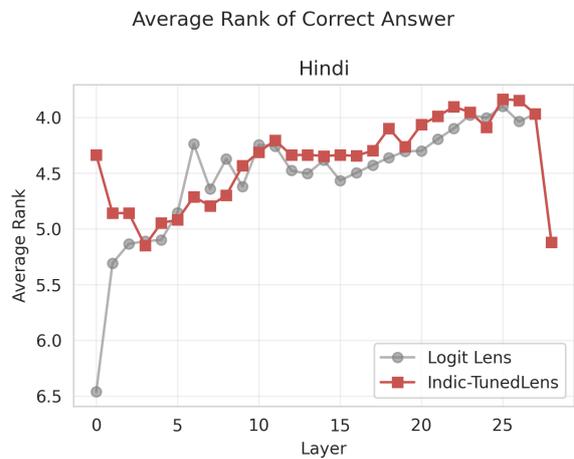


Figure 21: Average Rank of Correct Predictions for Hindi

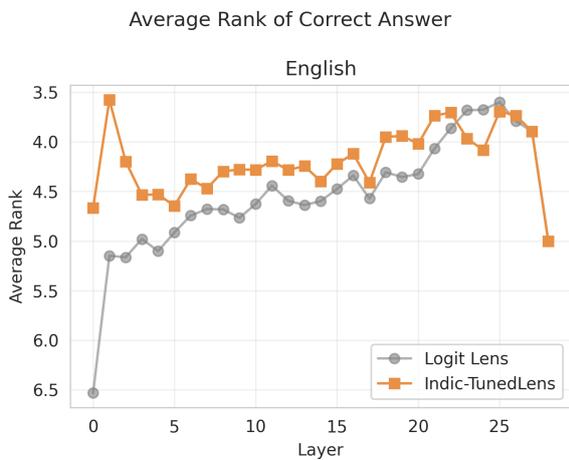


Figure 19: Average Rank of Correct Predictions for English

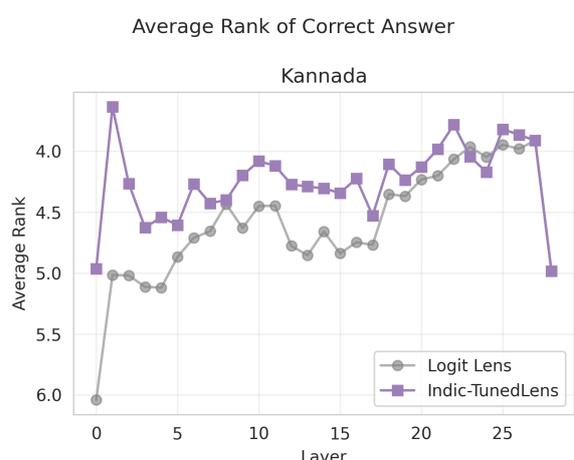


Figure 22: Average Rank of Correct Predictions for Kannada

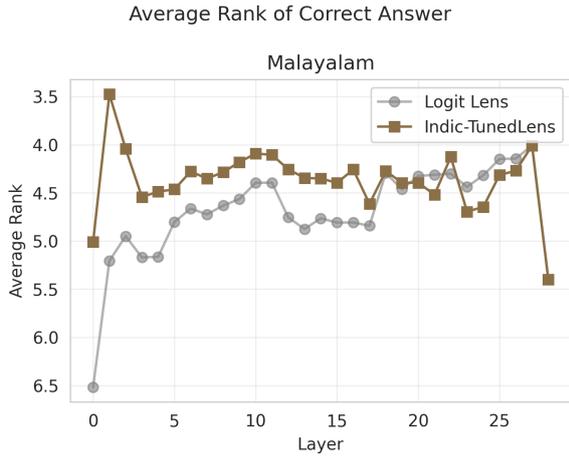


Figure 23: Average Rank of Correct Predictions for Malayalam

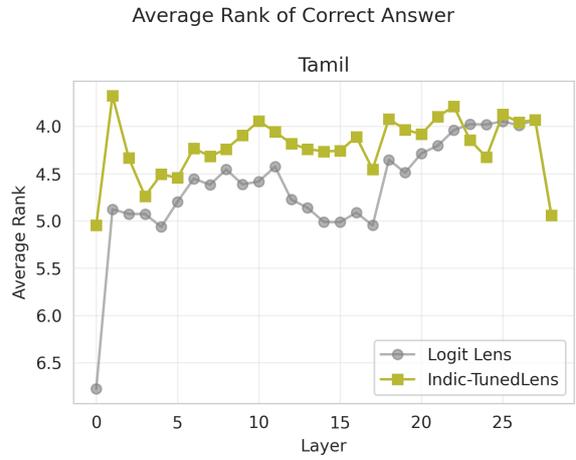


Figure 26: Average Rank of Correct Predictions for Tamil

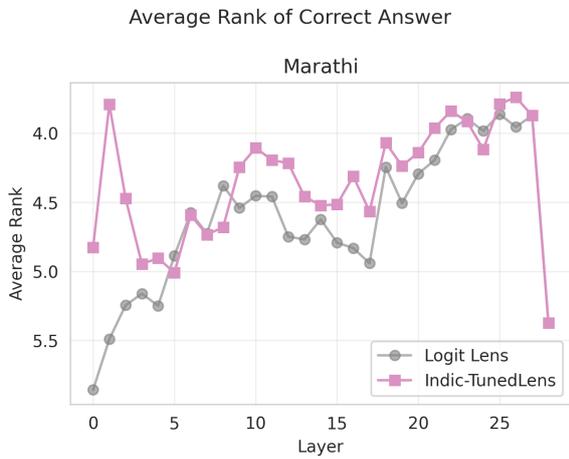


Figure 24: Average Rank of Correct Predictions for Marathi

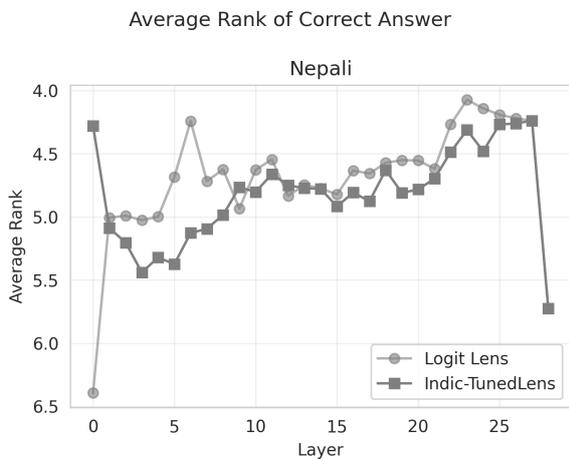


Figure 25: Average Rank of Correct Predictions for Nepali

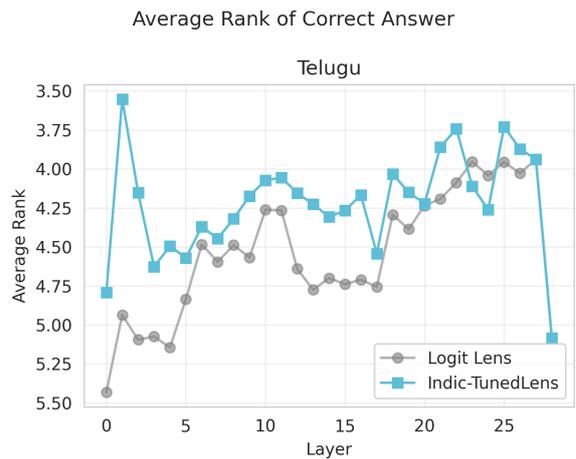


Figure 27: Average Rank of Correct Predictions for Telugu

Building ASR Resources for the Hutsul Dialect of Ukrainian

Roman Kyslyi, Artem Orlovskiy, Pavlo Khomenko, Bohdan Onyshchenko, Zakhar Guzii

Kyiv School of Economics

{rkyslyi, aorlovskiy, pkhomenko, bhnyshchenko, zguzii}@kse.org.ua

Abstract

Dialectal speech remains largely underexplored in Automatic Speech Recognition (ASR) research, particularly for Slavic languages. While Ukrainian ASR systems have rapidly improved in recent years with the adoption of Whisper, XLS-R, and Wav2Vec-based models, performance on dialectal variants remains unknown and often significantly degraded. In this work, we present the first dedicated effort to build ASR resources for the Hutsul dialect of Ukrainian. We develop a data preparation and segmentation pipeline, evaluate multiple forced alignment strategies, and benchmark state-of-the-art ASR models under zero-shot and fine-tuned conditions. We evaluate results using WER and CER demonstrating that large multilingual ASR models struggle with dialectal speech, while lightweight fine-tuning produces substantial improvements. All scripts, alignment tools, and training recipes are made publicly available to support future research on Ukrainian dialect speech.

1 Introduction

Ukrainian NLP has made significant progress in recent years, with new language models, ASR systems, and datasets becoming publicly available (Sereda, 2024). However, the majority of existing work targets standard Ukrainian as used in broadcast media or academic corpora, leaving dialectal variation largely unaddressed (Zhong et al., 2024). Yet dialects remain a central component of Ukrainian linguistic identity, especially in mountainous western regions such as Carpathian mountains, where phonetic and lexical differences from standard Ukrainian are quite huge¹.

Dialectal speech in general introduces challenges for ASR: vowel reduction, consonant softening, archaic lexical forms, code-switching with

other languages (Romanian, Polish, German, etc.) and highly variable pronunciation shaped by geography and speaker generation (Michelsanti and et al., 2019). Zero-shot multilingual models can transcribe such speech, but with significantly reduced accuracy (Adams and et al., 2020). At the same time, collecting labeled dialect data is difficult due to speaker scarcity, limited dialectal written tradition, and available recordings being long-form oral narratives requiring segmentation and alignment (Klejch and et al., 2025).

To address these gaps, we present the first systematic pipeline for one of the dialects (Hutsul²) ASR development, centered around a curated speech corpus derived from the recordings of "Dido Yvanchyk"³ - a unique novel written completely using dialect.

Our choice based first of all on availability of both recorded reading of the novel as well as a textual representation of it.

This paper introduces the dataset, describes alignment methodology, reports initial baselines, and outlines future work including dataset expansion, cross-dialect generalization, and LLM-powered transcript enhancement.

Our contributions in this work are:

1. Building a publicly available pipeline for segmentation, forced alignment, and dataset preparation for Hutsul dialect speech.
2. Evaluation of state of the art ASR models in zero-shot and fine-tuned settings using WER/CER metrics.
3. Release scripts and data, enabling reproducible training and evaluation for future research.

¹https://en.wikipedia.org/wiki/Ukrainian_dialects

²<https://en.wikipedia.org/wiki/Hutsuls>

³https://shron1.chtyvo.org.ua/Shekeryk-Donykiv_Petro/Dido_Yvanchik.pdf

2 Related Work

Research on ASR for low-resource languages and dialects has grown in Arabic, Hindi, Italian, and Turkic languages (Ali and et al., 2014; Kumar and et al., 2025), where phonetic variation and lexical differences significantly impact recognition. Recent regional work also introduced RoDia, a Romanian dialect speech dataset for dialect identification (Rotaru et al., 2024), highlighting increasing interest in dialect-focused speech resources.

In Ukrainian, existing work focuses primarily on standard language, including XLS-R and Wav2Vec2-UA models for broadcast and conversational speech (Paniv, 2023). Community projects have fine-tuned Whisper for Ukrainian (Shas, 2024), but dialect performance remains undocumented. No prior work has addressed Hutsul ASR, making this study the first benchmark of its kind.

Forced alignment for speech corpora is typically performed using Montreal Forced Aligner (McAuliffe et al., 2017) or Aeneas (Read-Beyond, 2015), while WhisperX (Shi and et al., 2023) recently demonstrated strong alignment performance for multilingual speech. Our work compares alignment approaches in a dialect context where standard lexicons may be insufficient.

3 Hutsul Dialect Speech Corpus

The used dataset is primarily based on recordings of a native Hutsul speaker reading "Dido Yvanchyk" novel, the most extensive and culturally important written sources of Hutsul dialect. The novel represents the largest continuous literary record of Hutsul speech, featuring authentic lexical, morphological, and phonetic elements that are rarely present in standard Ukrainian corpora. We use publicly available audio recordings of oral readings of the text, sourced from a native speaker and accessible on YouTube.⁴ The dataset is released on Hugging Face.⁵

Hutsul dialect is quite different comparing to standard Ukrainian due to its geographical location. Here are some of the linguistic Characteristics of Hutsul dialect:

- *Phonetics*: vowel transformations, such as /je/ instead of /a/ or /ja/ (e.g., yak → yek, yahoda → yehoda).

⁴<https://www.youtube.com/@didoyvanchik7322>

⁵<https://huggingface.co/datasets/KSE-RESEARCH-Group/Dido-Yvanchyk-Audio-Dataset-v2>

Property	Value
Total duration	19h 27m
Total samples	8,412
Average segment duration	8.35 s
Median segment duration	6.68 s
Duration range	0.11-78.84 s
Total words	162,204
Unique vocabulary	32,247 words
Average words per segment	19.4
Average characters per segment	108.8
Audio sample rate	16 kHz
Dialect marker coverage	72.4%
Total dialect markers	16,874

Table 1: Dido-Yvanchyk Hutsul Dialect Dataset Summary

- *Morphology*: unique case endings -yed, -si and preserved dual forms of a word apples, e.g., yablutsi instead of plural yabluka.
- *Lexicon*: Romanian, Polish and German borrowings such as brynza (cheese) and spaceruvaty (go for a walk, from German spazieren gehen).⁶

The audio is rich in dialect vocabulary, archaic expressions, and non-standard pronunciation, making it a valuable material for building dialectal ASR systems. Segmentation and alignment were performed on long-form recordings to create paired audio–text samples suitable for model training. All recordings are read by a single native Hutsul speaker, however, careful train–dev–test splitting and augmentation were applied to mitigate speaker memorization effects.

The current release serves as a baseline version of the corpus. In future work, we plan to expand the dataset with additional speakers, spontaneous speech, and regional variation, enabling multi-speaker modeling, dialect classification, and deeper linguistic analysis.

General dataset statistics are summarized in Table 1, with a detailed description provided in Appendix A.

4 Forced Alignment Pipeline

Recordings were normalized and resampled to 16kHz. Before the text–transcript alignment, the

⁶https://en.wikipedia.org/wiki/Eastern_Romance_influence_on_Slavic_languages

audio recordings are automatically transcribed with word-level time stamps. We evaluate two speech-to-text pipelines: WhisperX (Shi and et al., 2023) and the ElevenLabs (ElevenLabs, 2024) speech recognition system.

WhisperX extends Whisper (Radford et al., 2022) with forced alignment, enabling word-level timing extraction from neural acoustic models, while ElevenLabs provides native word-level timestamps as part of its transcription output. In practice, ElevenLabs yields more accurate and stable word boundaries (ElevenLabs, 2024), particularly in conversational and expressive speech, and is therefore used as the primary source of word-level timing information for subsequent alignment and segmentation steps.

We perform sentence-level alignment between ground-truth text and automatic speech transcription using a fuzzy string-matching strategy (Source, 2011). Both the transcript generated by the ASR system and the reference text are first normalized through lowercasing, punctuation removal, and whitespace standardization to reduce superficial mismatches.

For each ground-truth sentence, candidate hypothesis segments are searched within a sliding temporal window, and similarity is computed using token-based fuzzy matching (RapidFuzz (Bachmann, 2020)). The highest-scoring segment above a predefined threshold is selected as the alignment, enabling robust matching even in the presence of transcription errors, paraphrasing, or minor omissions (Gao and et al., 2025; Chen and et al., 2025; Abdjul and et al., 2025). The process is executed in batches while maintaining sentence and segment indices, ensuring sequential consistency and scalability to long recordings.

The selected ASR system produces a continuous transcript with word-level timestamps, which serves as the temporal backbone for subsequent sentence matching. The resulting aligned segments preserve the original temporal ordering of the recording and enable consistent downstream processing, including sentence-level segmentation and corpus construction for dialectal speech analysis.

Pipeline overview:

1. Transcribe raw audio using WhisperX or ElevenLabs STT.
2. Normalize reference text (lowercase, remove

punctuation, unify spacing).

3. Align reference sentences to ASR output using RapidFuzz similarity search.
4. Export word- or sentence-level timestamps in WebVTT/JSON format.
5. Filter low-confidence matches to construct clean training segments.

Initial observations indicate WhisperX yields the most stable results, while Montreal Forced Aligner (McAuliffe et al., 2017) requires dialect lexicon adaptation.

5 ASR Models and Training

We evaluate four ASR model families representing different architectural paradigms and multilingual capabilities:

- **OmniASR** – Meta’s new omnilingual speech model designed for cross-language generalization across hundreds of languages (team et al., 2025).
- **Wav2Vec2-XLSR-300M-UA** – a self-supervised multilingual wav2vec2 model with a Ukrainian CTC head, commonly used in Ukrainian ASR applications (Baevski et al., 2020; Babu and et al., 2021).
- **Whisper-large-v3** – a transformer-based encoder-decoder ASR model trained on 680K+ hours of weakly supervised multilingual data (Radford et al., 2022), widely adopted as a strong zero-shot baseline.
- **Wav2Vec2-BERT-UK-v2.1** – a hybrid SSL+MLM architecture integrating wav2vec2 acoustic features with BERT-style masked language modeling (Smoliakow and et al., 2024).

We select these models to cover most popular models in current ASR research and score them for the dialectal recognitions task (Radford et al., 2022; Baevski et al., 2020; Babu and et al., 2021; Smoliakow and et al., 2024; team et al., 2025).

Whisper serves as a strong zero-shot baseline (Radford et al., 2022), Wav2Vec2-XLSR and Wav2Vec2-BERT represent Ukrainian-centric self-supervised CTC approaches (Baevski et al., 2020; Babu and et al., 2021; Smoliakow and et al., 2024), enabling us to assess how well standard-Ukrainian

systems adapt to dialects, and OmniASR allows us to test whether a unified omnilingual speech encoder can generalize to dialect without prior exposure (team et al., 2025).

This spectrum enables comparison across (i) multilingual vs. Ukrainian-only pre-training, (ii) encoder–decoder vs. CTC decoding, and (iii) zero-shot vs. fine-tuned dialect adaptation (Radford et al., 2022; Baevski et al., 2020).

All models are fine-tuned on the aligned Hutsul dataset using an 80/10/10 train–dev–test split. For Whisper, we apply parameter-efficient LoRA fine-tuning (Hu et al., 2021).

For Wav2Vec2-BERT, lower convolutional and transformer layers are frozen to prevent catastrophic forgetting, while adapter modules and the CTC head remain trainable (Smoliakow and et al., 2024).

OmniASR models are fine-tuned using Meta’s official tri-stage learning rate schedule (team et al., 2025).

Also, given the limited amount of dialectal speech data, we use on-the-fly augmentation during training to improve robustness (Chen and et al., 2025). The pipeline includes Gaussian noise injection, pitch shifting, speed perturbation (0.8–1.2×), gain modulation, and time/frequency masking (Nguyen and et al., 2023). Such augmentation is essential for preventing overfitting and encouraged more stable convergence.

All training runs use the AdamW optimizer with FP16 mixed precision (Baevski et al., 2020). Whisper and Wav2Vec2 models employ a linear warm-up schedule followed by linear decay (Radford et al., 2022), while OmniASR follows a tri-stage scheduler (team et al., 2025). Checkpoints are evaluated every 500–1,000 steps, and the best checkpoint is selected based on development CER. Hyperparameters and training configurations for each model are summarized in next section.

6 Experimental Results

For the training we used mixed-precision training mode, with CER (character error rate) as the metric for selection of the best model. The issue between selecting WER and CER lies down to the nuances of the training ASR for the dialect task, since the changes are mostly in characters (Coll and et al., 2023; Thennal D K, 2024).

Ukrainian is a highly inflected Slavic language with:

- Seven grammatical cases
- Three genders with agreement across adjectives/participles
- Complex verb conjugation
- Extensive prefixation and suffixation

And other morphological complex structures and nuances (Pugh and Press, 2005; Sussex and Cubberley, 2006).

All models were evaluated on the same test split using WER and CER metrics.

Overall, fine-tuning consistently improved recognition quality on Hutsul speech across all architectures, with the best systems reaching sub-3% CER.

6.1 OmniASR

We fine-tuned two OmniASR CTC models released by Meta (300M and 1B parameters).⁷ Training was performed using the official (team et al., 2025) tutorial on the "Dido Yvanchyk" dataset.

Training configuration. Both models were trained with a learning rate $5e^{-5}$ using a tri-stage scheduler. The 300M model was trained on an RTX 5070 Ti (16GB) for 48k steps, and 1B model was trained on an RTX 4090 (48GB) for 36k steps. We used a per-device batch size of 8 with gradient accumulation of 4, and WER served as the primary optimization metric.

Model	WER _b	CER _b	WER _a	CER _a
300M	76.78	37.08	13.82	2.97
1B	80.09	51.24	13.09	2.75

Table 2: OmniASR fine-tuning results on Hutsul dialect (in %). WER_b/CER_b: before fine-tuning; WER_a/CER_a: after fine-tuning.

After fine-tuning, both OmniASR reduced WER from 76–80% down to around 13%, indicating that OmniASR benefits strongly from dialect adaptation.

6.2 Wav2Vec2-XLSR-300M-UA

We fine-tuned the Wav2Vec2-XLSR-300M-UA model (Conneau and et al., 2020; Smoliakow and et al., 2023) on the aligned Hutsul corpus. All audio was resampled to 16 kHz, normalized, and cleaned

⁷<https://huggingface.co/KSE-RESEARCH-Group/omniASR-CTC-300M-uk-dido-tuned-v2> <https://huggingface.co/KSE-RESEARCH-Group/omniASR-CTC-1B-uk-dido-tuned-v2>

using a regular-expression based text preprocessor. A Ukrainian CTC vocabulary was constructed to include all Cyrillic characters and dialectal orthographic forms.

Training setup. The model was trained on an RTX 4090 with an effective batch size of 64 (per-device BS=8, gradient accumulation=8), using FP16 mixed precision, AdamW optimizer, a peak learning rate of 1×10^{-4} , and a linear scheduler with 1,000 warmup steps. Training ran for 50

epochs (5,000 steps total).

Results. The best checkpoint was reached at step 4,700 with validation WER=12.99% / CER=2.53%, and the final test performance achieved **WER=13.61% / CER=2.43%**.

Fine-tuning reduced CER to below 3%, placing Wav2Vec2-XLSR performance close to OmniASR and confirming that Ukrainian-centric SSL models adapt well to dialectal speech.

Model	Base checkpoint	Max steps	Per-dev BS	Grad acc.	Eff. BS	Best-by
Small	openai/whisper-small	8000	4	4	16	CER
Medium	openai/whisper-medium	8000	8	4	32	CER
Large-v3	openai/whisper-large-v3	8000	4	4	16	CER
Large-v2 (UK)	arampacha/whisper-large-uk-2	4000	4	4	16	CER

Table 3: Fine-tuning configuration for Whisper models.

6.3 Whisper Family

We have fine-tuned four Whisper variants: whisper-small, whisper-medium, whisper-large-v3, and arampacha/whisper-large-uk-2, the latter already adapted to Ukrainian using Common Voice 11.0 (Ardila et al., 2022). We release our fine-tuned checkpoints on Hugging Face.⁸ Training was performed in mixed-precision mode, using CER as the selection criterion, as character-level variation better captures dialectal orthography and subword differences. The inflectional nature of Ukrainian further motivates character-aware adaptation for dialect modeling.

Training setup. All Whisper models were trained for up to 8,000 steps (4,000 for the Common Voice-adapted checkpoint). Batch size, gradient accumulation, and max steps per model are summarized in Table 3. Learning rate schedules and optimization configurations were kept consistent across models to enable fair comparison.

Results. Final WER/CER performance is shown in Table 4. Both whisper-large-v3 and the Ukrainian-adapted whisper-large-v2 achieved sub-4% CER and approximately 13% WER on the test set, outperforming the medium and small

variants.

Whisper fine-tuning consistently improved performance across all model sizes, with the large Ukrainian-adapted checkpoint achieving the best results.

6.4 Wav2Vec2-BERT

We additionally evaluated the Wav2Vec2-BERT architecture (Hsu et al., 2021; Baevski et al., 2020) using a Ukrainian-pretrained checkpoint⁹. The model was fine-tuned with adapters enabled, while freezing the feature encoder and BERT backbone to prevent catastrophic forgetting. Parameter-efficient training was selected following prior work on adapter-based optimization (Houlsby et al., 2019; Hu et al., 2021). A stronger augmentation pipeline was applied to increase robustness given the limited volume of dialectal speech data.

Results. Adapter-based fine-tuning achieved a test performance of **WER=18.24% / CER=3.47%**. In contrast, LoRA-based training converged less reliably, yielding **WER=43.08% / CER=9.93%** despite similar optimization settings. This suggests that adapter-based optimization is more suitable for this model family in low-resource dialect scenarios.

6.5 Overall Results Summary

Across all model families, fine-tuning led to large improvements in recognition quality on Hutsul speech. The strongest systems—OmniASR 1B, Whisper-Large-v3, Whisper-Large-v2-UK and

⁸<https://huggingface.co/KSE-RESEARCH-Group/whisper-small-dido-yvanchyk-v2> <https://huggingface.co/KSE-RESEARCH-Group/whisper-medium-dido-yvanchyk-v2> <https://huggingface.co/KSE-RESEARCH-Group/whisper-large-v3-dido-yvanchyk-v2> <https://huggingface.co/KSE-RESEARCH-Group/arampacha-whisper-large-v2-dido-yvanchyk-v2>

⁹<https://huggingface.co/Yehor/w2v-bert-uk-v2.1>

Model	Val CER↓	Val WER↓	Test CER↓	Test WER↓	Best ckpt
Small	4.71	17.60	4.72	17.84	6500
Medium	4.06	13.80	3.96	14.61	7000
Large-v3	3.73	12.40	3.90	13.20	7000
Large-v2 (UK)	3.73	12.60	3.69	13.03	4000

Table 4: Whisper ASR performance after fine-tuning. Lower is better.

Metric	Original	Finetuned	Improvement
Word Error Rate (WER)	60.79%	14.32%	↓ 76.4%
Character Error Rate (CER)	17.86%	5.14%	↓ 71.2%
Perfect Transcriptions	2 (0.2%)	217 (25.8%)	↑ 108×
Samples with Errors	840 (99.8%)	625 (74.2%)	↓ 25.6%

Table 5: Overall Performance Comparison: Original vs Finetuned Whisper-Large-V3

Wav2Vec2-XLSR—achieved **CER < 3%** on the test set, showing that high-quality ASR for a low-resource dialect can be obtained using a relatively small aligned corpus.

Figure 1 and Figure 2 visualize convergence for Whisper models, where larger architectures show a faster drop in CER and more stable late-stage training.

Models and resources. All trained models and datasets described are publicly released on Hugging Face: <https://huggingface.co/KSE-RESEARCH-Group>.

6.6 Qualitative Example Analysis

To illustrate model behavior beyond aggregate metrics, we provide qualitative recognition examples for OmniASR models, including Cyrillic text with transliteration. These examples highlight typical recognition patterns, including character alignment with occasional lexical substitutions and vowel variation.

This shows that most errors relate to dialect-specific morphology and orthographic variation rather than acoustic confusion. The complete samples are provided in Appendix B.

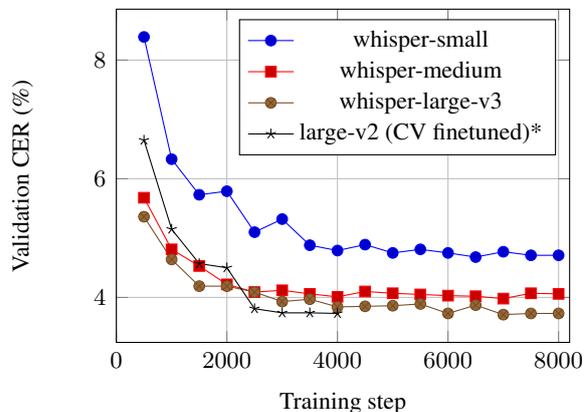


Figure 1: Validation CER over fine-tuning steps (evaluated every 500 steps). * – arampacha/whisper-large-v2, previously fine-tuned on Common Voice 11.0.

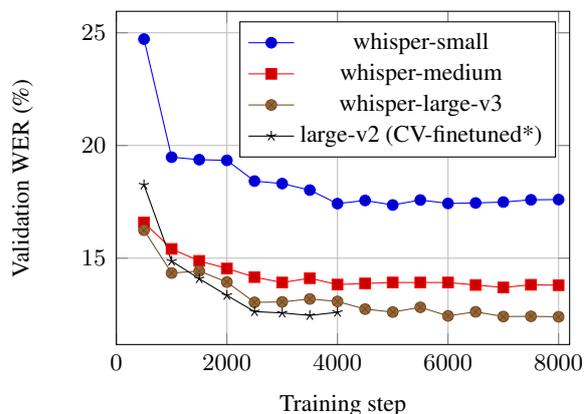


Figure 2: Validation WER over fine-tuning steps (evaluated every 500 steps). * – arampacha/whisper-large-v2, previously fine-tuned on Common Voice 11.0.

7 Discussion

Our results show that modern ASR models can adapt well to Ukrainian dialectal speech when fine-tuned on even relatively small aligned data. All tested model families improved substantially compared to zero-shot recognition, and (Omni-ASR 1B, Whisper-Large-v3, Whisper-Large-v2-UK, Wav2Vec2-XLSR) reached CER below 3%. This suggests that large multilingual and Ukrainian-centric self-supervised models can be effectively used for dialect ASR with limited supervision (Baevski et al., 2020; Shas, 2024; team et al., 2025).

During qualitative inspection, we observed that most remaining errors come from dialect-specific morphology, rare lexemes, vowel reduction, and non-standard spelling. Small changes in endings or palatalization often result in substitutions. This indicates that future work could explore phoneme-aware decoding, lexicon extension, or LM rescoreing to reduce errors caused by dialectal variation, as also noted for other Slavic speech systems (Sussex and Cubberley, 2006).

Although this work focuses on the Hutsul dialect, the methodology is general and can be applied to other Ukrainian regional varieties. The alignment pipeline, training scripts and evaluation setup can serve as a template for expanding research to other dialects. In future, combining several dialect corpora may allow dialect classification, cross-dialect transfer, and speech-style robustness studies.

The released dataset and pipeline represent a first step toward more complete Ukrainian dialect ASR resources. As new recordings are collected, it will be important to ensure balanced speaker representation, informed consent, and fair coverage across regions.

8 Conclusion

We present the first speech corpus with aligned text for the Hutsul dialect and evaluated four modern ASR model families on this data. Fine-tuned Whisper, OmniASR and Wav2Vec2-XLSR models achieved strong accuracy with CER below 3%, showing that high-quality recognition of Ukrainian dialectal speech is possible even with a single-speaker corpus when alignment and augmentation are applied effectively. Adapter-based training also improved Wav2Vec2-BERT, although it remained behind the top systems.

While our experiments focus on Hutsul, the same training pipeline can be applied to other dialects,

enabling work on multi-dialect modeling, dialect identification and cross-dialect transfer. In future we plan to extend the corpus with more speakers, spontaneous speech and additional regions, as well as explore phoneme-level decoding and LM rescoreing for better handling of dialect-specific forms.

We release the dataset, code and splits to facilitate reproducibility and further research on Ukrainian dialect ASR. We believe this work provides a starting point for building larger dialect resources and contributes toward speech technologies for low-resource Slavic varieties.

Limitations

This work represents an initial step toward ASR for Ukrainian dialects and carries a lot of limitations.

First, the current corpus is based on recordings from a small number of speakers, from a single Hutsul region, which may restrict dialectal and acoustic diversity. Broader demographic (age, gender, speaking style, recording conditions) will be necessary to ensure robust generalization.

Second, our training pipeline relies on automatic forced alignment. Although it is effective, alignment errors sometimes occurs in the fine-tuning data and may influence on achievable performance.

Third, we evaluate only end-to-end CTC and encoder–decoder architectures, leaving LM-rescoreing, shallow fusion, and hybrid ASR systems for future study.

Finally, evaluation is limited to WER and CER, without semantic or intelligibility assessments, which would provide a more complete measure of transcription quality.

As such, results should be interpreted as a strong baseline rather than a fully comprehensive solution for Hutsul or broader Ukrainian dialect recognition.

Ethics

We fully acknowledge the ACL Ethics Policy and commit to responsible research practice, including careful consideration of consent, potential harms, and responsible data release.¹⁰

Data provenance and consent. The Dido-Yvanchyk corpus is derived from publicly available audio recordings of a single narrator reading “*Dido Yvanchik*”, published on YouTube.¹¹ The literary work “*Dido Yvanchik*” by Petro Shekeryk-Donykiv

¹⁰https://www.aclweb.org/adminwiki/index.php/ACL_Policy_on_Publication_Ethics

¹¹<https://www.youtube.com/@didoyvanchik7322>

is a canonical text of Hutsul cultural heritage and is available in the public domain. The recordings were produced by the Ukrainian Cultural Fund,¹² which releases its materials under an open license permitting use for research purposes.¹³ Accordingly, the use of these recordings complies with applicable consent and data usage requirements for academic research.

Acknowledgments

We thank the Kyiv School of Economics (KSE) for institutional support and for providing computational resources used in this work. We also thank Vasyl Zelenchuk for his work on the reading "Dido Yvanchyk" and providing it to the public domain.¹⁴

References

- Rifqi Naufal Abdjul and et al. 2025. Indonesian speech content de-identification in low resource transcripts. In *Proceedings of SEALP 2025*.
- Oliver Adams and et al. 2020. A study of multidialect speech recognition with massively multilingual models. *Preprint*, arXiv:2007.03001.
- Ahmed Ali and et al. 2014. Arabic speech recognition for iwslt: Dialectal challenges and system description. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*.
- Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Natalia Meyer, Joelle Stevens, Michael Henretty, Lindsay Morais, Cătălina Saunders, Wan Xi Chua, and 1 others. 2022. Mozilla common voice corpus 11.0. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*. ELRA. A large-scale multilingual speech dataset.
- Arun Babu and et al. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.
- Max Bachmann. 2020. Rapidfuzz: A fast fuzzy string matching library. <https://github.com/maxbachmann/RapidFuzz>.
- Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *NeurIPS*.
- Li Chen and et al. 2025. Fuzzyseg: Segmentation and alignment for long-form speech. *Preprint*, arXiv:2505.15646.
- ¹²<https://ucf.in.ua/>
- ¹³https://ucf.in.ua/storage/docs/06012025/CUW%202025%20%D0%9B%D0%9E%D0%A2%201_4ca00d7c2239c56085a20d4f869de9cd4cb2b9fe.pdf
- ¹⁴<https://www.youtube.com/edidoyvanchik7322>
- Albert Coll and et al. 2023. Character-based or subword-based sequence labeling? a comparative study. In *Proceedings of CoNLL 2023*.
- Alexis Conneau and et al. 2020. Unsupervised cross-lingual representation learning for speech recognition. In *Interspeech*.
- ElevenLabs. 2024. Elevenlabs speech-to-text api documentation. <https://elevenlabs.io/docs/api/speech-to-text>.
- ElevenLabs. 2024. Meet scribe: Word error rates across 102 languages. <https://elevenlabs.io/blog/meet-scribe#fleurs-word-error-rate-102-languages>. Accessed: 2025-01-XX.
- Xing Gao and et al. 2025. Dynamic alignment for speech and text in low-resource settings. *Preprint*, arXiv:2509.24478.
- 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 *Proceedings of the 36th International Conference on Machine Learning (ICML)*.
- Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhota, and Ruslan Salakhutdinov. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Chen, Yu Li, Sijia Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint*, arXiv:2106.09685.
- Ondřej Klejch and et al. 2025. A practitioner's guide to building asr models for low-resource languages: A case study on scottish gaelic. *Preprint*, arXiv:2506.04915.
- Ashish Kumar and et al. 2025. Hindi dialect speech recognition in low-resource settings. *Preprint*, arXiv:2507.15272.
- Michael McAuliffe, Michaela Socolof, Steven Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kald. In *Proc. Interspeech*, pages 498–502.
- Davide Michelsanti and et al. 2019. Asr for unwritten languages: A survey. *Preprint*, arXiv:1907.13511.
- Hieu Nguyen and et al. 2023. Investigating the impact of data augmentation techniques for low-resource asr. *Preprint*, arXiv:2307.07948.
- Yurii Paniv. 2023. Ukrainian-tts: An open text-to-speech system for ukrainian. GitHub repository.

- Stefan Pugh and Ian Press. 2005. *Ukrainian: A Comprehensive Grammar*. Routledge.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. [Robust speech recognition via large-scale weak supervision](#). *arXiv preprint arXiv:2212.04356*.
- ReadBeyond. 2015. [Aeneas: Audio-text forced alignment toolkit](#). GitHub repository.
- Codruț Rotaru, Nicolae Ctilin Ristea, and Radu Tudor Ionescu. 2024. [RoDia: A new dataset for Romanian dialect identification from speech](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 279–286, Mexico City, Mexico. Association for Computational Linguistics.
- Taras Sereda. 2024. [Transcribe, align and segment: Creating speech datasets for low-resource languages](#). *Preprint*, arXiv:2406.12674.
- Anurag Shas. 2024. [Whisper-large-v2 fine-tuned for ukrainian speech recognition](#). HuggingFace model.
- Bowen Shi and et al. 2023. [Whisperx: Time-accurate speech transcription with word-level alignment](#). *Preprint*, arXiv:2303.00747.
- Yehor Smoliakow and et al. 2023. [Ukrainian xls-r speech model](#). Ukrainian adaptation of XLS-R pre-trained on Common Voice.
- Yehor Smoliakow and et al. 2024. [Wav2vec2-bert-uk-v2.1: Ukrainian wav2vec2 + bert integrative model](#). HuggingFace model.
- SeatGeek Open Source. 2011. [Fuzzywuzzy: Fuzzy string matching in python](#). <https://github.com/seatgeek/fuzzywuzzy>.
- Roland Sussex and Paul Cubberley. 2006. *The Slavic Languages*. Cambridge University Press.
- Omnilingual ASR team, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, Kevin Chan, Chierh Cheng, Joe Chuang, Caley Droof, Mark Duppenthaler, Paul-Ambroise Duquenne, Alexander Erben, Cynthia Gao, Gabriel Mejia Gonzalez, and 14 others. 2025. [Omnilingual asr: Open-source multilingual speech recognition for 1600+ languages](#). *Preprint*, arXiv:2511.09690.
- Jesin James Thennal D K. 2024. [Advocating character error rate for multilingual asr evaluation](#). <https://arxiv.org/abs/2410.07400>. ArXiv:2410.07400.
- Tianyu Zhong, Ziqi Yang, Zhen Liu, Rui Zhang, Yiheng Liu, Hanqi Sun, Yujia Pan, Yiming Li, and Yifan Zhou. 2024. [Opportunities and challenges of large language models for low-resource languages in humanities research](#). *arXiv preprint arXiv:2412.04497*.

Split	Mean	Median	Std	Min	Max	P95
Train	8.31	6.64	6.70	0.11	78.84	21.13
Validation	8.27	6.63	6.36	0.62	64.44	19.24
Test	8.47	6.76	6.72	0.11	42.98	21.79
Overall	8.35	6.68	6.59	0.11	78.84	20.72

Table 6: Audio duration statistics (in seconds)

Split	Characters	Words	Unique	Mean len
Train	736,512	131,142	21,816	108.1
Validation	81,715	14,582	4,936	107.9
Test	92,900	16,480	5,495	110.3
Total	911,127	162,204	32,247	108.8

Table 7: Text transcription statistics

A Dataset Characteristics

This section provides additional descriptive statistics and qualitative error analysis for the Hutsul speech corpus used in our experiments.

Tables 6 and 7 provide supplementary statistics for the dataset.

Table 7 reports transcription-level statistics for each data split, including character and word counts, vocabulary size, and average segment length.

A.1 Error Pattern Analysis

Character-level errors are dominated by vowel substitutions. The most frequent patterns include $i(i) \rightarrow u(y)$ (81), $u(y) \rightarrow i(i)$ (51), $y(u) \rightarrow B(v)$ (26), and $e(e) \rightarrow \epsilon(ye)$ (16), reflecting dialectal vowel variation and non-standard orthography.

Consonant-level errors are less frequent and include alternation between $y(u)$ and $B(v)$, voicing changes such as $\exists(z) \leftrightarrow c(s)$ and $\text{д}(d) \leftrightarrow \text{т}(t)$, as well as substitutions involving the Hutsul-specific consonant $r(g)$.

Vowel substitutions represent the most frequent error type overall. The dominant pattern is alternation between $u(y)$ and $i(i)$, reflecting the characteristic phenomenon in Hutsul speech. Other frequent changes include $e(e) \leftrightarrow u(y)$, $e(e) \rightarrow \epsilon(ye)$, and $a(a) \leftrightarrow o(o)$, consistent with known dialectal variation.

B Qualitative Recognition Examples

Below we show sample recognition outputs for OmniASR 300M and 1B models. Cyrillic text is shown with English transliteration.

OmniASR 300M

- “Жеріт, жер би вас гад, йкісте бешшесно проїсні.”
Zheryt, zher by vas had, ykiste beshshesno projisni.
“жеріт жер би вас гадий кісте бешесно про єсні”
Zheryt zher by vas hadyj kiste beshesno pro jesni.
WER=0.63, CER=0.11
- “Нима страху, ни завіситси».”
Nυμα strachu, ny zavisytsi.
“німа страху низавісиц”
Nυμα strachu nuzavisyts.
WER=0.50, CER=0.17

OmniASR 1B

- “Жеріт, жер би вас гад, йкісте бешшесно проїсні.”
Zheryt, zher by vas had, ykiste beshshesno projisni.
“жеріт жерби вас гади йкійсте бешесно проєсні”
Zheryt zherby vas hady ykiiste beshesno pro-jesni.
WER=0.75, CER=0.11
- “Нима страху, ни завіситси».”
Nυμα strachu, ny zavisytsi.
“німа страху ни завіситс”
Nυμα strachu ny zavisyts.
WER=0.25, CER=0.04

From FusHa to Folk: Exploring Cross-Lingual Transfer in Arabic Language Models

Abdulmuizz Khalak Abderrahmane Issam Gerasimos Spanakis

Department of Advanced Computing Sciences
Maastricht University

{a.khalak@student., abderrahmane.issam@, jerry.spanakis@}maastrichtuniversity.nl

Abstract

Arabic Language Models (LMs) are pretrained predominately on Modern Standard Arabic (MSA) and are expected to transfer to its dialects. While MSA as the standard written variety is commonly used in formal settings, people speak and write online in various dialects that are spread across the Arab region. This poses limitations for Arabic LMs, since its dialects vary in their similarity to MSA. In this work we study cross-lingual transfer of Arabic models using probing on 3 Natural Language Processing (NLP) Tasks, and representational similarity. Our results indicate that transfer is possible but disproportionate across dialects, which we find to be partially explained by their geographic proximity. Furthermore, we find evidence for negative interference in models trained to support all Arabic dialects. This questions their degree of similarity, and raises concerns for cross-lingual transfer in Arabic models.¹

1 Introduction

Arabic is a major world language with over 400 million speakers and a central role in cultural, historical, and religious life across the Middle East and North Africa (MENA) (Owens, 2006; Versteegh, 2014). Structurally, it is a *Semitic* language with a long written tradition, but for NLP the most salient property is its pervasive *diglossia*: a standardized written variety, Modern Standard Arabic (MSA), coexists with a rich spectrum of non-standardized dialects that vary across the MENA region (Ferguson, 1959). MSA (FusHa) dominates education, news, and formal writing, whereas dialects are the default medium of everyday communication and online interaction.

Although Modern Standard Arabic (MSA) is far less frequent in daily speech than Dialectal Arabic (DA), it remains the dominant variety in the

digital corpora used to train Large Language Models (LLMs). Consequently, most existing Arabic LMs are predominantly MSA-centric, with limited and disproportionate coverage of diverse dialects (Antoun et al., 2020; Inoue et al., 2021; Abdulmageed et al., 2021; Antoun et al., 2021; Sengupta et al., 2023; Lan et al., 2020). While these models are often implicitly expected to generalize across the Arabic continuum, cross-lingual transfer is not guaranteed, particularly as many dialects exhibit low mutual intelligibility with MSA and one another (Abu Farha and Magdy, 2022; Kegel et al., 2023). This linguistic imbalance has spurred the development of dialect-specific models (Qarah, 2024b; Gaanoun et al., 2025; Abdaoui et al., 2021; Qarah, 2024a; Shang et al., 2025; AlYami and Al-Zaidy, 2022). However, the necessity of such specialization remains an open question: **Do MSA-centric or multi-dialect models transfer equitably across the Arabic dialectal landscape?** To investigate this, we probe the internal representations of Arabic LMs to evaluate their transferability to various dialects across three core NLP tasks—Sentiment Analysis (SA), Named Entity Recognition (NER), and Part-of-Speech (POS) tagging—benchmarking their performance against specialized, dialect-specific counterparts.

While probing provides a functional account of cross-lingual transfer, its scope is often constrained by specific task selections and potential confounding dataset characteristics. Prior research has leveraged annotated linguistic features to predict inter-lingual transfer or interference (Lin et al., 2019; Eronen et al., 2023); however, the scarcity of granular linguistic annotations for Arabic dialects renders such approaches currently unfeasible. While Alsudais et al. (2022) utilized lexical overlap within the parallel MADAR corpus (Bouamor et al., 2018) as a proxy for dialectal similarity, surface-level overlap fails to adequately capture deeper syntactic and semantic nuances. Given the proficiency of

¹Our code: https://github.com/muizzkhalak/cross_lingual_transfer_arabic

LMs in encoding these high-level features (Conneau et al., 2018a), we propose employing Representational Similarity Analysis (RSA) between MSA and DA model representations. Specifically, we utilize Centered Kernel Alignment (CKA) (Kornblith et al., 2019) to quantify layer-wise similarity between DA and MSA models using parallel sentences from the MADAR dataset. We posit that this provides an intrinsic metric of cross-lingual transfer: high representational similarity between MSA and DA models serves as a robust signal for effective transferability and suggests that inter-dialectal interference will be minimized.

We integrate these two methodologies—Probing and Representational Similarity Analysis (RSA)—to provide a holistic evaluation of cross-lingual transfer between MSA and its dialects. Our findings indicate that, in general, MSA-centric models demonstrate strong transferability to DA, occasionally outperforming dialect-specific models. However, this transfer remains significantly disproportionate across the dialectal spectrum. To investigate the drivers of this disparity, we correlate model performance with factors such as geographic proximity and pretraining data volume. Our analysis reveals that dialect-specific models consistently exceed the performance of MSA and multi-dialect models only when supported by substantial DA pretraining data. Furthermore, geographic proximity serves as a strong predictor of transferability, aligning with the dialectal continuum hypothesis: as geographic distance increases, mutual intelligibility diminishes, posing a greater challenge for MSA-to-DA transfer.

2 Related Work

Cross-Lingual Transfer. Prior work consistently shows that transfer effectiveness depends strongly on source–target similarity: typologically or linguistically closer languages yield better zero-shot transfer, and language-distance metrics correlate with downstream accuracy (Eronen et al., 2023; Philippy et al., 2023). Consequently, selecting a related transfer source—rather than defaulting to English—can substantially improve performance for low-resource targets (Lin et al., 2019; Eronen et al., 2023). Multilingual LMs such as mBERT further demonstrate robust cross-lingual generalization without explicit alignment, achieving competitive zero-shot results across tasks like Natural Language Inference (NLI) and NER after fine-

tuning on a single language (Conneau et al., 2018b; Pires et al., 2019; Wu and Dredze, 2019). This behavior is often attributed to increasingly language-agnostic representations in higher layers (K et al., 2020; Artetxe et al., 2020); notably, retraining only the embedding layer of a monolingual transformer can effectively adapt it to a new language, suggesting that upper-layer features are largely transferable (Artetxe et al., 2020). However, multilingual training often leads to negative interference especially for high resource languages (Bapna and Firat, 2019; Chang et al., 2024; Alastruey et al., 2025). Overall, these findings motivate balanced transfer strategies, where leveraging one or a few related high-resource languages yields strong gains for low-resource targets without incurring excessive interference (Seto et al., 2025). In this study, we examine how well MSA-centric and multi-dialect Arabic models transfer to DA varieties, and whether negative interference occurs in dialectal context, even when linguistic similarity is assumed. Although our process follows cross-dialectal transfer across Arabic models, we use "cross-lingual" in the rest of the paper to align with literature norms.

Probing Arabic Language Models. Previous work has examined how linguistic information is distributed across layers and neurons in Arabic transformer models, including dialect-trained variants (Abdelali et al., 2022). This line of research reports that lower and middle layers primarily encode morphology, upper layers capture syntactic dependencies, and MSA-based models often fail to represent dialect-specific nuances despite substantial vocabulary overlap; embedding-layer neurons tend to be polysemous, whereas mid-layer neurons are more specialized for particular linguistic properties (Abdelali et al., 2022). By contrast, we focus on model-level representations: we assess how MSA, general-purpose Arabic (MIX/Multi-DA), and dialect-specific encoders encode morphological and syntactic information via controlled probing tasks, comparing best-layer performance across models rather than inspecting individual neurons.

Similarity of Arabic Dialects. Recent work in Arabic dialectometry links textual similarity to geography, showing that geographically proximate varieties tend to be linguistically closer, sometimes more so than political boundaries (Alsudais et al., 2022). Across dialectal corpora, classic NLP/IR similarity measures further reveal strong Levantine cohesion and suggest that Palestinian is par-

ticularly close to MSA (Kwaik et al., 2018), while comparisons with Classical Arabic provide a historical reference for modern variation (Abouzahr and Abouzahr, 2025). Analyses of the MADAR corpus also report Muscat (Spoken in Oman) as among the closest dialects to MSA and Sfax (Spoken in Tunisia) among the most distant (Bouamor et al., 2018). These lexical analyses remain limited at capturing syntactic and semantic similarity, therefore, our work studies whether such proximity relationships are internalized by pre-trained language models, which were shown to capture syntactic and semantic features.

3 Methodology

In this section, we outline our methodological framework for studying cross-lingual transfer of Arabic LMs. We combine two complementary methods for this study, namely: linear probes trained on frozen layer-wise embeddings (Section 3.1), and similarity of model representations measured using Centered Kernel Alignment (CKA) (Section 3.2). Furthermore, motivated by the linguistic theory of dialect continuum and its implications (Chambers and Trudgill, 1998), we propose a proxy for measuring geographic proximity to MSA (Section 3.3).

3.1 Probing Classifiers

Probing has been widely applied to analyze multilingual models, where it can reveal whether models encode language-specific features or universal cross-linguistic structures (Tenney et al., 2019b). In the case of Arabic, probing classifiers can help determine whether dialectal and MSA models capture similar linguistic phenomena and whether knowledge acquired in one variety (e.g., MSA) transfers to another (e.g., Egyptian or Levantine Arabic). By training probes on embeddings derived from dataset of one variety using a language model of different variety.

The standard probing setup involves training a lightweight classifier—commonly a linear or shallow feedforward model—on top of frozen representations extracted from different layers of the LM. We train probes to predict a supervised linguistic property (e.g., POS tags) using embeddings as input. If the probe achieves high accuracy, this suggests that the relevant linguistic information is encoded in the representations at that layer and is easily extractable by a simple function (Hewitt

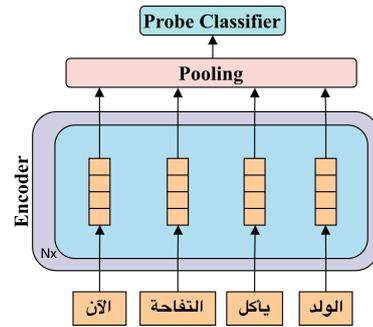


Figure 1: Architecture of the probing classifier for the example sentence “The boy is eating the apple now.” Sentence representations pass through N layers, and each layer is probed using the classifier in Eq. 1.

and Manning, 2019). Probing classifiers are deliberately kept simple to minimize the chance that the probe itself learns the task from scratch. Instead, their purpose is to act as a diagnostic tool that reveals the presence (or absence) of linguistic features in the LM representations (Pimentel et al., 2020). Figure 1 illustrates the probing architecture, where we employ a linear classifier with a softmax activation layer. For a given layer l , the probability distribution y_l is computed as:

$$y_l = \text{softmax}(Wz_l + b) \quad (1)$$

where z_l is a representation derived from the hidden states $H_l = \{h_{l,1}, \dots, h_{l,T}\}$. We define two pooling strategies depending on the probing task:

- **Token-level** to POS and NER tasks: $z_l = h_{l,i}$, the embedding at index i .
- **Sentence-level** for the SA task: $z_l = \frac{1}{T} \sum_{i=1}^T h_{k,i}$, the mean-pooled average of all token embeddings in the sequence.

3.2 Representation Similarity Analysis

Besides cross-lingual transfer as measured using probing on selected tasks, we also compute representation similarity, which can be viewed as an intrinsic signal for transfer. Primarily, we seek to study whether MSA LMs learn similar representations to their DA counterpart, which can be seen as a signal of similarity between DA varieties and MSA, and of the sufficiency of MSA for capturing the nuances of DA varieties.

Figure 2 illustrates our architecture for representation similarity, where we use Centered Kernel Alignment (CKA) (Kornblith et al., 2019) to measure the similarity of hidden representations. Let $X \in \mathbb{R}^{nd_x}$ and $Y \in \mathbb{R}^{nd_y}$ be representations of the

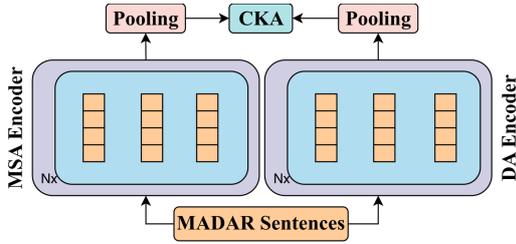


Figure 2: Architecture of CKA for representation similarity. MADAR parallel sentences are encoded by MSA and DA encoders through N layers, and the resulting representations are compared using linear CKA (Eq. 2).

same n inputs. For linear CKA, we set $K = XX^\top$ and $L = YY^\top$ and compute:

$$\text{CKA}(X, Y) = \frac{\|X^\top Y\|_F^2}{\|X^\top X\|_F \|Y^\top Y\|_F} \in [0, 1], \quad (2)$$

where higher values indicate the representations are similar. Linear CKA is invariant to orthogonal transformations and isotropic rescaling of features, making it robust to rotations and global scaling differences between models.

This yields a compact and task-agnostic view of how similarly MSA and dialectal models encode Arabic input, complementing our probing-based analysis of functional transfer.

3.3 Evaluating Dialectal Continuum

We test whether similarity measures track geographical proximity continuum, i.e., whether models associated with countries closer (geographically) to a chosen MSA anchor exhibit higher similarity to an MSA-pretrained encoder. Similar approaches have been used for finding a relationship between language distance and cross-lingual transfer (Eronen et al., 2023; Philippou et al., 2023). A key challenge is that MSA has no fixed geographic locus, as it is used in formal contexts across all Arab countries. The literature proposes several proxies: one strand argues that MSA is particularly close to Palestinian (Levantine) Arabic, based on lexical similarity within Levantine and North-African varieties (Kwaik et al., 2018), however, this study did not include Gulf dialects. Other work highlights the strong affinity between Classical Arabic (CA) and Yemeni Arabic, citing the preservation of archaic forms, phonological features, and relative isolation, and notes that Gulf Arabic, despite borrowings (e.g., from Persian), maintains close ties to Bedouin dialects (Boyi et al., 2024). A separate lexical / lexical-semantic comparison

with CA across regions also identifies Yemeni Arabic as the closest modern variety (Abouzahr and Abouzahr, 2025); although the authors attribute this in part to corpus-size bias for Yemeni data, they also show that Gulf Arabic—geographically close to Yemen—has high similarity to CA. Finally, analysis of the MADAR corpus reports that Muscat (Oman) is the dialect most similar to MSA (Bouamor et al., 2018), and Muscat is likewise geographically near Yemen. Given the widely attested closeness of MSA to CA, we therefore **use Yemen as a geographic proxy for the MSA anchor**, while explicitly acknowledging this as an operational choice.

4 Experimental Setup

4.1 Probing Datasets

For fair cross-dialect comparison, we select dialectal datasets per task and enforce equal sample sizes per probe: for **POS/NER** we balance by number of labeled tokens, and for **SA** by number of sentences. General-purpose Arabic models are consistently evaluated against dialect-specific models on the latter’s native test sets. Table 2 details the corpora utilized for each dialect; for multi-dialectal datasets, we partition and reuse relevant subsets for their corresponding varieties. Due to the absence of publicly available NER corpora for Gulf dialects (specifically Riyadh and Muscat), we derive a silver-standard dataset from the MADAR corpus (Bouamor et al., 2018). We extract 2,000 sentences per dialect and perform automated entity annotation using the **CAMEL-Lab/bert-base-arabic-camelbert-da-ner** model (Inoue et al., 2021).

For the POS tagging task, we apply preprocessing tailored to Arabic cliticization, where multiple grammatical units are written as a single orthographic word. For example, **وموبايل** (“**and mobile**”) combines **و** (*CONJ*) and **موبايل** (*NOUN*); we normalize such cases by aggregating clitic and stem tags into composite labels (e.g., *CONJ+NOUN*), following prior work in Arabic POS tagging (Darwish et al., 2018; Habash and Rambow, 2006). NER datasets are used with their original tags and no additional preprocessing. For sentiment analysis, where data often comes from tweets and other social media, we remove URLs, HTML tags, usernames, and strip emojis using the emoji² library.

²<https://pypi.org/project/emoji/>

Dialect identification filtering. Because many probing datasets originate from noisy online sources, they may contain *off-target* sentences (other dialects or MSA). To mitigate this, we run automatic dialect identification on every sentence using **CAMeL-Lab/bert-base-arabic-camelbert-mix-did-madar-corpus26** (Inoue et al., 2021). City-level predictions (e.g., *Riyadh, Jeddah, Cairo*) are mapped to countries, and we retain only sentences whose predicted country matches the dataset’s target dialect (or reassign them when the dataset explicitly spans multiple countries). This filtering step yields cleaner, better-aligned inputs for subsequent probing analyses.

4.2 Similarity Analysis Dataset

Our similarity analysis requires access to parallel sentences between MSA and dialects. Therefore, we use the **MADAR** corpus (Bouamor et al., 2018), which provides aligned versions of the same sentences in MSA and 25 city-level dialects. The corpus contains 2000 parallel sentences per city which helps avoid sample-size artifacts.

4.3 Models

We focus on BERT-style encoders throughout this study, so that cross-model comparisons are not confounded by architectural differences and transfer effects remain directly comparable. Table 1 summarizes the models used and their key properties that help contextualize their behavior and performance across our experiments.

4.4 Probing

Probe Training We employ a linear statistical classifier—multinomial logistic regression (softmax over a single affine layer)—implemented with the NEUROX toolkit (Dalvi et al., 2019). For the word-level tasks (POS and NER), we aggregate subword representations by selecting the embedding of the final subword token (Liu et al., 2019). For sentence-level Sentiment Analysis, representations are derived via mean-pooling across all tokens in the sequence. We use stratified 80/20 train/test splits, the Adam optimizer with a small learning rate. Each layer ℓ is probed independently, producing a performance curve per model and task. Token-level embeddings for POS/NER are obtained by aggregating subword representations to the word level (using the embedding of the last subword token (Liu et al., 2019)), while sentence-level embeddings for sentiment are computed by mean-pooling

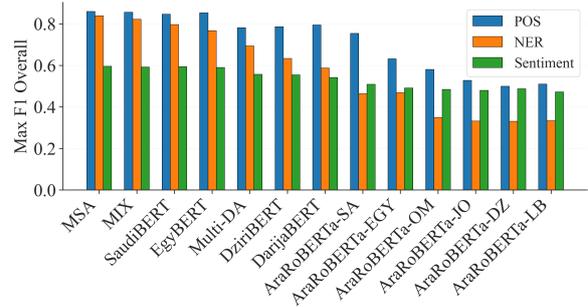


Figure 3: Performance of best performing layer on MSA Tasks.

token representations over the sequence.

Probe Evaluation We use the F1 score for the evaluation. We measure the performance of General models (i.e. MSA, Multi-DA and MIX) and DA models on both MSA only and DA datasets. We think that this offers a bidirectional look at the cross-lingual transfer of Arabic models. Where DA model that can transfer to MSA might indicate that the dialect is close to MSA and vice versa. Unless specified differently, we report the best F1 score across the model layers.

4.5 Representation Similarity

We use CKA to measure the representation similarity between MSA and DA models. We compute hidden states on the MADAR dataset, take the mean over the sentence tokens, then measure the layer-wise CKA similarity. The fact that MADAR is a parallel dataset allows us to investigate the following three scenarios:

1. MSA model vs. Dialect model representation of DA sentences.
2. MSA model vs. Dialect model representation of MSA sentences.
3. MSA model vs. Dialect model representation of MSA and DA sentences respectively.

Finally, since MADAR contains city level dialectal sentences and some countries are represented by more than one city (e.g., Egypt: Alexandria, Aswan, Cairo), we take the mean of CKA similarity over the cities to obtain the similarity per country, and consequently, per dialect.

Model Family	Model Name	Training Varieties	Tokens
CAMeLBERT	CAMeLBERT-MSA (Inoue et al., 2021)	MSA	12.6B
	CAMeLBERT-DA (Inoue et al., 2021)	DA	5.8B
	CAMeLBERT-MIX (Inoue et al., 2021)	MSA + DA + CA	17.3B
Dialect-Specific	SaudiBERT (Qarah, 2024b)	Saudi Dialect	2.7B
	EgyBERT (Qarah, 2024a)	Egyptian Dialect	1.07B
	DarijaBERT (Gaanoun et al., 2025)	Moroccan Dialect (Darija)	100M
	DziriBERT (Abdaoui et al., 2021)	Algerian Dialect	20M
AraRoBERTa	AraRoBERTa-SA (AlYami and Al-Zaidy, 2022)	Saudi Dialect	45.4M
	AraRoBERTa-EGY (AlYami and Al-Zaidy, 2022)	Egyptian Dialect	37.2M
	AraRoBERTa-OM (AlYami and Al-Zaidy, 2022)	Omani Dialect	3.8M
	AraRoBERTa-LB (AlYami and Al-Zaidy, 2022)	Lebanese Dialect	3.6M
	AraRoBERTa-JO (AlYami and Al-Zaidy, 2022)	Jordanian Dialect	2.6M
	AraRoBERTa-DZ (AlYami and Al-Zaidy, 2022)	Algerian Dialect	1.9M

Table 1: Overview of Arabic Pre-trained Language Models (PLMs)

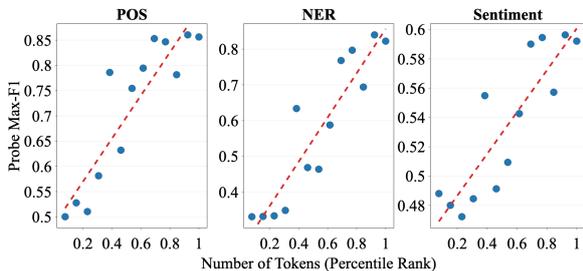


Figure 4: Impact of pretraining corpus size on probe performance across tasks. The Percentile Rank of the number of tokens is displayed for better visual interpretation.

5 Results and Analysis

5.1 Probing Analysis

For probing experiments implemented in Section 4.4, we analyze and compare the performance of the best layer for each model. This is based on evidence that models store different information disproportionately in different layers (Tenney et al., 2019a), while we are interested in comparing models in terms of their absolute performance rather than per-layer comparison.

Figure 3 reports the macro-F1 of the best-performing layer for each model and task on the MSA datasets, as discussed in Section 4.4. Across POS, NER, and SA, a consistent pattern emerges: the **MSA-pretrained** model achieves the strongest performance on MSA text, outperforming both the **MIX** and **Multi-DA** models. Among dialect-specific encoders, **Saudi** and **Egyptian** models typically rank closest to the MSA model, while other regional models lag more substantially. The same trend holds within the AraRoBERTa family, where Saudi and Egyptian variants systematically outperform other dialectal variants on MSA inputs. By

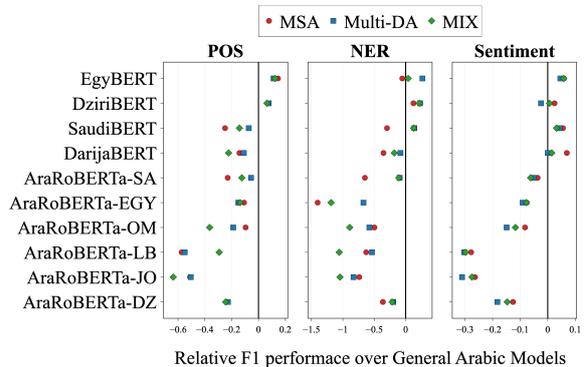


Figure 5: Relative performance of general vs. dialect-specific Arabic models on native dialectal datasets. Points to the right of the reference line denote cases where the dialect-specific model achieves higher performance, while points to the left indicate that the general model remains superior.³

contrast, models trained on North African varieties, such as DziriBERT and DarijaBERT, generalize poorly to MSA across all tasks, likely reflecting the pronounced lexical and morphological divergence between Maghrebi dialects and MSA. When we relate model training size, as summarized in Table 1, to performance on MSA data, we observe the trend visualized in Figure 4: models trained on larger and more diverse corpora tend to exhibit stronger generalization, underscoring the importance of corpus scale and linguistic variety for cross-dialectal robustness.

Figure 5 compares an **MSA** model, a **Multi-DA** model, and a **MIX** model against DA encoders on dialectal POS, NER, and SA. We compare models at their best-performing layer to control for layer specialization effects. We report the macro-

³Non-normalized F1 score results are provided in our code repository.

F1 score performance of each model relative to the dialect specific model. We compute the relative difference as follows:

$$\Delta = \frac{F1_{DA} - F1_M}{F1_{DA}} \quad (3)$$

where $M = \{\text{MSA, Multi-DA, MIX}\}$. We normalize by $F1_{DA}$ to make the results comparable across dialects.

Overall, DA models such as EgyBERT and DziriBERT consistently excel on their native dialects—especially for SA and for POS where morphology and cliticization diverge most from MSA—highlighting the benefits of dialect-specific pretraining. The MSA model, while weaker on SA for most dialects, remains competitive on structural tasks (POS, NER) for several varieties, indicating that MSA pretraining preserves transferable syntactic and entity-structure regularities. The Multi-DA model provides a strong dialect-agnostic baseline, where it often outperforms the MSA model on the SA task, but often falls short of the strongest DA models on their home dialect although it is trained on more data. This suggests that training on multiple Arabic dialects can still lead to negative interference (Wang et al., 2020), although they are regarded as similar in literature. The MIX model shows a similar breadth–sharpness trade-off: it can match or surpass many dialectal models on POS, but tends to underperform DA encoders on NER and SA when domain and style are highly dialect-specific. Across these comparisons, pretraining scale emerges as a key factor. DA models that reliably beat the general-purpose encoders (MSA, Multi-DA, MIX) are often trained on substantially larger pretraining corpora than lighter dialectal models (e.g., EgyBERT vs. AraRoBERTa-EGY, DziriBERT vs. AraRoBERTa-DZ). This suggests that the amount of pretraining data is crucial to improve the model’s encoding of dialect-specific morphological, syntactic, and semantic features.

5.2 Representation Similarity Analysis

Figure 6 summarizes the average layer-wise CKA between the MSA model and each DA model across all three scenarios discussed in Section 4.5. When both models encode the same dialectal sentences (Scenario 1), the highest alignment is observed for SaudiBERT on Saudi text, followed by DziriBERT, EgyBERT, and DarijaBERT, with AraRoBERTa-LB showing the closest alignment among the AraRoBERTa variants. When both

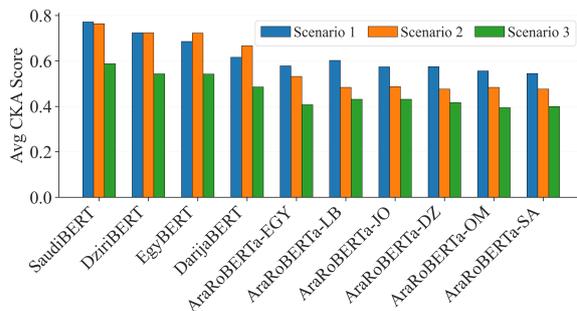


Figure 6: Average layer-wise CKA between the MSA model and DA models for the three scenarios described in Section 4.5.

encoders process the MSA versions of the same sentences (Scenario 2), the order of BERT-based dialect models remains similar, while within the AraRoBERTa family, AraRoBERTa-EGY becomes the closest to the MSA encoder and the remaining variants form a slightly lower, tightly clustered group. In scenario 3, where the MSA model encodes MSA sentences and dialectal models encode parallel dialect sentences, the BERT models ranking largely persists (SaudiBERT > DziriBERT > EgyBERT > DarijaBERT), whereas within AraRoBERTa, Levantine variants (AraRoBERTa-JO, AraRoBERTa-LB) show the strongest alignment, followed by North-African variants (AraRoBERTa-DZ, AraRoBERTa-EGY), then Gulf variants (AraRoBERTa-SA, AraRoBERTa-OM). Across scenarios, scenario 3 has the lowest similarity scores, showing the representations of DA and MSA models are not agnostic to their input variant. Furthermore, even when the input is the same variant (i.e. scenario 1 and 2), the similarity is always lower than 0.8, showing that MSA models fail to capture DA specific nuances and vice versa. The results of AraRoBERTa models suggest that this can be partly explained by the amount of training data, where the similarity of these models is lower than the BERT-based models, which are trained on more data.

5.3 Proximity Analysis

As discussed in Sections 5.2 and 5.1, there is significant disparity between dialect specific models in terms of their probing performance and similarity with MSA models. Motivated by the documented linguistic phenomena of dialect continuum (Chambers and Trudgill, 1998), and by previous studies that investigated the relationship between geographic proximity and lexical overlap (Alsu-

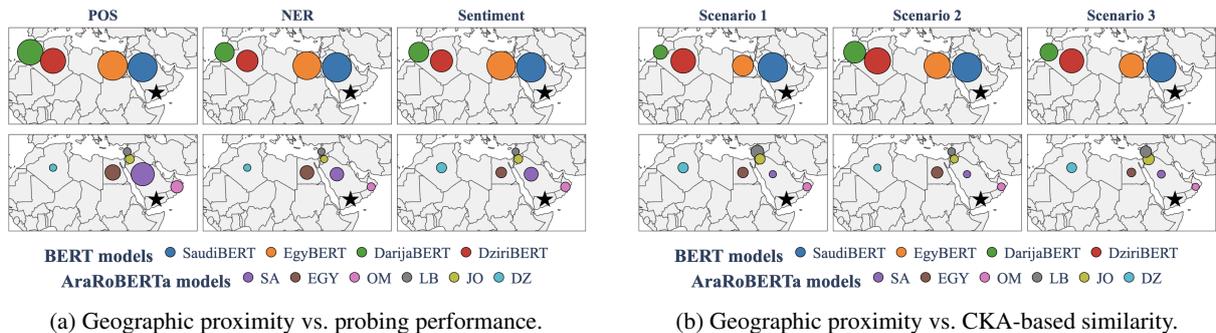


Figure 7: Comparison of Geographic Proximity with Probing Performance and CKA Similarity. The black star marks Yemen (the MSA anchor), and each bubble denotes a DA model; bubble size encodes (a) macro-F1 score on MSA and (b) average CKA score.

dais et al., 2022), we study whether geographic proximity can explain the disparity in probing and representation similarity results. Note that we use Yemen as the geographic location for MSA as explained in Section 3.3. Figure 7a shows dialectal models by their geographic location, with bubble sizes proportional to their macro-F1 score on MSA text from the probing tasks. Across all tasks, we observe a clear dialectal continuum: models associated with regions geographically closer to the MSA anchor are represented by larger bubbles, which gradually shrink with increasing distance. This pattern indicates that cross-lingual transfer in terms of *functional* similarity aligns well with geographic proximity. Figure 7b presents an analogous view for CKA-based similarity. Here, the dialectal continuum is evident primarily for models trained on large-scale corpora, while models with more limited pretraining data (i.e. AraRoBERTa models) do not reflect a dialect continuum effect. The results suggest that proximity plays a role in model performance and similarity, and is in line with the phenomena of dialect continuum, where more distant dialects and less mutually intelligible, which leads to weaker cross-lingual transfer and lower similarity. Specifically, Scenario 2 in Figure 7b is more comparable with Figure 7a since we compare the dialectal continuum of the MSA model and the DA models on MSA text.

6 Discussion

Based on the probing results on MSA datasets, we observe that the MSA model outperforms both MIX and Multi-DA models. The relative underperformance of the MIX model aligns with prior findings that monolingual MSA encoders surpass mixed-variant models on MSA tasks (Inoue et al., 2021).

Although the MIX model is trained on MSA data, the underperformance can be due to curse of multilinguality which adversely affects high-resource varieties (Chang et al., 2024). Among the DA models we observe that cross-lingual transfer to MSA is strongly dependent on the amount of data the model is trained on. This observation is further validated by geographical proximity analysis where models trained on large amount of data capture dialect continuum more efficiently. We also observe that it is easier for the language models to capture functional similarity even with less training data. But it is more difficult for Arabic language models to capture structural similarity as observed from geographical proximity analysis. Furthermore, probing and CKA reveal a clear functional–structural gap: some models (e.g., DziriBERT) can be structurally close to MSA yet fail to convert this into superior probing performance. This is in line with prior work showing that high CKA similarity does not guarantee functionally similar or equally useful features (Davari et al., 2023).

The probing performance on the DA datasets shows that the MSA-pretrained encoder is generally stronger on NER and POS tasks, while DA models tend to excel on SA tasks. The performance on POS suggests that Arabic and its dialects maintain a similar syntax, while the performance on NER can be attributed to code switching, where DA varieties often use MSA for entities. However, this does not translate to a higher performance on SA, where higher level semantic and pragmatic understanding is required to predict the sentiment. These findings suggest that Arabic dialects are similar to MSA in their syntax and use code switching for entities which explains their performance on POS and NER tasks respectively, however, they struggle on SA which is characterized by dialect

specific idioms, non-standard orthography and distinct negation patterns that are absent from their pretraining data.

Comparing DA models with *general* dialectal models (Multi-DA; MIX) on downstream tasks (Figures 5) shows that *general* models frequently achieve stronger average performance across dialects, a likely consequence of larger and more diverse pretraining. The MIX model commonly surpasses the Multi-DA model, suggesting that incorporating MSA and Classical Arabic alongside dialectal text improves the encoding of morphological and syntactic regularities. Nonetheless, there are notable exceptions: DziriBERT and EgyBERT often outperform *general* models on their native dialects. For SA, multiple DA models (with the exception of some AraRoBERTa variants) outperform general models, while the MIX model retains advantages on NER and POS tasks, where training on MSA text gives an advantage. We see that for most high resource dialects (Egyptian, Saudi Arabia), DA models usually outperform *general* dialect models. Whereas for low resource dialects (Lebanese, Jordan), *general* dialect models outperform native dialectal models. This observation is consistent with recent research on curse of multilinguality (Chang et al., 2024; Wu and Dredze, 2020), where multi-dialect pretraining benefits low resource dialects, but negatively affects high resource dialect performance.

7 Conclusion

In this work, we study how Arabic language models (LMs) capture cross-lingual transfer among dialects and their relationship to Modern Standard Arabic (MSA), combining **functional** evidence from probing classifiers (POS, NER, SA) and **structural** evidence from representation similarity (CKA), and linking both to **geographic proximity**. Our analysis covered MSA, MIX, and Multi-DA encoders alongside mono-dialectal DA models. We find consistent signals of a dialectal continuum: models associated with geographically closer varieties tend to align more closely with MSA in both probing and CKA, but this effect strongly depends on the amount of pretraining data. Interestingly, even among dialects of the same language we find evidence for negative interference, where Multi-DA models underperform mono-dialectal models of high resource dialects. Our findings question the assumption of cross-lingual transfer from MSA to

its dialects, where transfer is feasible but disproportionate across varieties. Furthermore, Multi-DA models might benefit from dialect or dialect-region specific parameters to avoid negative interference.

Limitations

Our work has the following limitations:

Dialectal dataset coverage: No publicly available Gulf-dialect NER dataset exists to our knowledge; we therefore leveraged MADAR and produced proxy annotations using an existing NER model. This introduces potential annotation noise and domain mismatch that may affect NER-related comparisons. Additionally, we rely on regional datasets (e.g., Gulf, Maghrebi) that are further segmented by country (e.g., Saudi, Moroccan) using automatic dialect identification. However, dialect identification is imperfect, and occasional false positives may introduce noise into the country-level splits and, consequently, into our downstream analyses.

Geographic proxy for MSA: Our choice of Yemen as a geographic proxy for MSA is literature-informed but not definitive. Results involving distance-to-anchor should be interpreted as operational rather than canonical; alternative anchors could yield different effect sizes.

Model heterogeneity: The compared models differ in training size, vocabularies, and pretraining hyperparameters. Such heterogeneity can confound attribution of observed effects. A more controlled comparison—equalizing corpus size, and training settings—would yield more robust findings.

Pretraining data composition: The paper does not focus the effect of the composition of pretraining data of dialectal models. Some models, such as DarijaBERT, were trained by filtering out MSA data, whereas SaudiBERT, DziriBERT, and EgyBERT were trained on data collected based on geolocation, which may include traces of MSA data. The presence of noisy MSA data in dialectal models may affect the model’s performance.

References

- Amine Abdaoui, Mohamed Berrimi, Mourad Oussalah, and Abdelouahab Moussaoui. 2021. [DziriBERT: a pre-trained language model for the algerian dialect](#). *CoRR*, abs/2109.12346.
- Ahmed Abdelali, Nadir Durrani, Fahim Dalvi, and Hassan Sajjad. 2022. [Post-hoc analysis of Arabic transformer models](#). In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting*

- Neural Networks for NLP*, pages 91–103, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- 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.
- Hossameddine Abouzahr and Mohammed Abouzahr. 2025. **Lexical and lexical-semantic comparisons of classical arabic and dialects**. *International Journal of Arabic Linguistics*, 11:66–98.
- Ibrahim Abu Farha and Walid Magdy. 2022. **The effect of Arabic dialect familiarity on data annotation**. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 399–408, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Kathrein Abu Kwaik, Motaz Saad, Stergios Chatzikiriakidis, and Simon Dobnik. 2018. **Shami: A corpus of Levantine Arabic dialects**. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Ashraf Ahmad, Mohammad Azzeh, Eman Alnagi, Qasem Abu Al-Haija, Dana Halabi, Abdullah Aref, and Yousef AbuHour. 2024. **Hate speech detection in the arabic language: corpus design, construction, and evaluation**. *Frontiers in Artificial Intelligence*, Volume 7 - 2024.
- Belen Alastruey, João Maria Janeiro, Alexandre Al-lauzen, Maha Elbayad, Loïc Barrault, and Marta R. Costa-jussà. 2025. **Interference matrix: Quantifying cross-lingual interference in transformer encoders**. *Preprint*, arXiv:2508.02256.
- Mashaël AlDuwais, Hend Al-Khalifa, and Abdulmalik AlSalman. 2024. **CLEANANERCorp: Identifying and correcting incorrect labels in the ANERcorp dataset**. In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation @ LREC-COLING 2024*, pages 13–19, Torino, Italia. ELRA and ICCL.
- Areeb Alowisheq, Nora Al-Twairish, Mawaheb Al-tuwaijri, Afnan Almoammar, Alhanouf Alsuwailam, Tarfa Albuhairi, Wejdan Alahaideb, and Sarah Alhummoud. 2021. **Marsa: Multi-domain arabic resources for sentiment analysis**. *IEEE Access*, 9:142718–142728.
- Abdulkareem Alsudais, Wafa Alotaibi, and Faye Alomary. 2022. **Similarities between arabic dialects: Investigating geographical proximity**. *Information Processing & Management*, 59(1):102770.
- Mohamed Aly and Amir Atiya. 2013. **LABR: A large scale Arabic book reviews dataset**. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 494–498, Sofia, Bulgaria. Association for Computational Linguistics.
- Reem AlYami and Rabah Al-Zaidy. 2022. **Weakly and semi-supervised learning for Arabic text classification using monodialectal language models**. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 260–272, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. **AraBERT: Transformer-based model for Arabic language understanding**. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.
- Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. **AraGPT2: Pre-trained transformer for Arabic language generation**. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 196–207, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.
- Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. **On the cross-lingual transferability of monolingual representations**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.
- Ankur Bapna and Orhan Firat. 2019. **Simple, scalable adaptation for neural machine translation**. 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 1538–1548, Hong Kong, China. Association for Computational Linguistics.
- Zakaria Benmounah, Abdennour Boulesnane, Abdeladim Fadheli, and Mustapha Khial. 2023. **Sentiment analysis on algerian dialect with transformers**. *Applied Sciences*, 13(20).
- Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouni, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhil Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. **The MADAR Arabic dialect corpus and lexicon**. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Ahmad Muhammad Boyi, Muhammad Badamasi Yusuf, and Mustapha Muhammad Isa. 2024. **The relevance**

- of arabic dialects and their diversity across the arab world. *Journal of Arts and Sociological Research*, 6.
- J. K. Chambers and Peter Trudgill. 1998. *Dialectology*, page 3–12. Cambridge Textbooks in Linguistics. Cambridge University Press.
- Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Ben Bergen. 2024. [When is multilinguality a curse? language modeling for 250 high- and low-resource languages](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4074–4096, Miami, Florida, USA. Association for Computational Linguistics.
- Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018a. [What you can cram into a single \\$&!#* vector: Probing sentence embeddings for linguistic properties](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.
- Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018b. [XNLI: Evaluating cross-lingual sentence representations](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.
- Abdelhalim Hafedh Dahou and Mohamed Amine Cheragui. 2023. [Dzner: A large algerian named entity recognition dataset](#). *Natural Language Processing Journal*, 3:100005.
- Fahim Dalvi, Avery Nortonsmith, Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, and James Glass. 2019. [Neurox: A toolkit for analyzing individual neurons in neural networks](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):9851–9852.
- Kareem Darwish, Hamdy Mubarak, Ahmed Abdelali, Mohamed Eldesouki, Younes Samih, Randah Alharbi, Mohammed Attia, Walid Magdy, and Laura Kallmeyer. 2018. [Multi-dialect Arabic POS tagging: A CRF approach](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- MohammadReza Davari, Stefan Horoi, Amine Natic, Guillaume Lajoie, Guy Wolf, and Eugene Belilovsky. 2023. [Reliability of CKA as a similarity measure in deep learning](#). In *The Eleventh International Conference on Learning Representations*.
- Niama Elkhbir, Urchade Zaratiana, Nadi Tomeh, and Thierry Charnois. 2023. [Cross-dialectal named entity recognition in Arabic](#). In *Proceedings of ArabicNLP 2023*, pages 140–149, Singapore (Hybrid). Association for Computational Linguistics.
- Juuso Eronen, Michal Ptaszynski, and Fumito Masui. 2023. [Zero-shot cross-lingual transfer language selection using linguistic similarity](#). *Information Processing & Management*, 60(3):103250.
- Charles A. Ferguson. 1959. [Diglossia](#). *WORD*, 15(2):325–340.
- Kamel Gaanoun, Abdou Mohamed Naira, Anass Al-lak, and Imade Benelallam. 2025. [Darijabert: a step forward in nlp for the written moroccan dialect](#). *International Journal of Data Science and Analytics*, 20(2):917–929.
- Nizar Habash, Muhammed AbuOdeh, Dima Taji, Reem Faraj, Jamila El Gizuli, and Omar Kallas. 2022. [Camel treebank: An open multi-genre Arabic dependency treebank](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2672–2681, Marseille, France. European Language Resources Association.
- Nizar Habash and Owen Rambow. 2006. [MAGEAD: A morphological analyzer and generator for the Arabic dialects](#). In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 681–688, Sydney, Australia. 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.
- 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*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.
- Mustafa Jarrar, Mohammed Khalilia, and Sana Ghanem. 2022. [Wojood: Nested Arabic named entity corpus and recognition using BERT](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3626–3636, Marseille, France. European Language Resources Association.
- Mouad Jbel, Mourad Jabrane, Imad Hafidi, and Abdulmutallib Mtrane. 2024. [Sentiment analysis dataset in moroccan dialect: bridging the gap between arabic and latin scripted dialect](#). *Lang. Resour. Eval.*, 59(2):1401–1430.
- Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. [Cross-lingual ability of multilingual bert: An empirical study](#). In *International Conference on Learning Representations*.

- Amr Keleg, Sharon Goldwater, and Walid Magdy. 2023. [ALDi: Quantifying the Arabic level of dialectness of text](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10597–10611, Singapore. Association for Computational Linguistics.
- Salam Khalifa, Nizar Habash, Dana Abdulrahim, and Sara Hassan. 2016. [A large scale corpus of Gulf Arabic](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4282–4289, Portorož, Slovenia. European Language Resources Association (ELRA).
- Rania Kora and Ammar Mohammed. 2019. [Corpus on Arabic Egyptian tweets](#).
- Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. [Similarity of neural network representations revisited](#). In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR.
- Kathrein Abu Kwaik, Motaz Saad, Stergios Chatzikyriakidis, and Simon Dobnik. 2018. [A lexical distance study of arabic dialects](#). *Procedia Computer Science*, 142:2–13. Arabic Computational Linguistics.
- Wuwei Lan, Yang Chen, Wei Xu, and Alan Ritter. 2020. [An empirical study of pre-trained transformers for Arabic information extraction](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4727–4734, Online. Association for Computational Linguistics.
- Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. [Choosing transfer languages for cross-lingual learning](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.
- Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. [Linguistic knowledge and transferability of contextual 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 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.
- Hanane Nour Moussa and Asmaa Mourhir. 2023. [Darnercorp: An annotated named entity recognition dataset in the moroccan dialect](#). *Data in Brief*, 48:109234.
- Hala Mulki, Hatem Haddad, Chedi Bechikh Ali, and Halima Alshabani. 2019. [L-HSAB: A Levantine Twitter dataset for hate speech and abusive language](#). In *Proceedings of the Third Workshop on Abusive Language Online*, pages 111–118, Florence, Italy. Association for Computational Linguistics.
- Jonathan Owens. 2006. *A Linguistic History of Arabic*. Oxford University Press.
- Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. [Identifying the correlation between language distance and cross-lingual transfer in a multilingual representation space](#). In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 22–29, Dubrovnik, Croatia. Association for Computational Linguistics.
- Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. [Information-theoretic probing for linguistic structure](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622, Online. Association for Computational Linguistics.
- Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. [How multilingual is multilingual BERT?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.
- Faisal Qarah. 2024a. [Egybert: A large language model pretrained on egyptian dialect corpora](#). *Preprint*, arXiv:2408.03524.
- Faisal Qarah. 2024b. [Saudibert: A large language model pretrained on saudi dialect corpora](#). *Preprint*, arXiv:2405.06239.
- Djamé Seddah, Farah Essaidi, Amal Fethi, Matthieu Futral, Benjamin Muller, Pedro Javier Ortiz Suárez, Benoît Sagot, and Abhishek Srivastava. 2020. [Building a user-generated content North-African Arabizi treebank: Tackling hell](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1139–1150, Online. Association for Computational Linguistics.
- Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Xudong Han, Sondos Mahmoud Bsharat, and 13 others. 2023. [Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models](#). *Preprint*, arXiv:2308.16149.
- Skyler Seto, Maartje Ter Hoeve, Richard He Bai, Natalie Schluter, and David Grangier. 2025. [Training bilingual LMs with data constraints in the targeted language](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19096–19122, Vienna, Austria. Association for Computational Linguistics.
- Guokan Shang, Hadi Abdine, Ahmad Chamma, Amr Mohamed, Mohamed Anwar, Abdelaziz Bounhar, Omar El Herraoui, Preslav Nakov, Michalis Vazirgiannis, and Eric P. Xing. 2025. [Nile-chat: Egyptian](#)

language models for Arabic and Latin scripts. In *Proceedings of The Third Arabic Natural Language Processing Conference*, pages 306–322, Suzhou, China. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. [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.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019b. [What do you learn from context? probing for sentence structure in contextualized word representations](#). In *International Conference on Learning Representations*.

Kees Versteegh. 2014. *The Arabic Language*, ned - new edition, 2 edition. Edinburgh University Press.

Zirui Wang, Zachary C. Lipton, and Yulia Tsvetkov. 2020. [On negative interference in multilingual models: Findings and a meta-learning treatment](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. [Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT](#). 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 833–844, Hong Kong, China. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. [Are all languages created equal in multilingual BERT?](#) In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

A Probing Datasets

Table 2 presents the datasets we use for probing, where we specify the dialects they support and the dialects we used them for.

Task	Dataset	Dataset Dialect	Used for Dialect	Total Samples	Dialect Samples
POS	NArabizi (Seddah et al., 2020)	Algeria	Algeria	1279	731
	QCRI Arabic POS Dialect (Darwish et al., 2018)	Egypt, Levant, Gulf, Maghreb	Egypt, Morocco	350, 350, 350, 350	338, 145
	Shami (Abu Kwaik et al., 2018)	Levant	Lebanon, Jordan	1069	240, 159
	CAMeL Treebank (Habash et al., 2022)	MSA and CA	MSA	5755 (MSA Variant, 21st Century)	2399
	GUMAR (Khalifa et al., 2016)	Gulf	Saudi, Oman	15205	1774, 1633
NER	CLEANANERcorp (AlDuwais et al., 2024)	MSA	MSA	4899	3050
	ACDNER (Elkhibir et al., 2023)	Egypt, Morocco, Syria	Egypt	353, 378, 361	210
	DzNER (Dahou and Cheragui, 2023)	Algeria	Algeria	5859	1153
	DarNERcorp (Moussa and Mourhir, 2023)	Morocco	Morocco	2511	2177
	Wojood (Jarrar et al., 2022)	MSA, Lebanon, Jordan	Lebanon, Jordan	24643, 383, 4705	190, 570
	MADAR (Tagged) (Bouamor et al., 2018)	Multi-Dialect, MSA	Saudi, Oman	2000 each	2000, 2000
SA	DZYT (Benmounah et al., 2023)	Algeria	Algeria	49942	7828
	AET (Kora and Mohammed, 2019)	Egypt	Egypt	39993	24546
	JHSC (Ahmad et al., 2024)	Jordan	Jordan	276003	30603
	L-HSAB (Mulki et al., 2019)	Lebanon	Lebanon	5377	1142
	LABR (Aly and Atiya, 2013)	MSA	MSA	63257	21696
	MYC (Jbel et al., 2024)	Morocco	Morocco	19455	6306
	MARSA (Alowisheq et al., 2021)	Gulf	Saudi, Oman	68128	10146, 2906

Table 2: Overview of Probing Datasets

Extending ASR Evaluation Resources for Modern Greek Dialects

Chara Tsoukala^{*1}, Stavros Bompolas^{*2}, Antigoni Margariti³, Konstantina Panagiotou⁴,
Maria Elisavet Plaiti⁴, Nefeli Tzanakaki³, Petros Karatsareas⁵, Angela Ralli^{2,6},
Antonios Anastasopoulos^{2,7}, Stella Markantonatou^{1,2}

¹ILSP/Athena R.C., Greece, ²Archimedes/Athena R.C., Greece, ³NKUA, Greece,

⁴University of Crete, Greece, ⁵University of Westminster, UK,

⁶University of Patras, Greece, ⁷George Mason University, USA

Correspondence: chara.tsoukala@athenarc.gr

Abstract

Recent progress in Automatic Speech Recognition (ASR) has primarily benefited high-resource standard languages, while dialectal speech remains challenging and underexplored. We present an expanded benchmark for low-resource Modern Greek dialects, covering Aperathiot, Cretan, Lesbian, and Cappadocian, spanning southern, northern, and contact-influenced varieties with varying degrees of divergence from Standard Modern Greek. The benchmark provides dialectal transcriptions in the Greek alphabet, following SMG-based orthographic conventions, while preserving dialectal lexical and morphophonological forms. Using this benchmark, we evaluate state-of-the-art multilingual ASR models in a zero-shot setting and by further fine-tuning per dialect. Zero-shot results reveal a clear performance gradient with dialectal distance from Standard Modern Greek, with best WERs ranging from about 60-70% for southern dialects to over 80% for Lesbian and nearly 97% for Cappadocian. Fine-tuning substantially reduces error rates (up to 47% relative WER improvement), with Cappadocian remaining the most challenging variety (best WER 68.17%). Overall, our results highlight persistent limitations of current pretrained ASR models under dialectal variation and the need for dedicated benchmarks and adaptation strategies.

1 Introduction

Recent advances in Automatic Speech Recognition (ASR), driven by end-to-end architectures and large-scale pretraining, have led to substantial improvements in performance for high-resource languages and standard language varieties (Babu et al., 2021; Radford et al., 2022; Omnilingual ASR Team et al., 2025). However, these improvements do not readily extend to dialectal (e.g., Torgbi et al., 2025) and other non-standard forms of speech (e.g., Koenecke et al.,

2020), where recognition accuracy degrades significantly. Dialectal variation poses challenges at multiple linguistic levels, including phonology, lexicon, morphology, syntax, and orthography, while the scarcity of annotated resources further limits the effectiveness of contemporary ASR models (Blaschke et al., 2025). As a result, speakers of regional and underrepresented varieties are still granted limited access to reliable speech technologies.

Large multilingual and self-supervised ASR models such as Whisper (Radford et al., 2022), XLS-R (Babu et al., 2021), and Omnilingual ASR (Omnilingual ASR Team et al., 2025) have become standard baselines for low-resource and cross-dialectal scenarios. Nevertheless, recent benchmarking efforts consistently show that even state-of-the-art models underperform on dialectal and accented speech compared to standard varieties, and typically require adaptation to achieve acceptable accuracy (Shi et al., 2024; Chen et al., 2025). Systematic evaluation across dialects is therefore essential to understand model behavior and remaining limitations.

Greek presents a particularly challenging test case due to its rich dialectal diversity (Kontosopoulos, 2008). Standard Modern Greek (SMG) co-exists with numerous regional and contact-influenced varieties that differ substantially in their linguistic properties, many of which are low-resource or endangered. While recent work has introduced an initial benchmark for Greek dialectal ASR (Vakirtzian et al., 2024), coverage remains limited, especially for varieties shaped by extensive language contact.

Building on this line of research, we present an expanded benchmark for ASR on low-resource Modern Greek (MG) dialects, incorporating four distinct varieties: Aperathiot, Cappadocian, Cretan, and Lesbian. Using this benchmark, we evaluate state-of-the-art ASR models from the Whis-

^{*}Equal contribution.

per and XLS-R families under zero-shot and fine-tuned conditions, and additionally assess Omnilingual ASR in a zero-shot setting.

Our contributions are threefold: (i) an expanded ASR benchmark covering previously unevaluated Greek dialects from both northern and southern groups, including the critically endangered Cappadocian; (ii) a systematic evaluation of multilingual ASR models with and without dialect-specific adaptation; and (iii) an empirical analysis demonstrating a strong relationship between dialectal distance and ASR performance, with word error rates ranging from approximately 60% for southern varieties to over 95% for structurally divergent and contact-influenced dialects.

2 Related Work

ASR for dialectal and non-standard speech remains a persistent challenge across many languages. A broad body of work shows that models trained primarily on standard or high-resource varieties degrade substantially when applied to regional dialects, even when those dialects are closely related. This pattern has been documented across typologically and sociolinguistically diverse settings, including Catalan (Hopton and Chodroff, 2025), Arabic (Khalafallah et al., 2024; Nasr et al., 2023), Japanese (Imaizumi et al., 2020; Toyama et al., 2024; Takahashi et al., 2024), German and Swiss German (Blaschke et al., 2025; Sicard et al., 2023), Scottish English (Torgbi et al., 2025), Telugu and other Indian languages (Aditya Yadavalli and Ganesh Mirishkar and Anil Kumar Vuppala, 2022; Bhardwaj et al., 2021; Alumäe et al., 2023), Tibetan (Qin et al., 2022), Pomak (Tsoukala et al., 2023), and numerous African languages surveyed by Imam et al. (2025).

Recent benchmarking efforts further confirm that dialectal and accented speech remains difficult even for state-of-the-art multilingual models. The ML-SUPERB 2.0 benchmark and challenge (Chen et al., 2025; Shi et al., 2024) has evaluated ASR and language identification across more than 200 languages, accents, and dialects, revealing substantial performance disparities and consistent degradation with non-standard varieties. More recently, Omnilingual ASR (Omnilingual ASR Team et al., 2025) has expanded multilingual coverage to over 1,600 languages, highlighting the growing emphasis on scale, although its robustness to dialectal variation within individual

languages remains largely unexplored.

A complementary line of work addresses these issues through dialect-specific data creation and modeling. New corpora and evaluation resources have been introduced for Arabic dialects (Khalafallah et al., 2024; Nasr et al., 2023), Punjabi and Telugu (Bhardwaj et al., 2021; Aditya Yadavalli and Ganesh Mirishkar and Anil Kumar Vuppala, 2022), German dialects (Blaschke et al., 2025), and low-resource varieties such as Tibetan and Pomak (Qin et al., 2022; Tsoukala et al., 2023), highlighting the central role of dedicated datasets in dialectal ASR research.

More recent work has focused on adapting end-to-end and self-supervised ASR models to dialectal speech. Increased dialectal diversity during fine-tuning has been shown to improve robustness for Catalan (Hopton and Chodroff, 2025), while dialect-aware modeling and adaptation of large pretrained models have returned consistent gains for Japanese (Imaizumi et al., 2020; Toyama et al., 2024; Takahashi et al., 2024), Swiss German (Sicard et al., 2023), Scottish accents (Torgbi et al., 2025), and dialect-rich Indian languages (Alumäe et al., 2023). Nonetheless, performance remains sensitive to the specific variety, data availability, and evaluation choices.

Despite this growing body of work across diverse language families, MG remains underrepresented. Existing ASR research has largely focused on SMG, with limited attention to dialectal variation. While domain-specific adaptation has been explored, for example in medical dictation (Georgilas and Stafylakis, 2025), systematic evaluation across MG dialects remains largely unexplored. To date, the benchmark introduced by Vakirtzian et al. (2024), covering Aivaliot, Eastern Cretan, Griko, and Messenian, constitutes the only dedicated effort in this direction.

We extend this line of work by introducing four additional varieties, including Cappadocian, a typological outlier shaped by extensive contact with Turkish, and providing systematic evaluation across a broader range of ASR models.

3 Dialectal Scope of the Study

This study examines four MG varieties: Aperathiot (spoken on the island of Naxos), Cappadocian, Cretan, and Lesbian. According to Trudgill (2003, 59–60), Aperathiot, as part of the Naxos varieties, belongs to the Northern



Figure 1: Geographic distribution of the Greek dialects examined in this study; squares indicate data collection sites.

Cyclades group, Cretan to the southern group, and Lesbian to the northern group of MG dialects. Cappadocian falls outside this classification due to its development in isolation and under strong Turkish influence (Karatsareas, 2011, 45). Together with SMG, these varieties form a heterogeneous dialectal landscape (Figure 1) well suited for evaluating ASR under conditions of dialectal variation.

Aperathiot (Glottocode: cyc11238) is spoken primarily in Apeiranthos, a mountainous village in central Naxos. Linguistically, it forms a distinct dialectal enclave, differing markedly from other varieties of Naxos across all linguistic levels (Oikonomidis, 1952, 216, 272–273). Its origins have been linked to historical Cretan settlement, based on linguistic and cultural parallels, though this remains debated (Xeferi, 2009). Increased mobility and tourism in recent decades have reduced intergenerational transmission, heightening the urgency of documentation and technological support.

Cappadocian (ISO 639-3: cpg; Glottocode: capp1239) refers to a group of MG dialects historically spoken in Cappadocia (present-day Turkey). Prolonged contact with Turkish, combined with long-term isolation from other Greek varieties, triggered extensive structural innovation, making Cappadocian a typological outlier within MG. Beyond pervasive lexical borrowing, Turkish influence is reported in morphosyntax and clause structure, including nominal and verbal morphology, argument marking, and constituent order. These innovations are widely discussed in the contact-linguistic literature (see Dawkins, 1916; Karatsareas, 2011; Thomason and Kaufman, 1988, 93–94, 215–222 for discussion and examples). Following the population exchange of the early 1920s, speak-

ers were resettled in Greece and rapidly shifted toward SMG. Although long considered extinct, native speakers were identified in the mid-2000s through fieldwork by Janse and Papazachariou. Today, only one variety survives, Mišiotika, which is generally regarded as being heavily influenced by Turkish (Dawkins, 1916, 209; Bompolas, 2023, 165–170). Mišiotika also exhibits phonological features resembling *northern vocalism*, discussed below in relation to Lesbian. However, these features are assumed to have developed independently and do not place Cappadocian within the traditional northern-southern division of MG dialects (Dawkins, 1916, 192–193). Mišiotika is spoken by only a few hundred, mostly elderly speakers and is classified as critically endangered (UNESCO, 2010).

Cretan (Glottocode: cret1244) is spoken across Crete and in diaspora communities. Its development reflects long-term geographic isolation and successive periods of foreign rule, including Arab, Venetian, and Ottoman domination (Pangalos, 1955; Kontosopoulos, 2008, 28–41). Cretan is traditionally divided into Eastern and Western subvarieties, with a boundary that roughly coincides with the administrative division between the prefectures of Rethymno and Heraklion (Pangalos, 1955, 143–151). The Eastern variety is generally described as more homogeneous, whereas the Western variety exhibits greater internal diversity (Kontosopoulos, 2008, 36). Unlike most MG dialects, Cretan remains robust and widely used. Our benchmark includes data from both subvarieties, extending previous resources that focused exclusively on Eastern Cretan (Vakirtzian et al., 2024).

Lesbian is the only northern MG dialect examined in this study (Glottocode: nort2600). It is characterized by the so-called *northern vocalism*, including vowel raising and deletion: unstressed mid vowels /e/ and /o/ are raised to [i] and [u], respectively, while unstressed high vowels /i/ and /u/ are deleted. These features clearly distinguish Lesbian from southern dialects, including SMG (Chatzidakis, 1905). The dialect has been shaped by prolonged contact with Italo-Romance and Turkish, primarily affecting the lexicon and morphology (Ralli, 2015, 2019a,b; Alexelli, 2021). Population movements between Lesbos and nearby Asia Minor (e.g., Ayvalik and Moschonisia), followed by refugee resettlement

after 1922, have contributed to notable intra-dialectal variation. Unlike many MG dialects, Lesbian remains vital and continues to function as the primary means of everyday communication on the island.

4 Datasets

Our benchmark consists of naturalistic speech datasets (narratives, conversations, and everyday-life stories) recorded from speakers of each variety, with transcriptions manually verified by trained linguists. Corpus statistics are reported in Table 1.

Aperathiot The Aperathiot dataset was collected in 2025 in the village of Apeiranthos and consists of narratives, conversations, and everyday-life stories from four native speakers (1 male, 3 female), ranging from middle-aged to elderly. Initial transcriptions were generated using Whisper Large-v3 and subsequently manually corrected by a trained linguist who is also a native speaker of the dialect.

Cappadocian The Cappadocian dataset consists of conversations, narratives, and everyday life stories recorded during fieldwork conducted in 2011 by twelve native or heritage speakers (8 male, 4 female) of Mišiotika in a village in Northern Greece. Speakers’ ages range from 19 to 94 years; the group includes both individuals born in Misti (Cappadocia) prior to the population exchange and later-generation speakers born and raised in Greece. The village where the recordings were collected is marked with a square on the map in Figure 1. The audio recordings were manually segmented into utterances and transcribed by two trained native speakers; in addition, parallel translations into SMG were provided as an auxiliary resource (and are not used as the reference transcription in our ASR evaluation).

Cretan The Cretan dataset combines previously published material with newly collected recordings. The previously published material consists of approximately two hours of already-processed transcribed radio broadcasts representing the Eastern Cretan variety, originally published by Vakirtzian et al. (2024)¹ and recorded between 1998 and 2001 by Radio Mires in the Messara region of Heraklion. To extend dialectal coverage beyond the Eastern variety, we collected

¹huggingface.co/datasets/ilsp/cretan-speech-corpus

Corpus	Tokens	Utterances	Audio Duration	
			Original	Processed
Aperathiot ²	8,830	798	1h 24m 17s	1h 3m 27s
Cappadocian ³	11,715	2,357	3h 29m 42s	1h 17m 5s
Cretan ⁴	36,594	6,897	4h 59m 47s ⁵	3h 56m 54s
Lesbian ⁶	11,652	2,294	2h 29m 14s	1h 6m 12s
Total	68,791	12,346	12h 23m 0s	7h 23m 38s

Table 1: Summary Statistics of the Speech Corpora.

additional recordings in 2025 from two elderly native speakers (1 male, 1 female) from Western Crete, consisting of natural conversations and narratives. The villages where the new Western Cretan data were recorded are marked with squares on the map in Figure 1. The transcriptions of the new data were generated using Whisper Large-v3 and manually corrected by trained native-speaker linguists. All transcriptions follow the orthography of SMG and adhere closely to the conventions used by Vakirtzian et al. (2024) in the Eastern Crete corpus to ensure comparability of subsets.

Lesbian The Lesbian dataset was collected through fieldwork conducted in 2023–2024 and consists of narratives, conversations, and everyday-life stories from eleven native speakers (5 male, 6 female) originating from eight villages in northern Lesbos. Initial audio-aligned transcriptions were generated using Whisper Large-v3 and subsequently manually corrected by a trained linguist who is also a native speaker of the dialect. Transcription follows the benchmark policy described below (Section 4.1): dialectal forms are rendered in the Greek alphabet using SMG-based grapheme–phoneme correspondences, with systematic adaptations to capture characteristic phonological patterns of northern vocalism. Specifically, the raising of /o/ (orthographic *o*, *ω*) is transcribed as *ou*, and the raising of /e/ (orthographic *α*, *ε*) as *ι*. The deletion of unstressed high vowels /u/ and /i/ (corresponding to the relevant Greek orthographic representations) is marked by an apostrophe only in word-final position; word-internal deletions are not represented, in order to preserve compatibility with standard orthographic conventions.

²huggingface.co/datasets/ilsp/aperathiot-speech-corpus

³huggingface.co/datasets/ilsp/cappadocian-speech-corpus

⁴huggingface.co/datasets/ilsp/cretan-extended-speech-corpus

⁵For Cretan, "Original" reflects the Eastern corpus as published plus unprocessed Western recordings.

⁶huggingface.co/datasets/ilsp/lesbian-speech-corpus

4.1 Transcription Policy and Benchmark Scope

The benchmark provides dialectal transcriptions (e.g., Kuparinen, 2025), not dialect-to-standard renderings (e.g., Blaschke et al., 2025; Ducceschi and Franzini, 2025; see also Dimakis et al., 2025 on dialect normalization for Greek in NLP). All varieties are transcribed using the 24-letter Greek alphabet and an orthography anchored in SMG grapheme-phoneme correspondences; i.e., we employ SMG-based spelling conventions as an orthographic substrate while preserving dialectal lexical and morphophonological forms. This choice is motivated by the fact that the dialects covered here do not have a widely adopted standardized orthography; therefore, we follow an ad-hoc pronunciation-oriented spelling practice that is common in modern dialectal texts for these varieties, aiming for consistency across corpora and readability for Greek-literate users. Importantly, we do not provide a parallel normalized to SMG reference transcription; accordingly, the benchmark is intended for dialectal ASR (recovering the dialectal wording in SMG-based Greek orthography), rather than dialect-to-standard ASR/speech-to-standard translation.

4.2 Data Anonymization

All speech data were anonymized prior to release, at both the audio and transcription levels. For Aperathiot and Cretan, identifying information was removed by muting audio segments and deleting aligned transcriptions, while for Cappadocian and Lesbian, sensitive segments were excluded during transcription. This process was performed manually by trained linguists to ensure speaker privacy while preserving linguistic integrity.

5 Experiments

5.1 Models

To assess the robustness of current ASR systems to dialectal variation in MG, we evaluate a range of widely used multilingual speech recognition models under both inference-only and fine-tuned settings. Our benchmark focuses on models from the Whisper (Radford et al., 2022) and XLS-R (Babu et al., 2021) families, as well as Omnilingual ASR (Omnilingual ASR Team et al., 2025), all of which have emerged as standard reference points for low-resource and cross-dialectal ASR.

Inference-only evaluation All dialects are evaluated under zero-shot (inference-only) conditions using pretrained models without dialect-specific adaptation. We report results for representative models from the Whisper and XLS-R families, as well as Omnilingual ASR, enabling comparison across architectures, model sizes, and levels of language-specific adaptation. Specifically, we evaluate Whisper Large-v3, Whisper Large-v2, Whisper-medium, XLS-R-53-greek, XLS-R-300-greek, and Omnilingual ASR.

XLS-R is a multilingual speech encoder trained on approximately 56k hours of audio from 53 languages. In our experiments, we use Greek-adapted XLS-R variants, which provide stronger baselines for MG and closely related varieties than the original multilingual checkpoints. Whisper, in contrast, is a large-scale multilingual sequence-to-sequence model trained on substantially larger and more heterogeneous data: Whisper Large-v2 was trained on roughly 680k hours of weakly supervised audio, while the most recent Large-v3 extends training to approximately 1 million hours of weakly labeled audio and an additional 4 million hours of pseudo-labeled data generated using Large-v2. Finally, Omnilingual ASR is a large-scale multilingual model supporting over 1,600 languages and is evaluated exclusively under inference-only conditions.

Fine-tuning In addition to inference-only evaluation, we examine the impact of dialect-specific adaptation by fine-tuning XLS-R-53-greek and Whisper-medium for 35 epochs. Fine-tuning is performed separately for each dialect included in the benchmark.

Larger Whisper variants (Large-v2 and Large-v3) were not fine-tuned due to their high computational cost and the increased risk of overfitting given the limited amount of available dialectal data. Omnilingual ASR was likewise evaluated only in a zero-shot setting, as its scale and training setup make dialect-specific fine-tuning impractical in low-resource scenarios. Similarly, XLS-R-300-greek was not fine-tuned, since preliminary experiments showed no clear gains over the smaller XLS-R-53-greek variant under low-resource conditions.

5.2 Preprocessing

All datasets were processed using a unified preprocessing pipeline to ensure consistency across dialects. All preprocessing, fine-tuning, and evalu-

ation scripts are available at <https://github.com/athena-ils/greek-dialects-asr>. Audio recordings were converted to mono WAV files at a sampling rate of 16 kHz. Text normalization consisted of lowercasing the text and removing punctuation.

5.3 Segmentation

ASR models require relatively short audio segments for both training and evaluation. All recordings were therefore segmented into utterances with a maximum duration of 30 seconds. For Cappadocian, where time-aligned transcriptions were already available, segmentation was performed automatically using the existing annotations.

For the remaining dialects, which include newly collected material, audio recordings were first transcribed using Whisper, and the resulting timestamp information was converted into Praat TextGrid files (Boersma and Weenink, 2001). Next, these files were manually reviewed and corrected by trained native-speaker linguists, who refined both segment boundaries and transcriptions. Non-speech material, such as long pauses, untranscribed portions, and, where applicable, the utterances of the fieldworker, was removed. As a result, the effective audio duration was reduced for some datasets, as reflected in Table 1.

5.4 Dataset Construction

Following segmentation, a separate dataset was constructed for each dialect using the resulting audio–text pairs. Utterances exceeding 30 seconds were excluded. The material was then split into training, development, and test sets using an 80/10/10 split to enable supervised adaptation and evaluation.

5.5 Fine-tuning

Fine-tuning experiments were carried out using XLS-R-53-greek and Whisper-medium with model selection based on validation WER (best checkpoint loaded at the end of training). XLS-R-53-greek was trained on an NVIDIA GeForce RTX 3090 GPU for up to 35 epochs with early stopping, using a learning rate of 3×10^{-4} , batch size of 8, and gradient accumulation of 2. Whisper-medium was trained on an NVIDIA A100 with a learning rate of 10^{-5} and the same batch configuration; training used step-based evaluation and checkpointing every 1000 steps up to a maximum of 10000 update steps. Parameter-efficient fine-tuning methods (e.g., LoRA) are left for future

work and would enable extending fine-tuning to larger checkpoints under the same compute budget.

6 Results

This section reports the performance of the evaluated ASR models on Greek dialectal speech. We first present inference-only (zero-shot) results obtained with pretrained models, followed by results after dialect-specific fine-tuning. Performance is measured using Word Error Rate (WER) and Character Error Rate (CER), computed on normalized text without punctuation.

6.1 Inference-only Evaluation

ASR performance under inference-only conditions, measured by WER and CER, is reported in the upper parts of Tables 2–5. Across dialects, WER and CER exhibit distinct but complementary patterns that reflect both the degree of dialectal divergence from SMG and the ability of different model architectures to capture subword structure under mismatch.

For Aperathiot and Cretan, both southern dialects sharing core phonological and morphosyntactic properties with SMG, error rates are consistently lower than for the remaining varieties. Whisper Large-v3 achieves the best zero-shot performance in both cases (Aperathiot: WER 61.85%, CER 33.08%; Cretan: WER 70.24%, CER 41.74%), indicating that large multilingual sequence-to-sequence models are relatively robust when dialectal variation remains close to the standard language.

The Omnilingual ASR model performs competitively for Aperathiot, slightly outperforming Whisper Large-v3 at both the word and character level (WER 60.22%, CER 31.04%), but degrades more substantially on Cretan (WER 86.31%, CER 49.45%). In contrast, the Greek-adapted XLS-R models yield WERs exceeding 100% in both dialects, despite somewhat lower CERs, suggesting that while some subword patterns are captured, word-level reconstruction fails under zero-shot conditions.

For Lesbian, the only northern dialect in the benchmark, both WER and CER increase markedly across all models. Whisper Large-v3 again performs best (WER 80.87%, CER 57.70%), while Omnilingual ASR shows comparable character-level accuracy (CER 56.07%) but

Model	Checkpoint	WER (%)	CER (%)
Large-v3	pretrained	61.85%	33.08%
Large-v2	pretrained	69.68%	40.84%
Whisper-medium	pretrained	70.32%	39.73%
XLS-R-53-greek	pretrained	104.18%	91.89%
XLS-R-300-greek	pretrained	101.08%	90.48%
Omnilingual ASR	pretrained	60.22%	31.04%
XLS-R-53-greek ⁷	epoch 34	45.64%	16.45%
Whisper-medium ⁸	step 2000	37.41%	15.54%

Table 2: Aperathiot Model Performance Comparison.

Model	Checkpoint	WER (%)	CER (%)
Large-v3	pretrained	96.65%	59.72%
Large-v2	pretrained	105.94%	70.41%
Whisper-medium	pretrained	117.13%	91.25%
XLS-R-53-greek	pretrained	109.4%	97.48%
XLS-R-300-greek	pretrained	103.1%	94.03%
Omnilingual ASR	pretrained	101.24%	59.72%
XLS-R-53-greek ⁹	epoch 16	68.17%	30.66%
Whisper-medium ¹⁰	step 4000	77.79%	47.80%

Table 3: Cappadocian Model Performance Comparison.

substantially higher WER (95.05%). The elevated CER values across models indicate systematic difficulties in modeling vowel reduction and deletion associated with northern vocalism, which affects not only word segmentation but also character-level alignment. The persistent gap between southern and northern dialects thus emerges clearly in both evaluation metrics.

Cappadocian represents the most challenging case under inference-only evaluation. All pretrained models exhibit extremely high WERs, reflecting severe word-level mismatch. Whisper Large-v3 attains the lowest WER (96.65%) and a comparatively lower CER (59.72%), closely matched at the character level by the Omnilingual ASR (CER 59.72%) but with higher WER (101.24%). Other Whisper variants and the Greek-adapted XLS-R models perform progressively worse, with CER values exceeding 70% and in some cases approaching 95%. The divergence between WER and CER for Cappadocian is particularly informative: although most models fail to recover correct word forms, lower CERs for Whisper Large-v3 and Omnilingual ASR sug-

⁷huggingface.co/ilsp/xls-r-53-greek-aperathiot

⁸huggingface.co/ilsp/whisper-aperathiot-asr

⁹huggingface.co/ilsp/xls-r-53-greek-cappadocian

¹⁰huggingface.co/ilsp/whisper-cappadocian-asr

¹¹huggingface.co/ilsp/xls-r-53-greek-cretan-extended

¹²huggingface.co/ilsp/whisper-cretan-extended-asr

¹³huggingface.co/ilsp/xls-r-53-greek-lesbian

¹⁴huggingface.co/ilsp/whisper-lesbian-asr

Model	Checkpoint	WER (%)	CER (%)
Large-v3	pretrained	70.24%	41.74%
Large-v2	pretrained	78.90%	50.67%
Whisper-medium	pretrained	81.69%	55.92%
XLS-R-53-greek	pretrained	106.29%	94.81%
XLS-R-300-greek	pretrained	102.55%	93.04%
Omnilingual ASR	pretrained	86.31%	49.45%
XLS-R-53-greek ¹¹	epoch 6	55.04%	22.08%
Whisper-medium ¹²	step 2000	37.02%	17.02%

Table 4: Cretan Model Performance Comparison.

Model	Checkpoint	WER (%)	CER (%)
Large-v3	pretrained	80.87%	57.70%
Large-v2	pretrained	92.02%	57.70%
Whisper-medium	pretrained	87.89%	60.86%
XLS-R-53-greek	pretrained	107.96%	97.02%
XLS-R-300-greek	pretrained	102.03%	93.6%
Omnilingual ASR	pretrained	95.05%	56.07%
XLS-R-53-greek ¹³	epoch 22	71.79%	32.97%
Whisper-medium ¹⁴	step 4000	54.73%	26.70%

Table 5: Lesbian Model Performance Comparison.

gest partial preservation of phonotactic or subword structure even under extreme mismatch. This outcome is expected given that Cappadocian is a typological outlier within the Greek dialectal continuum: beyond prolonged contact with Turkish, it has undergone extensive structural changes across all linguistic levels, resulting in substantial lexical, phonological, and morphosyntactic divergence from the data used to pretrain current ASR models.

Overall, inference-only results point to a strong interaction between dialectal distance and error type. Varieties closer to SMG yield lower WER and CER, whereas increased phonological divergence and contact-induced change lead to sharp degradation, most visibly at the word level. At the same time, systematic WER–CER gaps indicate that some architectures preserve limited subword structure even when word-level reconstruction fails. Interpretation of zero-shot scores must also account for the benchmark target: our references encode dialectal lexical and morphophonological forms rendered in SMG-based orthography (cf. Section 4.1), so performance reflects not only acoustic mismatch but also lexical and orthographic mismatch with decoders biased toward standard-language distributions. Models exposed predominantly to SMG text may therefore normalize toward SMG-like outputs or fail to reproduce dialect-specific spellings, inflating WER under inference-only settings.

Taken together, these observations suggest that

current pretrained ASR systems do not adequately handle Greek dialectal speech—especially structurally divergent varieties—without dialect-aware adaptation, motivating the fine-tuning experiments presented in the following section.

6.2 Fine-tuning

Dialect-specific fine-tuning leads to substantial and consistent improvements across all varieties (Tables 2-5, lower rows), confirming that even limited amounts of dialectal data are sufficient to markedly reduce both WER and CER.

For the southern dialects (i.e., Aperathiot and Cretan), fine-tuning Whisper-medium yields the strongest gains. In Aperathiot, WER is reduced from 60.22% under the best inference-only setting (Omnilingual ASR) to 37.41%, corresponding to a 39% relative improvement, while CER drops to 15.54%. Cretan shows the largest absolute improvement: Whisper-medium achieves 37.02% WER and 17.02% CER, representing a 47% relative WER reduction compared to the best zero-shot model (Whisper Large-v3 at 70.24%). These results indicate that dialects structurally close to SMG benefit strongly from lightweight adaptation, with fine-tuning substantially closing the gap between dialectal and standard speech recognition.

For Lesbian, fine-tuning also yields clear improvements, though performance remains lower than for southern dialects. Whisper-medium reduces WER from 80.87% to 54.73% (32% relative improvement) and CER from 57.70% to 26.70%. The remaining gap reflects persistent phonological divergence due to northern vocalism, which continues to challenge word-level recognition even after adaptation, despite substantial gains at the character level.

Cappadocian remains the most challenging variety after fine-tuning. The best-performing adapted model is XLS-R-53-greek, achieving 68.17% WER and 30.66% CER, corresponding to a 29% relative WER reduction compared to the best zero-shot result (Whisper Large-v3 at 96.65%). Notably, for Cappadocian, XLS-R-53-greek outperforms Whisper-medium after fine-tuning, reversing the pattern observed in the other dialects. This suggests that the wav2vec-style architecture may be better suited to extreme out-of-distribution settings when sufficient dialect-specific acoustic evidence is provided, particularly at the subword level, as reflected in the larger CER reductions.

Across all dialects, fine-tuning consistently nar-

rows the gap between WER and CER, indicating improved alignment between acoustic modeling and lexical prediction. However, the magnitude of improvement varies systematically with dialectal distance: while adaptation largely mitigates mismatch for southern dialects, varieties involving deep structural divergence and extensive contact-induced change remain substantially more difficult.

Overall, these results demonstrate that fine-tuning is essential for robust ASR on Greek dialects, but also that model architecture and dialect typology jointly shape the limits of adaptation under low-resource conditions.

7 Discussion and Outlook

Across both inference-only and fine-tuned settings, dialectal distance from SMG is a strong predictor of ASR difficulty, with systematic differences between southern, northern, and contact-influenced varieties. Southern dialects (Aperathiot, Cretan) are consistently easier, while the northern Lesbian and especially Cappadocian remain substantially harder, even after adaptation.

A key result is the growing divergence between WER and CER as dialectal distance increases. For the southern dialects, both metrics are comparatively lower, suggesting that models recover not only phonotactics but also a meaningful portion of the word inventory. For Lesbian and Cappadocian, WER rises sharply—often approaching or exceeding 100% in zero-shot evaluation—while CER remains noticeably lower. This is clearest in Cappadocian, where Whisper Large-v3 and Omnilingual ASR both reach 59.72% CER but still exhibit very high WER (96.65% and 101.24%), indicating that subword structure is partially captured even when word-level recognition breaks down under severe lexical and structural mismatch.

Model comparisons further point to architectural effects. Whisper Large-v3 is the strongest zero-shot Whisper model across dialects, while Omnilingual ASR generalizes unevenly: it slightly outperforms Large-v3 on Aperathiot, but degrades on Cretan, where Large-v3 is clearly better. In contrast, both Greek-adapted XLS-R models perform poorly in zero-shot conditions across dialects, and scaling from XLS-R-53-greek to XLS-R-300-greek yields no clear improvement, suggesting that model scaling alone does not resolve strong dialectal mismatch.

Fine-tuning improves all varieties but with uneven returns. Whisper-medium reaches $\approx 37\%$ WER on both Aperathiot and Cretan (about 40–47% relative reduction), while Lesbian (54.73%) and Cappadocian (68.17%) remain substantially harder. Notably, Cappadocian is the only case where fine-tuned XLS-R-53-greek outperforms fine-tuned Whisper-medium, suggesting that CTC-style adaptation may be more robust than seq2seq decoding under extreme out-of-distribution conditions (for similar results and discussion in dialectal/low-resource settings, see Williams et al., 2023; Adnan and Hassani, 2025; see also Barcovski et al., 2023; Vázquez-Correa and Álvarez Muñain, 2023). This suggests architectural differences in handling extreme out-of-distribution data, a question worth exploring in future work.

Our findings align with prior Greek dialect ASR benchmarking (Vakirtzian et al., 2024). Their zero-shot Whisper Large-v3 result on Eastern Cretan (58.42% WER; single speaker) is lower than ours on combined Eastern–Western Cretan (70.24%), plausibly reflecting greater internal variation in Western Cretan (Kontosopoulos, 2008) without changing the overall pattern. Similarly, our results for the dialect of Lesbos match the high-error profile reported for the closely related Aivaliot variety (zero-shot WER > 100%), reinforcing that northern vocalism and related SMG-mismatch phenomena pose a systematic challenge for pretrained ASR. Beyond ASR, similar distance-sensitive transfer effects have been observed in morphosyntactic parsing, with stronger transfer from SMG to Cretan than to Lesbian (Bompolas et al., 2025; Vakirtzian et al., 2025). Contact-driven divergence remains the hardest case: Vakirtzian et al. (2024) report WER > 100% for Griko (Italo-Romance contact plus Latin orthography), while Cappadocian (contact with Turkish plus extensive structural change) is the most challenging variety in our benchmark even after fine-tuning. Unlike Aivaliot, where Turkish influence is largely lexical (Vakirtzian et al., 2024), Cappadocian shows multi-level restructuring that amplifies lexical and subword mismatch.

Overall, performance ranges from WERs near 37% for fine-tuned southern dialects to nearly 70% for Cappadocian, highlighting the need for dialect-specific benchmarks and adaptation strategies. For the most divergent varieties, where orthographic and lexical gaps persist, approaches beyond straightforward fine-tuning (e.g., dialect-

aware pretraining or modeling) may be necessary, particularly for endangered dialects where ASR can support documentation and revitalization.

Limitations

As with most dialectal ASR studies, our results are shaped by practical constraints related to data availability, transcription conventions, and computational resources. The limitations outlined below contextualize the reported WER and CER scores and highlight directions in which future work could improve reliability, comparability, and linguistic interpretability.

Orthography and transcription variability.

None of the examined dialects has a fully standardized orthography, and transcriptions therefore necessarily reflect local conventions and annotator decisions. Although we aimed to keep transcriptions as close as possible to SMG orthography, including systematic choices for northern vocalism in Lesbian, residual inconsistencies remain and may inflate both WER and CER, particularly for varieties with larger lexical and phonological gaps from SMG. This issue is especially salient for Cappadocian, where extensive contact-induced and internal restructuring places additional pressure on spelling conventions, and where multiple transcribers may adopt slightly different practices. Moreover, inter-annotator agreement was not formally measured, which limits our ability to quantify transcription consistency and disentangle annotation-related variability from model-induced errors.

Heterogeneous transcription pipelines. In addition to orthographic variability across dialects, the benchmark currently includes subcollections produced via different annotation workflows (e.g., direct manual transcription vs. manual correction of ASR-bootstrapped drafts). Even when the same transcription policy is targeted, workflow differences may yield subtle systematic effects (e.g., segmentation preferences or persistence of model-like spellings). Users who wish to train on the benchmark as a single pooled dataset should therefore treat transcription workflow as a potential confound and consider (i) training and evaluating per dialect, (ii) adding workflow metadata as a control variable, and/or (iii) applying post-hoc normalization of segmentation and tokenization conventions. Extending the benchmark with parallel

re-annotation under a single unified pipeline is an important direction for future work.

Data size, speaker coverage, and domain differences. While the benchmark expands dialectal coverage, the amount of processed audio remains limited for some varieties (e.g., approximately one hour for Aperathiot and Lesbian, and about 1h17m for Cappadocian), and speaker diversity is uneven across datasets. For instance, the Western Cretan subset is based on recordings from a small number of speakers, whereas the Eastern Cretan data originate from radio broadcasts. These factors constrain the robustness of fine-tuning results and may limit generalization to broader speaker populations and communicative contexts within each dialect.

Limited error diagnosis beyond aggregate metrics. Evaluation is based on aggregate WER and CER computed over normalized text (lowercased, punctuation removed). While this enables consistent quantitative comparison, it does not reveal which specific dialectal phenomena, such as phonological alternations, morphological variation, lexical borrowing, or word boundary effects, contribute most to recognition errors. A more detailed qualitative error analysis could provide deeper insight into model behavior and better inform future adaptation strategies (see, for example, [Parsons et al., 2023](#); [Blaschke et al., 2025](#)).

Speaker overlap across splits. The training, development, and test partitions are constructed at the utterance level, which implies that the same speakers appear in all splits. This split strategy is common in low-resource ASR benchmarks, but it can lead to optimistic scores because models may partially leverage speaker-specific characteristics observed during training. As a result, these results primarily reflect within-speaker generalization under dialectal mismatch; evaluating cross-speaker robustness would require speaker-disjoint splits, which we leave for future benchmark extensions and ablation experiments.

Computational constraints. Computational resources limited the range of adaptation strategies explored. In particular, large models such as Whisper Large-v2/Large-v3 and Omnilingual ASR were evaluated only in zero-shot settings, and fine-tuning was restricted to Whisper-medium and XLS-R-53-greek. As a result, we did not investigate more computationally demanding approaches

such as continued pretraining, larger-scale hyperparameter optimization, or dialect-aware model variants, which may be necessary for highly divergent varieties such as Cappadocian.

Ethical Considerations

All speech data used in this study were collected with informed consent from participants and anonymized prior to inclusion in the benchmark. Personal names and identifiable information were removed at both the audio and transcription levels through manual review by trained linguists, and no sensitive personal data are released. In the case of the Cappadocian recordings, which were collected during earlier fieldwork, consent for research use and subsequent reuse of the data was obtained by the original fieldworker.

Several of the examined varieties are low-resource or endangered (notably Cappadocian), and the goal of this work is to support linguistic documentation and technological inclusion rather than deployment in real-world applications. The reported ASR models exhibit high error rates for several dialects, and their outputs should not be interpreted as suitable for practical use or decision-making involving speakers of these varieties.

Finally, dialectal ASR systems may reproduce or amplify existing linguistic biases if deployed without adequate adaptation and evaluation. We therefore frame this benchmark as a research resource aimed at understanding model limitations and guiding future work, rather than as an endorsement of current systems for dialectal speech processing.

Acknowledgments

This work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program.

References

- Aditya Yadavalli and Ganesh Mirishkar and Anil Kumar Vuppala. 2022. [Multi-Task End-to-End Model for Telugu Dialect and Speech Recognition](#). In *Interspeech 2022*, pages 1387–1391.
- Renas Adnan and Hossein Hassani. 2025. [Which one Performs Better? Wav2Vec or Whisper? Applying both in Badini Kurdish Speech to Text \(BKSTT\)](#). *arXiv:2508.09957*.

- Vasileia Alexelli. 2021. *Chartografisi tis glossikis poikilias tis Lesvou [Mapping the linguistic variety of Lesbos]*. Ph.D. thesis, University of Patras, School of Humanities and Social Sciences, Department of Philology, Linguistics Section.
- Tanel Alumäe, Jiaming Kong, and Daniil Robnikov. 2023. *Dialect Adaptation and Data Augmentation for Low-Resource ASR: TalTech Systems for the MADASR 2023 Challenge*. *Preprint*, arXiv:2310.17448.
- Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. *XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale*. *Preprint*, arXiv:2111.09296.
- Andrei Barcovschi, Rishabh Jain, and Peter Corcoran. 2023. *A comparative analysis between Conformer-Transducer, Whisper, and wav2vec2 for improving the child speech recognition*. *Preprint*, arXiv:2311.04936.
- Vivek Bhardwaj, Vinay Kukreja, Navjeet Kaur, and Nandini Modi. 2021. *Building an ASR System for Indian (Punjabi) language and its evaluation for Malwa and Majha dialect: Preliminary Results*. In *2021 12th International Conference on Computing Communication and Networking Technologies (ICCN-T)*, pages 1–5.
- Verena Blaschke, Miriam Winkler, Constantin Förster, Gabriele Wenger-Glemser, and Barbara Plank. 2025. *A Multi-Dialectal Dataset for German Dialect ASR and Dialect-to-Standard Speech Translation*. In *Interspeech 2025*, pages 913–917. ISCA.
- Paul Boersma and David Weenink. 2001. Praat: Doing phonetics by computer. Computer program.
- Stavros Bompolas. 2023. *Computational dialectology in the linguistic varieties of Cappadocian, Phara-siot, and Silliot*. Ph.D. thesis, University of Patras, School of Humanities and Social Sciences, Department of Philology, Linguistics Section.
- Stavros Bompolas, Stella Markantonatou, Angela Ralli, and Antonios Anastasopoulos. 2025. *Crossing Dialectal Boundaries: Building a Treebank for the Dialect of Lesbos through Knowledge Transfer from Standard Modern Greek*. In *Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*, pages 39–51, Ljubljana, Slovenia. Association for Computational Linguistics.
- Georgios Chatzidakis. 1905. *Mesaionika kai Nea Ellinika [Medieval and Modern Greek]*, volume 12. Sakellarios, Athens.
- William Chen, Chutong Meng, Jiatong Shi, Martijn Bartelds, Shih-Heng Wang, Hsiu-Hsuan Wang, Rafael Mosquera, Sara Hincapie, Dan Jurafsky, Antonis Anastasopoulos, Hung yi Lee, Karen Livescu, and Shinji Watanabe. 2025. *The ML-SUPERB 2.0 Challenge: Towards Inclusive ASR Benchmarking for All Language Varieties*. In *Interspeech 2025*, pages 2093–2097.
- Richard MacGillivray Dawkins. 1916. *Modern Greek in Asia Minor: a study of the dialects of Silli, Cap-padocia and Phárasa with grammar, texts, translations and glossary*. Cambridge University Press, Cambridge.
- Antonios Dimakis, John Pavlopoulos, and Antonios Anastasopoulos. 2025. *Dialect Normalization using Large Language Models and Morphological Rules*. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23696–23714, Vienna, Austria. Association for Computational Linguistics.
- Luca Ducceschi and Greta H. Franzini. 2025. *Speech transcription from South Tyrolean Dialect to Standard German with Whisper*. In *Interspeech 2025*, pages 1–5. ISCA.
- Vardis Georgilas and Themis Stafylakis. 2025. *Automatic Speech Recognition for Greek medical dictation*. *Preprint*, arXiv:2509.23550.
- Zachary Hopton and Eleanor Chodroff. 2025. *The Impact of Dialect Variation on Robust Automatic Speech Recognition for Catalan*. In *Proceedings of the 22nd SIGMORPHON workshop on Computational Morphology, Phonology, and Phonetics*, pages 23–33, Albuquerque, New Mexico, USA. Association for Computational Linguistics.
- Ryo Imaizumi, Ryo Masumura, Sayaka Shiota, and Hitoshi Kiya. 2020. *Dialect-Aware Modeling for End-to-End Japanese Dialect Speech Recognition*. In *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (AP-SIPA ASC)*, pages 297–301.
- Sukairaj Hafiz Imam, Tadesse Destaw Belay, Kedir Yassin Husse, Ibrahim Said Ahmad, Idris Abdulummin, Hadiza Ali Umar, Muhammad Yahuza Bello, Joyce Nakatumba-Nabende, Seid Muhie Yimam, and Shamsuddeen Hassan Muhammad. 2025. *Automatic Speech Recognition (ASR) for African Low-Resource Languages: A Systematic Literature Review*. *Preprint*, arXiv:2510.01145.
- Petros Karatsareas. 2011. *A study of Cappadocian Greek nominal morphology from a diachronic and dialectological perspective*. Publisher: Apollo - University of Cambridge Repository.
- Haneen Bahjat Khalafallah, Mohamed Abdel Fattah, and Ruqayya Abdulrahman. 2024. *Speech corpus for Medina dialect*. *Journal of King Saud University - Computer and Information Sciences*, 36(2):101864.
- Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Touns, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. *Racial disparities in automated speech recognition*. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

- Nikolaos G. Kontosopoulos. 2008. *Dialektoi kai idiomata tis Neas Ellinikis [Dialects and Idioms of Modern Greek]*. Ekdoseis Grigori, Athens.
- Olli Kuparinen. 2025. [Automatic Dialectal Transcription: An Evaluation on Finnish and Norwegian](#). In *Interspeech 2025*, pages 2390–2394. ISCA.
- Seham Nasr, Rehab Duwairi, and Muhannad Quwaider. 2023. [End-to-End Speech Recognition For Arabic Dialects](#). *Arabian Journal for Science and Engineering*, 48(8):10617–10633.
- Dimitrios V. Oikonomidis. 1952. *Peri tou glossikou idiomatos Aperathou–Naksou [On the Linguistic Idiom of Aperathos, Naxos]*. Typografeion Myrtidou, Athens.
- Omnilingual ASR Team, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, Kevin Chan, Chierh Cheng, Joe Chuang, Caley Droof, Mark Dupenthaler, Paul-Ambroise Duquenne, Alexander Erben, Cynthia Gao, Gabriel Mejia Gonzalez, and 14 others. 2025. [Omnilingual ASR: Open-Source Multilingual Speech Recognition for 1600+ Languages](#). *Preprint*, arXiv:2511.09690.
- Georgios Emmanouil Pangalos. 1955. *Peri tou glossikou idiomatos tis Kritis, itoi diagramma grammatikis kai glossarion tou simerinou glossikou idiomatos tis Kritis [On the Linguistic Idiom of Crete: A Grammatical Outline and Glossary of the Contemporary Cretan Dialect]*. n.p., Athens.
- Phoebe Parsons, Knut Kvale, Torbjørn Svendsen, and Giampiero Salvi. 2023. A character-based analysis of impacts of dialects on end-to-end Norwegian ASR. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 467–476.
- Siqing Qin, Longbiao Wang, Sheng Li, Jianwu Dang, and Lixin Pan. 2022. [Improving low-resource Tibetan end-to-end ASR by multilingual and multi-level unit modeling](#). *EURASIP Journal on Audio, Speech, and Music Processing*, 2022(1):2.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. [Robust Speech Recognition via Large-Scale Weak Supervision](#). *Preprint*, arXiv:2212.04356.
- Angela Ralli. 2015. Strategies and Patterns of Loan Verb Integration in Modern Greek Varieties. In Angela Ralli, editor, *Contact Morphology in Modern Greek Dialects*, pages 73–88. Cambridge Scholars Publishing.
- Angela Ralli. 2019a. [Affixoids and Verb Borrowing in Aivaliot Morphology](#). In Angela Ralli, editor, *The Morphology of Asia Minor Greek*, pages 221–254. BRILL.
- Angela Ralli. 2019b. [Greek in Contact with Romance](#). In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press.
- Jiatong Shi, Shih-Heng Wang, William Chen, Martijn Bartelds, Vanya Bannihatti Kumar, Jinchuan Tian, Xuankai Chang, Dan Jurafsky, Karen Livescu, Hung yi Lee, and Shinji Watanabe. 2024. [ML-SUPERB 2.0: Benchmarking Multilingual Speech Models Across Modeling Constraints, Languages, and Datasets](#). *Preprint*, arXiv:2406.08641.
- Clement Sicard, Kajetan Pyszkowski, and Victor Gillioz. 2023. [Spaiche: Extending State-of-the-Art ASR Models to Swiss German Dialects](#). *Preprint*, arXiv:2304.11075.
- Naoki Takahashi, Shogo Miwa, Yuta Kamiya, Takumi Toyama, Raufun Nahar, and Atsuhiko Kai. 2024. [Comparison of Large Pre-trained Models and Adaptation Methods for Japanese Dialects ASR](#). In *2024 IEEE 13th Global Conference on Consumer Electronics (GCCE)*, pages 811–814.
- Sarah Grey Thomason and Terrence Kaufman. 1988. *Language contact, creolization, and genetic linguistics*, 1. paperback print edition. Univ. of California Press, Berkeley.
- Melissa Torgbi, Andrew Clayman, Jordan J. Speight, and Harish Tayyar Madabushi. 2025. [Adapting Whisper for Regional Dialects: Enhancing Public Services for Vulnerable Populations in the United Kingdom](#). *Preprint*, arXiv:2501.08502.
- T. Toyama, A. Kai, Y. Kamiya, and N. Takahashi. 2024. [Adapting Large-Scale Pre-trained Models for Unified Dialect Speech Recognition Model](#). *Acta Physica Polonica A*, 146(4):413–418.
- Peter Trudgill. 2003. [Modern Greek dialects: A preliminary classification](#). *Journal of Greek Linguistics*, 4(1):45 – 63. Number: 1 Place: Leiden, The Netherlands Publisher: Brill.
- Chara Tsoukala, Kosmas Kritsis, Ioannis Douros, Athanasios Katsamanis, Nikolaos Kokkas, Vasileios Arampatzakis, Vasileios Sevetlidis, Stella Markantonatou, and George Pavlidis. 2023. [ASR pipeline for low-resourced languages: A case study on pomak](#). In *Proceedings of the Second Workshop on NLP Applications to Field Linguistics*, pages 40–45, Dubrovnik, Croatia. Association for Computational Linguistics.
- UNESCO, editor. 2010. *Atlas of the World's Languages in Danger*. United Nations Educational, Scientific and Cultural Organization, Paris.
- Socrates Vakirtzian, Vivian Stamou, Yannis Kazos, and Stella Markantonatou. 2025. [Dialectal treebanks and their relation with the standard variety: The case of East Cretan and Standard Modern Greek](#). In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th*

Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 776–784, Tallinn, Estonia. University of Tartu Library.

Socrates Vakirtzian, Chara Tsoukala, Stavros Bompolas, Katerina Mouzou, Vivian Stamou, Georgios Paraskevopoulos, Antonios Dimakis, Stella Markantonatou, Angela Ralli, and Antonios Anastasopoulos. 2024. *Speech Recognition for Greek Dialects: A Challenging Benchmark*. In *Proc. Interspeech 2024*, pages 3974–3978.

Juan Camilo Vásquez-Correa and Aitor Álvarez Muñain. 2023. *Novel Speech Recognition Systems Applied to Forensics within Child Exploitation: Wav2vec2.0 vs. Whisper*. *Sensors*, 23(4).

Aiden Williams, Andrea Demarco, and Claudia Borg. 2023. *The Applicability of Wav2Vec2 and Whisper for Low-Resource Maltese ASR*. In *2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL 2023)*, pages 39–43. ISCA.

Maria Xefteri. 2009. *Koinonioglossiki proseggisi tou idiomatos t' Aperathou Naksou (opos diatireitai stin Athina) [(A Sociolinguistic Approach to the Aperathou Dialect of Naxos as Preserved in Athens)]*. In *Proceedings of the 4th Graduate Student Meeting of the Department of Philology, University of Athens*, Athens, Greece. Department of Philology, National and Kapodistrian University of Athens. Available at Academia.edu.

How Should We Model the Probability of a Language?

Rasul Dent¹, Pedro Ortiz Suarez², Thibault Clérice¹, Benoît Sagot¹

¹Inria, Paris, {firstname.lastname}@inria.fr

²Common Crawl Foundation, Paris, pedro@commoncrawl.org

Abstract

Of the over 7,000 languages spoken in the world, commercial language identification (LID) systems only reliably identify a few hundred in written form. Research-grade systems extend this coverage under certain circumstances, but for most languages coverage remains patchy or nonexistent. This position paper argues that this situation is largely self-imposed. In particular, it arises from a persistent framing of LID as decontextualized text classification, which obscures the central role of prior probability estimation and is reinforced by institutional incentives that favor global, fixed-prior models. We argue that improving coverage for tail languages requires rethinking LID as a routing problem and developing principled ways to incorporate environmental cues that make languages locally plausible.

1 Introduction

To use many Natural Language Processing (NLP) systems, we must first specify the language. The task of inferring this information is known as automatic language identification (LID). Following substantial progress in the 1990s, McNamee (2005) famously described LID for lengthy European-language documents as “a solved problem suitable for undergraduate instruction.” Attention has since shifted to closely-related varieties (Aepli et al., 2023), rare languages (Caswell et al., 2020), short texts (Murthy and Kumar, 2006), and code-switching (Burchell et al., 2024).

Despite improvements on benchmarks, the practical reality for most languages has changed little in the past 20 years. New models are occasionally released for a region or family (e.g. Adebara et al., 2022), and some attempt to reach “the next 1000 languages” (e.g. Brown, 2013; Kargaran et al., 2023). Nonetheless, production-grade LID systems at companies like Meta and Google still focus on only a few hundred widely spoken languages.

This stagnation reflects two core issues. First, in research settings, LID is typically framed as a decontextualized inference task, in which systems aim to map directly text to a label from a global set. This encourages approaches that perform well on benchmarks but fail when applied to real data. Second, field-level incentive structures reward developing *novel* methods over revisiting basic tools.

We argue that recentering context will be crucial for expanding the *effective* coverage of LID. For “local languages” (see Bird, 2022) in particular, success should be defined relative to local contexts and resource constraints. This entails rethinking the values that shape research in LID.

2 The Received Framing of LID

At its core, LID is intended to route content to users who manage different languages. Due to differences between the modalities of text, speech, and sign, LID in each modality is generally approached separately, with text-based LID typically viewed as the easiest (Jauhiainen et al., 2024). For Rau (1974), the prototypical LID operator was a data entry clerk filing documents in unfamiliar languages for human patrons. With the explosion of born-digital content, the immediate recipient is now often a computer program.¹

2.1 Standard Approaches

LID is typically approached as a supervised classification problem (Jauhiainen et al., 2024). This framing builds in two critical assumptions:

1. Labels represent a global hypothesis space.
2. Inference can be performed solely from text.

As in other classification tasks, there are two main modeling approaches. Per-class approaches

¹However, Rau’s vision still lives on in some institutions, like libraries.

such as Multinomial Naïve Bayes (e.g. [Lui and Baldwin, 2012](#)) fit the data to each class independently, and take the best fitting class as the label.² In contrast, discriminative models like fastText ([Joulin et al., 2017](#); [Grave et al., 2018](#)) and CLD3 ([Salcianu et al., 2023](#)) learn decision boundaries between all classes simultaneously.

Hierarchical architectures also garnered interest. Many models, such as IndicLID ([Madhani et al., 2023](#)), are region-specific and depend on external software to ensure that they are used over the correct language set. Others, especially those based on fastText such as OpenLID ([Burchell et al., 2023](#)) and GlotLID ([Kargaran et al., 2023](#)), take a flat approach and model a wide variety of languages using one combined model. Yet others, such as LIMIT ([Agarwal et al., 2023](#)), try to learn hierarchical classification schemas directly from model errors.

2.2 Alternative Framings

The standard assumptions, and the modeling constraints that come with them, are not unreasonable in many common scenarios. In particular, they produce strong results on well-written monolingual documents in widely spoken languages ([McNamee, 2005](#)). However, their reliability deteriorates in several scenarios. [Caswell et al. \(2020\)](#) and [Kreutzer et al. \(2022\)](#) showed that the models that perform the best on common benchmarks often struggle when applied to noisy web data. Similarly, when working with closely-related varieties, or very short texts, it is often impossible to select just one correct label on the basis of input alone. The recent Shared Task on Improving Language Identification for Web Text at the 1st Workshop on Multilingual Data Quality Signals confirmed these issues remain relevant ([Suarez et al., 2026](#)).

In response to such issues, alternate framings have been proposed. For example, [Baimukan et al. \(2022\)](#) show that hierarchical labels are important for fine-grained dialect classification. [Bernier-Colborne et al. \(2023\)](#) and [Keleg and Magdy \(2023\)](#) extend this insight, reconceptualizing the task of dialect identification as one of multi-label classification. From a different angle, [Dent et al. \(2025\)](#) contend that building web corpora for rare languages is more of a mining task than strict classification.

Together, these reframings touch on a much more general issue, but do not completely resolve

²Traditionally, this is called generative modeling, but as an anonymous reviewer noted, the term has become ambiguous.

it. Namely, general-purpose LID models need a label set fixed enough to allow training and flexible enough to handle great variation in inference-time granularity. In Section 3, we argue that the diversity of inference-time conditions has often been overlooked, leading to structural limitations with significant practical and scientific costs.

3 Probability Problems

Both per-class and discriminative approaches to LID ultimately estimate a conditional probability distribution: given input features X , estimate the probability P of each language ℓ , and choose the most likely label. Bayesian framing reveals that $P(\ell | X)$ actually depends on two terms, the probability of X given ℓ , and the prior probability of ℓ itself. The decision rule can be written in the familiar argmax form:

$$\hat{\ell} = \arg \max_{\ell} P(X | \ell) P(\ell), \quad (1)$$

or equivalently in log-space:

$$\hat{\ell} = \arg \max_{\ell} [\log P(X | \ell) + \log P(\ell)]. \quad (2)$$

This decomposition highlights a highly non-trivial modeling problem. Namely, what does $P(\ell)$ actually represent?

3.1 Global Frequency?

In the classic Bayesian formulation, the answer is simple enough. $P(\ell)$ is the number of texts in a given language over the number of examples in the entire corpus. In principle, this should correspond to a real difference in the frequency of languages, rather than a sampling artifact. However, under naïve frequency-based estimates, we immediately face two related problems when dealing with massive class imbalances.

1. Rare classes become nearly undetectable.
2. Ignoring rare classes has minimal effect on macro-performance.

To see this concretely, suppose we compare English ($\ell = \text{en}$) with a rare language ($\ell = r$). Let the global priors be:

$$P(\text{en}) = 0.40, \quad P(r) = 10^{-6}.$$

Assume for illustration that the likelihoods of a given text under the two languages are of comparable magnitude—for example:

$$P(X | \text{en}) = 10^{-4}, \quad P(X | r) = 10^{-2}.$$

Even though the rare-language model assigns a *hundred times higher* likelihood to the input text, the posterior still overwhelmingly favors English:

$$P(X | \text{en})P(\text{en}) = 10^{-4} \cdot 0.40 = 4 \times 10^{-5},$$

$$P(X | r)P(r) = 10^{-2} \cdot 10^{-6} = 10^{-8}.$$

Thus,

$$4 \times 10^{-5} \gg 10^{-8},$$

and the argmax will select English. The only way the likelihood evidence can compensate for a prior that is effectively zero is if the likelihood for more common languages is *also* effectively zero (likely due to a script mismatch). Because language r occurs less than a thousandth of the time, getting it completely wrong will barely show up as a rounding error in overall accuracy.

3.1.1 Attenuated Frequency?

Due to the extreme distribution of true frequencies, one common mitigation strategy is to upsample rare classes and/or downsample common ones (Burchell et al., 2023). This can help prevent a low $P(\ell)$ from instantly overpowering the likelihood evidence. However, even with this, there is a double bind. On the one hand, the prior still favors classifying data without obvious hints under a dominant class. On the other hand, the model can rely on a few shallow features and thus lack robustness when such features appear as noise in real data. These two issues have led recent studies to emphasize False Positive Rate (FPR) as a key metric.

3.1.2 False Positives at Scale

To contextualize this issue, we revisit the example of Caswell et al. (2020), who note that a LID classifier that upsamples r enough to achieve 99% precision and 99% recall on a balanced data, could be expected to recover 9.9K out of 10K true positives. However, if this is applied at web scale over 100B documents with a 0.01% FPR, the resulting 10M false positives would render the resulting corpus unusable. Thus, adjusting training data ratios and

carefully choosing metrics can reduce the damage, but they do not solve the underlying issue.

Naturally, one is tempted to ask where these false positives come from. While the obvious answers of high-resource and out-of-model cousins do indeed play a role, Caswell et al. (2020) also document various forms of noise that are difficult to *thoroughly* account for at training time. While a full review is beyond the scope of this article, important classes include emoji, misrendered PDFs, non-Unicode fonts, chance n-gram overlaps between unrelated languages, ‘A N T S P E A K’, and mixed-script boilerplate. In response to this, they, as well as later filtered crawls like Fineweb-2 (Penedo et al., 2025), emphasize *post-filtering* for document-level consistency and token-precision in minority languages. Once again, this mitigates the fallout, but does not solve the issue.

3.2 Local Priors?

Once we appreciate the magnitude of the problem of false positives at scale, the central question is no longer how well languages can be discriminated in the abstract, but rather how to avoid suggesting labels that are not even plausible candidates in a given context. We must first remember that, even for widely spoken languages, global frequency is often irrelevant. For instance, English, Mandarin, and Russian each account for a substantial share of web text, but each is typically written in a distinct script. Observing the script instantly render demographic dominance of the others nearly meaningless, and Bayesian models like CLD2 often incorporate preliminary script classification for this very reason.

In theory, similar effects should arise when incorporating metadata such as geography, social networks, or user feedback. These contextual signals are useful in general, but crucial for assigning usable prior probabilities to local languages, as well as managing the sheer size of the label inventory. In practice, however, converting context into usable prior probabilities remains challenging. Although some recent work has explored the utility of geodata (Dunn and Edwards-Brown, 2024), current benchmarks make it difficult to compare such context-aware models. The challenge is not simply that contextual information is missing; rather, the logistical requirements are fundamentally incompatible with how multilingual datasets are generally collected, labeled, and shared today.

3.2.1 Where Do Languages Live?

Firstly, there is the question of whether tail languages are actually used enough that misclassification is a practical, rather than theoretical, issue. Recently, many languages that have traditionally been considered “unwritten”, ranging from European varieties like Luxembourgish (Wagner, 2013) to the Zapotec languages of Mexico (Lillehaugen, 2019), are increasingly written (and spoken) on social media platforms and messaging services. Over the last few years, both and privacy regulation and contractual restrictions have curtailed research access, but there is reason to believe this trend continues.

3.2.2 Dataset Difficulties

Given the limited availability of user data, public LID datasets turn to sources that are easier to work with, like Wikipedia, the Bible, or newswire.³ Here, providing metadata about sources, by saying for example that the passage comes from a Saint Lucian Bible, would often solve the task, so the source is not given. Even the full, document-length text would still be to easy, so the tendency is to simulate harder scenarios by decomposing data from these sources into short, freestanding snippets, as in FLORES-200 (NLLB_Team et al., 2022) and GlotLID-C (Kargaran et al., 2023).

These practices can lead to a non-trivial circularity problem. Creating reliable labels for languages beyond the top 100–200 demands annotators with broad linguistic coverage, including familiarity with lesser-known scripts and regional varieties, as exemplified by Kreutzer et al. (2022). Even when such expertise can be assembled, annotators typically lack access to contextual metadata for the reasons we have just discussed. In practice, this is mitigated by beginning with pre-compiled corpora, which cover precisely the kinds of data for which LID already works.

3.2.3 Model Expectations

Older per-class models like CLD2 allowed users to provide hints like the expected language(s), URL sources, top-level domain name, original encoding, and embedded language tags. In contrast, newer discriminative classifiers like fastText and CLD3 that are deeply embedded in industrial and academic preprocessing pipelines cannot easily take local context into account. While it is possible to restrict the output space of such models to a few

³Although even these are increasingly the subject of copyright challenges

candidate languages, this involves an opaque manipulation of the softmax layer that can lead to unstable behavior.

3.2.4 Towards Context-Aware LID

In light of these operational difficulties, it is entirely understandable that public benchmarks do not focus on the relationship between the idealized task and real-world deployments. However, this encourages the development of models optimized for an version of LID that simultaneously overestimates the difficulty posed by rare languages in their local contexts and underestimates the impact of noise at scale. This makes it difficult to justify context-aware methods that target certain kinds of confusion that mainly surface when raw data is treated across diverse global sources. In Section 5, we will argue that this technical misalignment is reinforced by powerful institutional incentives, which further entrench the decontextualized framing of LID. However, we first present two case studies in Section 4 which help illustrate the problem.

4 Case Studies

As a brief demonstration, we briefly consider two contact languages where the global prior probability is effectively zero, resulting in exclusion from all SOTA models. The first is Louisiana Creole (LC), a critically endangered language used online in the context of language revitalization. The second is the now-extinct Lingua Franca, long suspected to have influenced contact languages around the world.

4.1 Language Revitalization

Following over a century of language shift to English, LC is spoken today by a few thousand people. However, there is a revitalization movement with a strong online presence on several social media platforms, such as Facebook and Discord. In revitalization spaces, the main languages are LC and English. Some groups tolerate discussion of potentially relevant languages (Louisiana) French, Spanish, and other French-based Creoles, while others, like the Louisiana Creole Virtual Classroom prefer a strictly bilingual learning environment (Mayeux, 2024). In any case, $P(LC)$ is much higher in dedicated groups than its global rarity would suggest.

However, as the LID label set for existing models is defined at training time, the modeling assumption is that $P(LC) = 0$. When downstream services like auto-translate are activated, the language



Figure 1: Interfaces allow override, but not for LC. (Google Translate)

is frequently misrecognized as French, which predictably leads to misunderstandings. Some platforms now have options to rate translations, guide the machine translation to a slightly closer target (i.e. Haitian Creole) or disable auto-translation. However, at the time of writing, there is no user-facing option to identify a sentence as LC, reflecting the absence of LC from the underlying LID label set. As a result, users can mitigate downstream effects but cannot directly correct the initial language identification.

To explore this issue, we present a sample from Google Translate in Figure 1 that is functionally similar to social media, but deliberately not taken from a real group due to privacy considerations. The input text is *Mo sòr fé ègzemp-çila pou montré kofè li marsh pa*⁴. The language is misidentified as Bambara and incorrectly translated as ‘I’m sure I did example to show coffee li marsh pa’.

What is of interest here is not the mistranslation, which is expected since the language is not supported, but rather the LID interface. Previous work by (Yang et al., 2025a) has taken the absence of Native American languages as labels in Google Translate (and similar industrial models) as an indication that LID is a bottleneck. Yet, there *is* an option for manual override. Crucially, this only extends to fully supported languages like Haitian Creole and French. Moreover, we are able to create a shortlist of languages that have recently been used, and are thus presumed to appear again. Although the shortlist does not seem to bias the LID classifier, this interface showcases that for common languages, manual override is accepted and does not require a 7,000 language menu every time. Once we observe this capacity for manual selection, it becomes clearer that the question is not necessarily whether languages can be recognized, but rather whether recognition can be decoupled from full support. We return to this issue in Sections 5.2 and 6.3.

⁴I just made this example to show why it does not work’

4.2 Retracing Language Contact

Whinnom (1956) once argued that the Mediterranean Lingua Franca (LF) may have been an inspiration for contact varieties in Africa and Asia. Although the general idea has been largely rejected since the 1980s, Nolan (2023) claims LF may have influenced the 19th-20th century French military register known as Français Tirailleur (FT). However, it has been extremely difficult to validate even this limited claim two main reasons. Firstly, LF was a rarely-written low status contact language based on — yet distinct from — prestige Romance languages like Italian and Spanish. Secondly, it was mostly used in North Africa and the Levant, where local writing was typically in the Arabic script and furthermore has not necessarily been digitized to the same extent.

Overall, current LID systems are not optimized for recovering historical contact varieties as heterogeneous as LF, and existing models do not attempt to. Even if they did, the small extant corpus and high lexical overlap with much better documented Romance varieties as well as Romanized Arabic would likely lead to confusion with these more frequent categories.

However, we have found reports which show that the French military served as a powerful vector for the global transmission of LF, also known as Sabir, following the French occupation of Algiers. In particular, the *Relation historique de la campagne du Sénégal (1861) : extrait du journal d’un capitaine de tirailleurs algériens*⁵ (Dubois, 1863) includes several passages of dialogue in the contact variety during an expedition in Senegal around the time FT is thought to have developed. For example, *Toi courageux kif r’azaloun, toi parlar kif barougue, toi tornar M’Boro, toi fasir meillour Lackmar*.⁶ In this excerpt alone, the pronouns *toi* ‘you’ is French, the names of the animals are Arabic, the verbs are Mediterranean Romance, and M’Boro is the name of a town in Senegal.

In a decontextualized setting, this combination would look like a code-switching nightmare. Yet, in the context of the 19th French *Armée d’Afrique* ‘Army of [North] Africa’, the main task is to distinguish LF from French and to a much lesser extent, Arabic. This is a much more tractable sub-

⁵Historical Relation of the Senegal Campaign (1861) : Excerpt from the Journal of a Captain of the Tirailleurs Algériens

⁶You who are courageous like a gazelle and speak like a parrot, you go back to M’Boro, you will do better than Lackmar.

problem. In doing so, we identify a historically plausible mechanism of transmission in the form of Sabir-speaking Algerian troops transported to West Africa and placed in daily contact with newly conscripted Senegalese recruits. Although validating such historical connections is clearly feasible with lightweight NLP methods, such questions remain relatively marginal for reasons we explore in Section 5.

5 Why Is This Hard To Change?

Even if the limitations of decontextualized LID are clear, several distinct incentive structures across academia, industry, and public institutions prevent these problems from being addressed in practice.

5.1 Academic Incentives

Within academic NLP, it is often assumed that meaningful contributions should involve novel architectures, training strategies, or datasets. Demonstrating that a well-understood technique still works is typically viewed as incremental rather than interesting. As a result, researchers are structurally disincentivized from relying on simple or well-established methods. This pattern is reflected in persistent under-citation of earlier foundational work (Singh et al., 2023) and a broader tendency to treat re-engineering as innovation, had been amplified by the rapid growth of the field and the prestige hierarchy surrounding machine learning conferences.

More specifically, *Excitement* has been explicitly recognized as a key criterion for top-conference acceptance within the *ACL umbrella (Rogers et al., 2023). This is well suited to promoting major breakthroughs, but it also deprioritizes contributions that are merely *Sound*. Since LLMs are not yet a practical or cost-effective substitute for (shallow) general LID systems, there is an increasing gap between what the interests of the broader field and what actually works for this particular task.

To be clear, this dynamic does not strictly *prevent* work like van der Goot (2025) from being published in main conferences. However, it does tend to route LID papers toward Findings and especially specialized workshops. The issue is not the quality of these venues, but their position in the prestige hierarchy. Because such venues lack the visibility and career incentives associated with the main ACL-branded conferences, there is little incentive for early-career researchers to specialize

in LID, as opposed to other, “more contemporary” problems.

5.2 Commercial Incentives

Equally, if not more important, is the reward structure of private industry. Here, the primary issue is not necessarily novelty or prestige. Rather, it is that existing LID solutions work well enough for commercially attractive languages, and improving LID for tail languages does not clearly lead to a viable revenue stream. Thus, even though there have been efforts such as Caswell et al. (2020), LID is not often treated as a pressing issue in its own right, but rather as a routing component for machine translation (e.g. NLLB_Team et al., 2022), as explored in Section 4.

Furthermore, accurately identifying a language implicitly commits a platform to a certain level of support for that language. Once a system signals that it can reliably recognize a language, users may reasonably expect corresponding improvements in areas such as content moderation, abuse detection, policy enforcement, and ranking. However, fully supporting even one additional language often requires coordinated work across multiple teams such as policy, annotation, trust and safety, and regional compliance. This organizational effort can exceed the technical effort by a significant margin.

From a business perspective, the combination of increasing operational and compliance costs with no clear additional revenue is particularly discouraging. This in turn casts a large shadow on the data asymmetry problem, since, as noted in Sections 3.2.1 and 4.1, industry platforms often contain precisely the high-quality, contextual data for rare languages necessary to build and evaluate robust LID systems.

5.3 (Absence of) Vertical Integration

In principle, progress in ASR, transcription, text processing, and LID should be mutually reinforcing. ASR systems require textual resources; text collection is far more efficient when ASR exists; and accurate LID is a prerequisite for building both. This feedback loop operates naturally for many large languages.

For local languages, however, the previously named incentives lead to structural deadlock. Each component waits for progress in the others, and only the largest actors have incentive to coordinate the pipeline. For example, some government and national security agencies manage to overcome

these issues and build vertically integrated systems, they cannot release them. It is true that some commercial endeavors like [Omnilingual ASR team et al. \(2025\)](#) are starting to pay more attention to these interdependencies, but they have yet extend them to textual LID. Thus, even though public radio archives and contemporary broadcasts in principle provide enough material to put many languages into medium-resource territory ([Doumbouya et al., 2021](#); [Havard et al., 2025](#)), a lack of vertical integrations renders many of these resources effectively unusable for the moment.

6 What Can We Do?

In Section 3, we explored how using invariant priors poses special problems for languages that are globally rare but locally common. In Section 4 we saw that such languages present particular challenges across social media and archives that are not insurmountable at the local level. In Section 5, we explored some institutional obstacles to incorporating contextual cues in LID models. Now, we pivot from describing problems to outlining alternative solutions for these interwoven issues.

6.1 Two Coherent Paths Forward

In introductory computer science, we are explicitly taught to place correctness at the center of program design. A large part of correctness concerns preempting edge cases and testing them. In LID, we recognize that most languages are edge cases, and simply eliminate them the label space during training. This is often justified as a necessity caused by data and/or model limitations. But is doing so correct?

The answer depends on one’s conception of LID. If one accepts, as we have observed, that LID typically functions to route content within a larger system, then success consists of reliably moving content from Source A to Target B. This framing naturally treats LID as an engineering problem where solutions are evaluated relative to downstream requirements. If, however, we wish to maintain that LID remains a scientific problem, then it is imperative that the determination of prior plausibility be treated as problem in its own right.

To make the role of prior specification concrete, we briefly sketch how local context can be incorporated into a standard per-class language identification model without introducing new architectures or training procedures. We define a *local frame of*

reference as any mechanism that restricts the hypothesis space of candidate languages or modifies their prior probabilities before observing the text itself, based on external signals such as document provenance, interface-level language settings, discourse continuity, previous inferences, or explicit user hints. Within such a frame, language identification consists of selecting the most plausible language relative to a locally valid prior distribution, rather than with respect to a single, globally fixed prior over all languages.

Concretely, let us (re)consider the Naive Bayes LID model represented in Section 3. Under a local frame of reference, the global prior, which implicitly encodes assumptions about worldwide language frequency, is replaced by a context-dependent distribution $P_{\text{local}}(\ell)$, yielding

$$\hat{\ell} = \arg \max_{\ell \in \mathcal{L}_{\text{local}}} [\log P(X | \ell) + \log P_{\text{local}}(\ell)].$$

In the simplest case, this local context can act as a hard gate: languages deemed implausible in the current setting are removed from the hypothesis space, while remaining languages receive a uniform local prior. More intricate formulations allow for soft gating, in which locally plausible languages are upweighted relative to a global prior, while others retain a small but non-zero probability mass. Importantly, both variants substitute local priors for global ones at inference time, and do not require retraining the underlying language models. For coarse metadata on most languages, including alternate names, [LinguaMeta \(Ritchie et al., 2024\)](#) provides a usable starting point. More localized data, such as placing Native Alaskan languages in Alaska specifically, would likely help limit false positives (cf. [Yang et al., 2025b](#)).

We recognize that discrimination between very similar varieties remains challenging, and that even creating likelihood models requires a certain amount of data. This approach is fully compatible with hierarchical and multi-label setups that leverage specific similarities and differences between related languages for fine-grained classification.

6.2 Evaluation and Benchmarks

One implication of the fork outlined above is that current evaluation practices conflate two distinct questions: whether a system can discriminate among languages given a fixed hypothesis space, and whether that hypothesis space is itself appropriate for a given use case. Under both paths, this

conflation obscures the source of error and limits meaningful comparison.

As mentioned in Section 6.1, if LID is treated primarily as an engineering problem, then evaluation should be aligned with deployment conditions. Rather than optimizing for global accuracy over a fixed label set, systems should be assessed relative to the routing decisions they are intended to support, including downstream error tolerance, latency constraints, and the availability of auxiliary signals. In this setting, comparisons across platforms or domains are not expected to be universal. Instead, success is defined locally, in terms of whether content is routed to an appropriate handler with acceptable cost.

If, by contrast, LID is to remain a scientific task, then evaluation must explicitly separate prior determination from text-based inference. Benchmarks that present decontextualized text snippets implicitly encode an often arbitrary prior of convenience over the label space based on the available data. A scientifically meaningful alternative would require benchmarks in which non-linguistic signals are exposed as inputs and evaluated in their own right. This would allow models to be compared on their ability to restrict the hypothesis space before observing the text, as well as on their performance once that restriction has been applied.

Crucially, such benchmarks need not rely on sensitive or proprietary user data. As demonstrated by deployed systems, coarse signals such as interface-level language selection, document provenance, discourse continuity, or deliberately abstracted metadata can already produce substantial reductions in uncertainty. Making these signals explicit would not simplify the task; rather, it would clarify which aspects of LID difficulty arise from linguistic ambiguity and which arise from prior misspecification.

Under either interpretation, the prevailing practice of evaluating LID systems solely on context-free text classification conflates engineering convenience with scientific validity. The result is not merely suboptimal performance for long-tail languages, but an evaluation procedure that is unable to distinguish between failures of modeling and failures of task definition.

6.3 Transparency and User Interaction

As observed in Section 5, institutions have various commitments that can make it difficult or unnecessary to support every language uniformly. For

web platforms in particular, successful LID without corresponding downstream support can lead to complicated operational and organizational situations. However, in many spaces where tail languages are actually used, it can appear that LID is inherently difficult, when in fact performance could be meaningfully improved through local context injection.

It is worth reiterating that the issue is not simply reduced accuracy for these languages. Rather, they are systematically eliminated from consideration altogether, often through implicit prior assumptions. This occurs despite the fact that mechanisms such as manual override or constrained language selection (potentially combined with verification) are already accepted for widely supported languages. As such, even when such compensatory mechanisms are not implemented, making the exclusion of a language explicit is important. Setting $P(\ell) = 0$ *everywhere* is a strong modeling decision, and leaving it implicit can lead to confusion when tested on languages where identification is often trivial due to script, such as Cherokee (Eggleston et al., 2025).

7 Conclusion

Despite the perception that language identification is an easy task, the majority of the world’s languages are not still not consistently identified in writing. Correctly identifying such languages is a key step for reducing digital disparities across speech communities. One of the root causes of this gap is the difficulty of defining the local probability of a language. To move forward, we must pay special attention to the social contexts these languages are used in, and reconsider institutional practices that center novelty and modeling convenience.

Limitations

We recognize that empirical demonstrations may have helped to make some of our arguments more concrete. However, due to the very nature of SOTA discriminative classifiers, directly exploring the effects of dynamic priors would have taken us beyond the scope of this paper.

Ethical Considerations

Whether or not one agrees with our position, it should be clear that there is a certain tension between accurate coverage of highly localized language on one hand and privacy considerations on the other. In this paper, we have argued that using

metadata like especially community and discourse structure is likely necessary to overcome technical obstacles to identifying languages that are not covered by existing large-scale systems. This does not imply that LID *must* be implemented even in communities that express explicit concerns about privacy and/or surveillance. However, in cases where such considerations are a relevant factor, they should be stated alongside other engineering constraints.

Acknowledgments

References

- Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Inciarte. 2022. [AfroLID: A Neural Language Identification Tool for African Languages](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1958–1981, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Noëmi Aepli, Çağrı Çöltekin, Rob Van Der Goot, Tommi Jauhiainen, Mourhaf Kazzaz, Nikola Ljubešić, Kai North, Barbara Plank, Yves Scherrer, and Marcos Zampieri. 2023. [Findings of the VarDial Evaluation Campaign 2023](#). In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 251–261, Dubrovnik, Croatia. Association for Computational Linguistics.
- Milind Agarwal, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2023. [LIMIT: Language Identification, Misidentification, and Translation using Hierarchical Models in 350+ Languages](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14496–14519, Singapore. Association for Computational Linguistics.
- Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2022. Hierarchical Aggregation of Dialectal Data for Arabic Dialect Identification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4586–4596, Marseille, France. European Language Resources Association.
- Gabriel Bernier-Colborne, Cyril Goutte, and Serge Leger. 2023. [Dialect and Variant Identification as a Multi-Label Classification Task: A Proposal Based on Near-Duplicate Analysis](#). In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 142–151, Dubrovnik, Croatia. Association for Computational Linguistics.
- Steven Bird. 2022. [Local Languages, Third Spaces, and other High-Resource Scenarios: 60th Annual Meeting of the Association for Computational Linguistics, ACL 2022](#). *ACL 2022 - 60th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, pages 7817–7829. Publisher Copyright:
© 2022 Association for Computational Linguistics.
- Ralf D. Brown. 2013. [Selecting and Weighting N-Grams to Identify 1100 Languages](#). In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Ivan Habernal, and Václav Matoušek, editors, *Text, Speech, and Dialogue*, volume 8082, pages 475–483. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. [An Open Dataset and Model for Language Identification](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 865–879, Toronto, Canada. Association for Computational Linguistics.
- Laurie Burchell, Alexandra Birch, Robert Thompson, and Kenneth Heafield. 2024. [Code-Switched Language Identification is Harder Than You Think](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 646–658, St. Julian’s, Malta. Association for Computational Linguistics.
- Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. 2020. [Language ID in the Wild: Unexpected Challenges on the Path to a Thousand-Language Web Text Corpus](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6588–6608, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Rasul Dent, Pedro Ortiz Suarez, Thibault Clérice, and Benoît Sagot. 2025. [Identifying Rare Languages in Common Crawl Data is a Needles-in-a-Haystack Problem](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1460–1473, Suzhou, China. Association for Computational Linguistics.
- Moussa Doumbouya, Lisa Einstein, and Chris Piech. 2021. Using radio archives for low-resource speech recognition: Towards an intelligent virtual assistant for illiterate users. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14757–14765.
- E. Dubois. 1863. *Relation historique de la campagne du Sénégal (1861) : extrait du journal d’un capitaine de tirailleurs algériens*. P. Arnavon, Blidah.
- Jonathan Dunn and Lane Edwards-Brown. 2024. Geographically-Informed Language Identification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7672–7682, Torino, Italia. ELRA and ICCL.
- Liam Enzo Eggleston, Michael P Cacioli, Jatin Sarabu, Ivory Yang, and Kevin Zhu. 2025. Bridging the

- Gap: Leveraging Cherokee to Improve Language Identification for Endangered Iroquoian Languages. In *Proceedings of the 1st Workshop on Advancing NLP for Low-Resource Languages*, Varna, Bulgaria.
- Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning Word Vectors for 157 Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- William N. Havard, Renauld Govain, Benjamin Lecouteux, and Emmanuel Schang. 2025. Self-Supervised Models of Speech Processing for Haitian Creole. *BABEL*, 1091(547):544.
- Tommi Jauhiainen, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2024. *Automatic Language Identification in Texts*. Synthesis Lectures on Human Language Technologies. Springer International Publishing.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.
- Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. *GlottLID: Language Identification for Low-Resource Languages*. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218, Singapore. Association for Computational Linguistics.
- Amr Keleg and Walid Magdy. 2023. Arabic Dialect Identification under Scrutiny: Limitations of Single-label Classification. In *Proceedings of ArabicNLP 2023*, pages 385–398, Singapore (Hybrid). Association for Computational Linguistics.
- Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, and 33 others. 2022. *Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets*. *Transactions of the Association for Computational Linguistics*, 10:50–72.
- Brook Danielle Lillehaugen. 2019. Tweeting in Zapotec: Social Media as a Tool for Language Activists. In Jennifer Carolina Gómez Menjívar and Gloria E. Chacón, editors, *Indigenous Interfaces: Spaces, Technology, and Social Networks in Mexico and Central America*, pages 202–226. University of Arizona Press, Tuscon.
- Marco Lui and Timothy Baldwin. 2012. *Langid.py: An Off-the-shelf Language Identification Tool*. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.
- Yash Madhani, Mitesh M. Khapra, and Anoop Kunchukuttan. 2023. *Bhasa-Abhijnaanam: Native-script and romanized Language Identification for 22 Indic languages*. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 816–826, Toronto, Canada. Association for Computational Linguistics.
- Oliver Mayeux. 2024. *Language revitalization on social media: Ten years in the Louisiana Creole Virtual Classroom*. *Language Documentation and Description*, 23(1).
- Paul McNamee. 2005. Language identification: A solved problem suitable for undergraduate instruction. *Journal of computing sciences in colleges*, 20(3):94–101.
- Kavi Narayana Murthy and G. Bharadwaja Kumar. 2006. *Language identification from small text samples*. *Journal of Quantitative Linguistics*, 13(1):57–80.
- NLLB_Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. *No Language Left Behind: Scaling Human-Centered Machine Translation*. *Preprint*, arXiv:2207.04672.
- Joanna Nolan. 2023. *Lingua Franca and Français Tirailleuse: From Sea Jargon to Military Order?* Springer International Publishing, Cham.
- Omnilingual ASR team, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, Kevin Chan, Chierh Cheng, Joe Chuang, Caley Droof, Mark Duppenthaler, Paul-Ambroise Duquenne, Alexander Erben, Cynthia Gao, Gabriel Mejia Gonzalez, and 14 others. 2025. *Omnilingual ASR: Open-Source Multilingual Speech Recognition for 1600+ Languages*. *Preprint*, arXiv:2511.09690.
- Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. *FineWeb2: One Pipeline to Scale Them All – Adapting Pre-Training Data Processing to Every Language*.
- Morton David Rau. 1974. *Language Identification by Statistical Analysis*. Ph.D. thesis, Monterey, California. Naval Postgraduate School.

- Sandy Ritchie, Daan van Esch, Uche Okonkwo, Shikhar Vashishth, and Emily Drummond. 2024. *LinguaMeta: Unified Metadata for Thousands of Languages*. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10530–10538, Torino, Italia. ELRA and ICCL.
- Anna Rogers, Marzena Karpinska, Jordan Boyd-Graber, and Naoaki Okazaki. 2023. Program Chairs’ Report on Peer Review at ACL 2023. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages xl–lxxv, Toronto, Canada. Association for Computational Linguistics.
- Alex Salcianu, Andy Golding, Anton Bakalov, Chris Alberti, Daniel Andor, David Weiss, Emily Pitler, Greg Coppola, Jason Riesa, Kuzman Ganchev, Michael Ringgaard, Nan Hua, Ryan McDonald, Slav Petrov, Stefan Istrate, and Terry Koo. 2023. Compact Language Detector v3 (CLD3). Google.
- Janvijay Singh, Mukund Rungta, Diyi Yang, and Saif Mohammad. 2023. *Forgotten Knowledge: Examining the Citational Amnesia in NLP*. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6192–6208, Toronto, Canada. Association for Computational Linguistics.
- Pedro Ortiz Suarez, Laurie Burchell, Catherine Arnett, Rafael Mosquera-Gómez, Sara Hincapie-Monsalve, Thom Vaughan, Damian Stewart, Malte Ostendorff, Idris Abdulmumin, Vukosi Marivate, Shamsuddeen Hassan Muhammad, Atnafu Lambebo Tonja, Hend Al-Khalifa, Nadia Ghezaiel Hamouda, Verrah Otiende, Tack Hwa Wong, Jakhongir Saydaliev, Melika Nobakhtian, Muhammad Ravi Shulthan Habibi, and 78 others. 2026. *CommonLID: Re-evaluating State-of-the-Art Language Identification Performance on Web Data*. *Preprint*, arXiv:2601.18026.
- Rob van der Goot. 2025. *Identifying Open Challenges in Language Identification*. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18207–18227, Vienna, Austria. Association for Computational Linguistics.
- Melanie Wagner. 2013. Luxembourgish on Facebook: Language Ideologies and Writing Strategies. *Social Media and Minority Languages: Convergence and the Creative Industries*, 152:87.
- Keith Whinnom. 1956. *Spanish Language Contact in the Phillipines*. Hong Kong University Press, London.
- Ivory Yang, Weicheng Ma, Chunhui Zhang, and Soroush Vosoughi. 2025a. *Is It Navajo? Accurate Language Detection for Endangered Athabaskan Languages*. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 277–284, Albuquerque, New Mexico. Association for Computational Linguistics.
- Ivory Yang, Chunhui Zhang, Yuxin Wang, Zhongyu Ouyang, and Soroush Vosoughi. 2025b. *Visibility as Survival: Generalizing NLP for Native Alaskan Language Identification*. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6965–6979, Vienna, Austria. Association for Computational Linguistics.

Bridging Dialectal Variation: A Phonetic Transcription Tool for Tamil

Ahrane Mahaganapathy¹, Sumirtha Karunakaran¹,
Kavitha Navakulan², Kengatharaiyer Sarveswaran¹

¹Department of Computer Science, University of Jaffna

²Department of Linguistics and English, University of Jaffna

Correspondence: ahrane@univ.jfn.ac.lk

Abstract

Phonetic transcription is vital for speech processing and linguistic documentation, particularly in languages like Tamil with complex phonology and dialectal variation. Challenges such as consonant gemination, retroflexion, vowel length, and one-to-many grapheme-phoneme mappings are compounded by limited data on Sri Lankan Tamil dialects. We present a dialect-aware, rule-based transcription tool for Tamil that supports Indian and Jaffna Tamil, with extensions underway for other dialects. Using a two-stage pipeline: Tamil script to Latin, then to IPA with context-sensitive rules, the tool handles dialect shifts. A real-time interface enables dialect selection. Evaluated on a 7,830-word corpus, it achieves 94.54% accuracy for Jaffna Tamil and is higher than other tools like eSpeak NG, advancing linguistic preservation and accessible speech technology for Tamil communities.

1 Introduction

Tamil is a classical South Dravidian language (Steever, 2019) spoken by nearly 80 million people across India, Sri Lanka, Malaysia, Singapore, and all over the world. Tamil has a proven literary history that spans more than two millennia, during which it passed through different phases of evolution: Old Tamil, Middle Tamil, and Modern Tamil, resulting in a dichotomy between literary Tamil, adhered to codified grammar (Lehmann, 2019) and spoken Tamil, which can vary significantly between regions and communities (Schiffman, 1998). This dialectal diversity, especially in low-resource languages such as Sri Lankan Tamil, presents significant challenges for linguistic documentation and computational modeling.

In the Tamil language, a single grapheme can correspond to multiple phonemic realizations, depending on its phonological context, morphological environment, or dialectal variation. This phe-

nomenon is known as one-to-many grapheme-to-phoneme (G2P) mapping.

Phonetic transcription, which provides a standardized representation of speech sounds, is fundamental in both linguistics and speech technology. The International Phonetic Alphabet (IPA) offers a cross-linguistic framework for transcribing phonemes (Association, 2015). However, Tamil phonology, marked by vowel length distinctions, consonant gemination, retroflexion, and context-sensitive plosive variation, demands transcription models that are context-sensitive and dialect-sensitive. These complexities are amplified in dialects like Sri Lankan Jaffna Tamil, which preserve conservative features no longer prominent in other Tamil dialects or varieties (Suseendrarajah, 1993).

There were extensive research exists on dialect-specific transcription in other languages such as BanglaIPA (Hasan et al., 2026) models six regional Bengali dialects using state-alignment algorithms to handle phonetic variability and out-of-vocabulary characters. TransDic (Garrido et al., 2018) supports multiple dialects of Spanish and Catalan by incorporating rules for primary and secondary phonetic phenomena, such as seseo and yeísmo. Similarly, Arabic (Hawasly et al., 2025), and American English (Károly, 2019) have been modeled with dialect-sensitive G2P systems.

In contrast, existing tools, such as eSpeak NG (Bernard and Titeux, 2021) *Anunaadam* (Virtu-alvinodh, 2014), offer IPA transcription for Tamil but lack support for dialectal nuances, a gap that disproportionately affects speakers of marginalized dialects. To address this gap, we present a rule-based phonetic transcription tool for Tamil, with particular emphasis on Sri Lankan Jaffna Tamil—a dialect that remains largely excluded from existing computational resources, despite being widely spoken not only in Sri Lanka but also by diaspora communities worldwide over the past

four decades.

Our key contributions include:

- **Dialect-Specific Modeling:** The first Tamil phonetic transcription system designed by considering dialectal variations, which captures key phonological features such as retroflexion, intervocalic alternation, and other allophones in vowels and consonants. The Language Technology Group (LTG) website ¹ contains a dedicated “Projects” section where the “Text to IPA transcription” tool and the relevant code can be found.
- **Linguistic Preservation:** By formalizing and digitizing features of a dialect, the tool supports phonetic documentation and speech technology development in low-resourced Tamil communities.
- **Scalability and Generalization:** The modular rule-based architecture of the tool enables adaptation to other low-resourced dialects of Tamil and similar phonetically rich languages.

Currently, most of the existing speech processing tools utilize ESpeak NG as the phonetic transcription tool, which does not support Tamil dialectal variations. By modeling Tamil’s phonological complexity and linguistic diversity, the proposed tool also contributes to inclusive NLP. It also supports applications in Text-to-Speech (TTS) and Automatic Speech Recognition (ASR).

2 Phonetics and Phonological Features of Tamil

Tamil consists of a rich phonetic system characterized by a diverse set of vowels and consonants that contribute to its unique sound structure. One of the important features of the language is its clear distinction between short and long vowels, as well as the phonemic role of consonant gemination (Kuno, 1958). The vowel inventory includes five basic vowels (/i/, /e/, /a/, /o/, /u/), each appearing in both short and long forms, resulting in ten contrastive vowel phonemes. This vowel length distinction plays a significant role in lexical differentiation (Kuno, 1958). Beyond these core phonemes, dialectal variation introduces additional phonetic complexity. For instance, Jaffna Tamil includes a

¹<https://www.ltg.jfn.ac.lk/>

broader set of vowel allophones, such as [ɛ], [i̠], and [ə], which are not typically found in standard varieties (Suseendirajah, 1993), as demonstrated in Table 1.

Consonants of the language include stops, nasals, fricatives, and liquids, with a notable distinction between voiced and voiceless obstruents. The language’s phonological rules are highly context-sensitive (Krishnamurthy, 1977), and consonant gemination further highlights the importance of phonological context (Balasubramanian, 1984; Heselwood, 2013). Krishnamurthy (1977) emphasizes the efficiency of Tamil’s orthographic system, wherein a single grapheme can represent multiple phonemic values depending on contextual cues. This characteristic reflects Zipf’s principle of least effort, which suggests that languages tend to minimize articulatory and cognitive load by reducing the number of distinct symbols while preserving expressive capacity (Cherry, 1971).

3 Tamil Phonetic Transcription

Tamil phonetics and phonology form a well-organized system that is essential for analyzing the linguistic structure. The phonological framework has several key features, including context-dependent phonetic variations, regional dialectal patterns, and the incorporation of lexical borrowings, all of which contribute to Tamil’s distinctive phonetic characteristics (Annamalai, 1979; Keane, 2004; Suseendirajah, 1993). This section provides a comprehensive examination of these aspects.

3.1 Context-Sensitive Pronunciations

Tamil phonology is characterized by significant context-sensitive variations, where the pronunciation of consonants changes based on their position within a word and the surrounding phonetic environment. A distinctive characteristic of Tamil phonology is the distribution of obstruents, particularly stops and fricatives. Voiced obstruents generally appear in word-internal positions, most often following nasal consonants. For instance, sequences such as a nasal followed by a voiced plosive, e.g., [nɖ] or [nj] are frequently observed within words. A few examples of context-sensitive pronunciation in Tamil, as handled by the proposed system, are presented in Table 2.

These variations present challenges in phonetic transcription tools, as they necessitate sophisti-

Phoneme	Allophone	Example
/a/	[ɛ]	அல்லி <i>alli</i> ‘water lily’ [ɛ l l i]
/a:/	[ɛ:]	நாய் <i>naai</i> ‘dog’ [nɛ:j]
/i/	[i]	நிறை <i>nirai</i> ‘weight’ [n i r a i]
/i:/	[i:]	நீளம் <i>neelam</i> ‘length’ [n i : l a m]
/u/	[u]	உப்பு <i>uppu</i> ‘salt’ [u p p u]
/e/	[ə]	எண் <i>en</i> ‘number’ [ə ŋ]
/e:/	[ɛ:]	ஏணி <i>e:ni</i> ‘ladder’ [ɛ:ŋi]

Table 1: Allophonic vowel realizations in Jaffna Tamil with example IPA transcriptions.

Grapheme	Allophone	Example
க்	[k]	கல் <i>kal</i> ‘stone’ [k ɛ l]
க்	[g]	கங்கை <i>gangai</i> ‘river Ganga’ [k a ŋ g a i]
க்	[x]	பகல் <i>pagal</i> ‘day time’ [p a x ɛ l]
ச்	[tʃ]	பச்சை <i>pachai</i> ‘green’ [p ɛ tʃ tʃ a i]
ச்	[dʒ]	மஞ்சள் <i>manjal</i> ‘yellow’ [m a n dʒ a l]
ச்	[s]	சட்டி <i>satti</i> ‘pot’ [s a tʃi]
த்	[t]	தம்பி <i>thambi</i> ‘younger brother’ [t a m b i]
த்	[d]	சிந்தனை <i>sindhanai</i> ‘thought’ [s i ŋ d ɛ n a i]
த்	[ð]	உதவி <i>uthavi</i> ‘help’ [u ð a v i]
ப்	[p]	பல் <i>pal</i> ‘tooth’ [p ɛ l]
ப்	[b]	பம்பரம் <i>pambaram</i> ‘spinning top’ [p a m b a r a m]
ப்	[β]	அபாயம் <i>abaayam</i> ‘danger’ [a β a: j a m]

Table 2: Some of the examples of one-to-many grapheme-to-phoneme mappings in Tamil across phonological contexts

cated modeling that can dynamically account for these positional and contextual changes. For instance, a single orthographic representation of a consonant might correspond to different phonetic realizations depending on its context, making accurate transcription complex without a context-aware system.

3.2 Dialectal Variations

Tamil exhibits considerable regional dialect variation across India, Sri Lanka, and other geographical regions. For instance, Jaffna Tamil is conservative and has unique vowel allophones that are generally absent in Indian Tamil dialects (Suseendirarajah, 1993). In contrast, urban varieties such as Chennai Tamil demonstrate significant phonetic and lexical influence from languages such as English, Hindi, Urdu, and Telugu (Smirnitskaya, 2018). These forms of variation underscore the need for transcription systems that are sensitive to dialectal variation across Tamil-speaking communities.

3.3 Loanword Influence

The integration of loanwords from Indo-Aryan languages such as words contain *Grantha* phonemes from Sanskrit and English has introduced new phonological elements into Tamil, altering traditional sound patterns. Even though, voiced plosives such as /b/ and /d/ are rare at the beginnings

of native Tamil words, they frequently occur in borrowed terms from languages like Sanskrit, Hindi, and English (Annamalai, 1979).

Therefore, the transcription systems must accommodate these non-native phonemes to accurately represent the evolving phonetic inventory. For instance, voiced fricatives such as /v/ and /z/, which are uncommon in the indigenous Tamil lexicon, are often present in borrowed vocabulary.

3.4 Notational Variances in Tamil Phonology

The notational system used to represent Tamil phonemes varies. Especially when using the International Phonetic Alphabet (IPA) to depict these sounds, ISO 15919 symbols are also used. The notational differences highlight the complexities of standardizing Tamil phonetic transcription, particularly when denoting dialectal variations. Some of the examples are listed in Table 3.

4 Methodology

According to Figure 1, the proposed Tamil Phonetic Transcription Tool adopts a rule-based architecture designed to preserve linguistic information while enabling dialect-sensitive phonetic realization. LLM based methods were not adopted because of the lack of annotated dialect specific corpora and their bias towards Indian Tamil varieties. The pipeline consists of (i) Tamil-to-

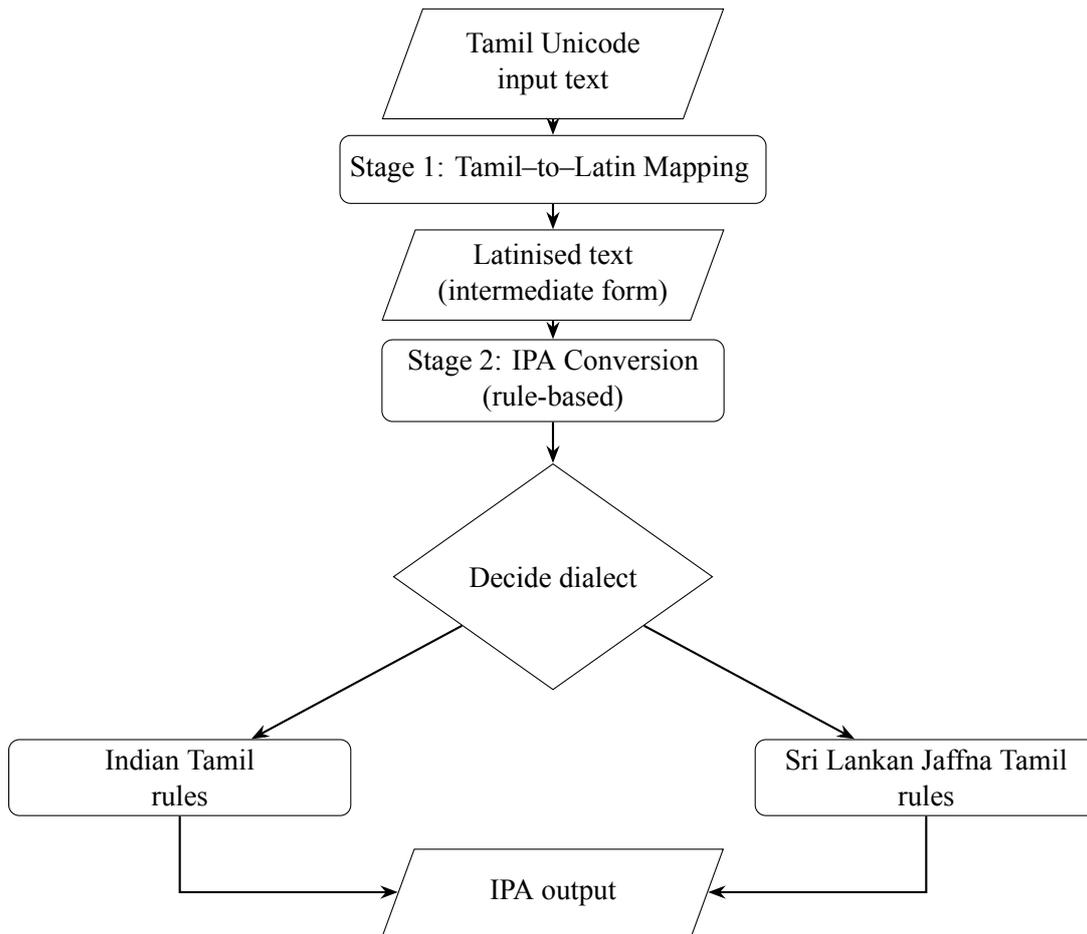


Figure 1: Pipeline for dialect-aware Tamil-to-IPA transcription.

IPA	ISO 15919
d	t
ɖ	ʈ
dz	c
j	y
l	ɭ
n	ɳ
ɲ	ɳ̠
ɳ	ɳ̠
ʈ	ʈ
tʂ	c
a:	ā
e:	ē
i:	ī
o:	ō
u:	ū

Table 3: Tamil Phonemes: IPA and ISO 15919 Equivalents

Latin grapheme mapping, followed by (ii) context-aware IPA conversion. It differs from eSpeakNG and *Anunaadam* as they do not cover dialectal vari-

ations in Tamil by adding dialect specific rule-set for Sri Lankan Jaffna Tamil.

4.1 Stage 1: Tamil-to-Latin Mapping (Information-Preserving Layer)

In the first stage, the `tam2lat()` module converts the Tamil script into an intermediate Latin-based representation. This mapping is strictly orthography-preserving: no segments are deleted, merged, or reinterpreted at this stage. The resulting representation remains structurally aligned with the original Tamil orthography, and all graphemic distinctions are explicitly retained. This conversion serves as a preprocessing step that facilitates the application of regular expressions, which are otherwise difficult to apply directly to Tamil Unicode characters. In addition, this intermediary representation enables easier extension of the system to other languages in the future.

4.2 Stage 2: IPA Conversion and Contextual Rule Application

In the second stage, the `ipa()` module maps the Latin intermediary to phonetic output through a sequence of context-sensitive transformations. These include intervocalic alternation of stops, nasal-stop voicing sequences, gemination stabilisation, and vowel length realisation. The transformations are applied only at this stage, ensuring that the phonemic contrasts encoded earlier are available for accurate contextual interpretation.

This stage introduces a dialect-selection mechanism that allows users to generate IPA transcriptions for different Tamil dialects. Currently, the system supports Indian Tamil (with *Anunaadam*-compatible phonetic behavior) and Sri Lankan Jaffna Tamil, with additional dialects being integrated in future extensions.

4.3 Dialect-Specific Rule Layer

The Jaffna Tamil mode extends the baseline rules by introducing a dedicated dialect layer implemented on top of the shared core rules. This layer contains 11 Sri Lankan Jaffna Tamil specific phonological and allophonic rules, including intervocalic situations and additional vowel allophones documented in dialectal descriptions (Suseendirajah, 1993). These rules are applied only when the Jaffna dialect is selected; the Indian Tamil option retains the *Anunaadam*-style realization. Other dialects can be extended in this layer by applying their dialect specific rules when their options are selected.

The Sri Lankan Jaffna Tamil specific rules introduced in the proposed system are listed in Table 4.

5 Results and Discussion

As the Indian Tamil rules are based on *Anunaadam* tool, the system was evaluated for the newly added Sri Lankan Jaffna Tamil dialect using a phonetic lexicon consisting of 7,830 words, including intervocalic situations, allophonic variations, *Grantha* phonemes, and other context-sensitive phonological contexts and the phonetic transcription accuracy was manually verified by a Sri Lankan native linguist. Out of them, 7403 were stated to be precise by the linguist, resulting in the word-level accuracy of 94.54%. A few cases in free variation in /t:r/ and /tt/, which are not captured in the proposed system, led to the reduction in ac-

curacy. The proposed system only results /t:r/ as output, does not cater to other variations in such instances. For example, in Sri Lankan Jaffna Tamil, the word முற்றிலும் *murṛilum* ‘all inclusive’ exhibits phonetic free variation and is pronounced either as [mut:rɪlum] or [muttilum]. This variation, which occurs in fluent speech without altering semantic content, reflects a broader tendency to simplify consonant clusters, particularly the /tr/ sequence. These free variations are not captured on the proposed tool, resulting the reduction of accuracy.

A detailed analysis revealed notable phonetic differences between Sri Lankan Jaffna Tamil and Indian Tamil based on the provided dataset. One key distinction lies in the use of fricative versus plosive sounds: Indian Tamil employs the voiced velar fricative [ɣ], as in [aɣaŋga:ra], while Sri Lankan Jaffna Tamil substitutes it with the voiceless velar plosive [x], yielding [axaŋga:ra]. This substitution is a characteristic feature of Jaffna Tamil pronunciation. Vowel variation is another significant feature. Indian Tamil typically uses the low central vowel [a], as in [aɣandaiya:na], whereas Sri Lankan Jaffna Tamil often shifts to the mid-front vowel [ɛ], producing forms such as ax-endaiya:na.

For comparative purposes, eSpeak NG was also evaluated using the same corpus for the Sri Lankan Jaffna dialect. Its phonetic outputs were analyzed, enabling a direct comparison of transcription accuracy and dialectal adequacy between the proposed system and an established text-to-speech framework. Out of 7,830 words, 5,212 were judged to be correctly transcribed by the linguist. Hence, the word-level accuracy was 66.56%.

When eSpeak NG output is analysed, the grapheme ஃ is realized only as the context-sensitive variants [k] and [g]. In intervocalic positions, it is realized as [g], whereas in Sri Lankan Jaffna Tamil, the fricative [x] occurs in intervocalic positions (Suseendirajah, 1993).

In some instances, word-initial positions are also realized as g, for example, the word கேலி *kēli* ‘teasing’ is also transcribed by eSpeak NG as [ge:li] and குதி *kuti* ‘jump’ is transcribed as [gudi]. However, in Jaffna Tamil, [g], voiced dorso-velar plosive, occurs medially, in the sequence, and [k] voiceless dorso-velar plosive, occurs in word-initial position and in medial position following [ɾ] (Suseendirajah, 1993).

eSpeak NG also transcribes ஸ as [b] in the

Target	Environment
/k/ to [x]	intervocalic, பகல் <i>pagal</i> ‘day time’ [p a x ε l]
/p/ to [β]	intervocalic, அபாயம் <i>abaayam</i> ‘danger’ [a β a: j a m]
/t/ to [ð]	intervocalic, உதவி <i>uthavi</i> ‘help’ [u ð a v i]
/k/ to [g]	only after nasal, கங்கை <i>gangai</i> ‘river Ganga’ [k a ŋ g a i]
/a/ to [ε]	before alveolar consonants /r/, /l/, /n/, அல்லி <i>alli</i> ‘water lily’ [ε l l i]
/a:/ to [ε:]	before /j/, /r:/, நாப <i>naai</i> ‘dog’ [nε:j]
/i/ to [i]	before retroflex consonants and /r/, நிறை <i>nirai</i> ‘weight’ [n i r a i]
/i:/ to [i:]	before retroflex consonants and /r/, நீளம் <i>neelam</i> ‘length’ [n i : a m]
/u/ to [i]	if not in first syllable, உப்பு <i>uppu</i> ‘salt’ [u p p i]
/e/ to [ə]	before retroflex consonants, /p/, /k/, /m/, /v/, /ŋ/ and /r/, எண் <i>en</i> ‘number’ [ə ŋ]
/e:/ to [ə:]	before retroflex consonants, /p/, /k/, /m/, /v/, /ŋ/ and /r/, ஏணி <i>e:ni</i> ‘ladder’ [ə:ni]

Table 4: Jaffna-Tamil specific rules with examples

word-final positions. Eg. குருவைப் *kuruvaip* ‘teacher.ACC’, [guɻvəɪb] ends with [b]. In some instances, it results as [b] in the word’s first position. Eg. பத்திரமாக *pattiramaaka* ‘carefully’ is transcribed as [battira:ma:ga] through eSpeak NG. However, in Sri Lankan Jaffna Tamil, [b], voiced bilabial plosive occurs in the sequence of [mp], [np], and voicing of [b] is not so heavy as in Indian Tamil (Suseendirarajah, 1993).

நான் *naan* ‘I’ is transcribed as [na:n] which does not reflect a phonemic distinction between the orthographic forms ந and ன orthographies. In contrast, the proposed system distinguishes between the two (for example, [nān]). These distinctions are important for second-language learners to avoid confusion.

There are issues regarding vowel length from eSpeak NG, in some contexts words such as சுத்தியூட்ட *cahtiyuutta* [sʌtʃijʊʈʌ] ‘to energize’ resulted in [ɥ] instead of a long vowel in the transcription. Whereas, the proposed system works well in this context.

In addition to that, eSpeak NG does not cover the phonological rules specific to the Sri Lankan Jaffna dialect listed in Table 4, including allophonic vowel variations. Even though the [k] to [g] transformation occurs after nasals, it does not adequately capture the Jaffna Tamil context, as in Jaffna Tamil, [g] occurs only after nasals and not in word-initial position.

The *Anunaadam* tool itself was not evaluated for comparative purposes, as our system extends *Anunaadam* to account for Sri Lankan Jaffna Tamil dialectal phonological phenomena, making a direct comparison with the tool irrelevant

6 Challenges in Tamil Phonetic Transcription

Automating phonetic transcription for Tamil involves navigating a range of linguistic complexities rooted in its phonological structure. Tamil exhibits strong context sensitivity in its sound system, where the realization of consonants and vowels depends heavily on their syntactic and phonetic environments (Krishnamurthy, 1977). For example, ப is [p] in பல் *pal* ‘teeth’ but becomes [b] in பம்பரம் *pamparam* ‘spinning top’ as mentioned in Table 2. These phenomena are addressed through syntax-directed transcription rules employing a one-symbol look-ahead mechanism, thereby providing a foundation for context-aware phonetic modeling.

Loanwords from Indo-Aryan languages, European and other languages came into contact due to regional contact and the migration of Tamil diaspora all over the world, adding another layer of complexity, often introducing phonemes not native to Tamil. As a result, transcription remains an ongoing process that must continually adapt to changing phonotactic patterns. For example [f] in ஃபோன் *fon* from ‘phone’. Foundational frameworks such as Caldwell’s Law (Caldwell, 1931) remain relevant, but modern transcription systems must adapt to a broader spectrum of linguistic inputs.

Altogether, these challenges underscore the need for transcription tools that are dialect-sensitive, context-aware, and capable of handling both standard and non-standard phonetic phenomena. Addressing them is critical for developing inclusive NLP systems.

7 Conclusion

This study presents a dialect-sensitive phonetic transcription tool for Tamil, called *ThamizhiIPA-*

Trans, addressing the phonological complexity and regional variability inherent in the language. Unlike existing tools, which often overlook dialectal nuances of low-resource languages such as Tamil, the proposed system is designed to model features specific to individual Tamil dialects, with a particular focus on Sri Lankan Tamil. A user-friendly interface allows real-time input and dialect selection, making the system useful for researchers, educators, and speech technology developers alike.

The developed system was evaluated on a lexicon of 7,830 words, and the outputs were assessed by a trained linguist, yielding an overall accuracy of 94.54%, which is substantially higher than the widely used eSpeak NG, which achieved 66.56% for the Sri Lankan Jaffna dialect. *ThamizhiIPA-Trans* can be accessed through the Language Technology Group (LTG) website². Future directions include extending the system to cover additional Tamil dialects in Sri Lanka and abroad and integrating prosodic features.

Limitations

Even though the proposed system demonstrates strong performance for Sri Lankan Tamil phonetic transcription, certain limitations remain. Some instances of free variation, speaker-specific pronunciation, code-mixed forms, and style-dependent phonetic shifts are not captured by the current rule set — addressing these may require a data-driven approach. The present approach models phonetic realization primarily at the segmental level, while prosodic phenomena such as stress, rhythm, and intonation remain outside the scope of this work. These limitations point to promising directions for future extensions, but do not detract from the central contribution of this study, namely, demonstrating that a dialect-aware, rule-based framework can effectively model key phonological processes in Tamil phonetic transcription.

Declaration on Generative AI

We used AI-assisted tools (e.g., ChatGPT) solely for language refinement and carefully rechecked and edited by the authors; all scientific contributions were carried out by the authors.

²<https://www.ltg.jfn.ac.lk/>

Acknowledgements

This research study was carried out under the DigSAL project, supported by the German Academic Exchange Service (DAAD) and funded by the Federal Ministry for Economic Cooperation and Development (BMZ) through SDG Partnerships (2024-2027).

References

- E. Annamalai. 1979. *Linguistic Complexity of Tamil Phonology*. Madras University Press.
- International Phonetic Association. 2015. *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press.
- T. Balasubramanian. 1984. Tamil Phonetic Characteristics: A Comparative Study. *Phonology Journal*.
- Mathieu Bernard and Hadrien Titeux. 2021. [Phonemizer: Text to Phones Transcription for Multiple Languages in Python](#). *Journal of Open Source Software*, 6(68):3958.
- Robert Caldwell. 1931. *Comparative Grammar of the Dravidian or South Indian Family of Languages*. University Press, London.
- C. Cherry. 1971. *On Human Communication*. MIT Press.
- Juan María Garrido, Marta Codina, and Kimber Fodge. 2018. TransDic, a public domain tool for the generation of phonetic dictionaries in standard and dialectal Spanish and Catalan. In *IberSPEECH*, pages 291–295.
- Jakir Hasan, Shrestha Datta, Md Saiful Islam, Shubhashis Roy Dipta, and Ameya Debnath. 2026. [BanglaIPA: Towards Robust Text-to-IPA Transcription with Contextual Rewriting in Bengali](#). *Preprint*, arXiv:2601.01778.
- Majd Hawasly, Hamdy Mubarak, Ahmed Abdelali, and Ahmed Ali. 2025. DialG2P: Dialectal Grapheme-to-Phoneme. Arabic as a Case Study. In *Proceedings of The Third Arabic Natural Language Processing Conference*, pages 466–471.
- Barry Heselwood. 2013. *Phonetic Transcription in Theory and Practice*. Edinburgh University Press.
- László Károly. 2019. An integrated dialect analysis tool using phonetics and acoustics. *Lingua*, 221:37–48.
- Elinor Keane. 2004. Illustrations of the IPA: Tamil. *Journal of the International Phonetic Association*, 34(1):111–116.
- E.V. Krishnamurthy. 1977. Automatic Phonetic Transcription of Tamil in Roman Script. *Proceedings of the Indian Academy of Sciences*, 86(6):503–512.

- Susumu Kuno. 1958. *Phonemic Structure of Colloquial Tamil*. *Journal of Linguistic Studies*.
- Thomas Lehmann. 2019. Old tamil. In *The Dravidian languages*, pages 81–103. Routledge.
- Harold F Schiffman. 1998. Standardization or re-standardization: The case for “standard” spoken Tamil. *Language in Society*, 27(3):359–385.
- Anna Aleksandrovna Smirnitskaya. 2018. Diglossia and Tamil varieties in Chennai. *Acta Linguistica Petropolitana. Труды института лингвистических исследований*, 14(3):318–334.
- Sanford B Steever. 2019. Introduction to the Dravidian languages. In *The Dravidian languages*, pages 1–44. Routledge.
- S. Suseendirajah. 1993. *Jaffna Tamil (1) Phonology and Morphology*. University of Jaffna Publication.
- Virtualvinodh. 2014. Anunaadam. GitHub repository. <https://github.com/virtualvinodh/anunaadam>.

Regional Variation in the Performance of ASR Models on Croatian and Serbian

Tanja Samardžić
IDSIA USI-SUPSI,
Lugano, Switzerland
tanja.samardzic@supsi.ch

Peter Rupnik and Nikola Ljubešić
Department of Knowledge Technologies,
Jožef Stefan Institute, Ljubljana, Slovenia
{peter.rupnik|nikola.ljubestic}@ijs.si

Abstract

Regional variation was a limiting factor for automatic speech recognition (ASR) before large language models. With the new technology, speech processing becomes more general, which opens the question of how to use data in similar languages such as Croatian and Serbian. In this paper, we analyse model performance in the currently available train-test scenarios with the goal of better understanding the mutual interference of these two languages. Our findings suggest that better performing models are not very sensitive to the regional variation. Models trained from scratch in one of the languages can give good results on both of them, while fine-tuning large pre-trained multilingual models on smaller data sets does not give the expected results.

1 Introduction

For a long time, automatic speech recognition (ASR) was only possible in very limited domains targeting a particular speaker and a particular topic. This lack of generalisation was famously captured in the comedy sketch *Voice Activated Elevator*¹ by the BBC Scotland Burnistoun show first aired in 2009. The sketch shows two persons trying to activate a lift by pronouncing the number 11. Since the system does not react, they guess that the problem is their Scottish accent and try to mimic other accents (“American”, “English”) with no success. The sketch captures not only the limitations of the ASR technology at that time but also important societal implications of these limitations.

Regional variation in the ASR performance is certainly present in all languages, but it has not been extensively studied as this kind of research is particularly demanding in terms of resources (both data and computing) and study design (representative sampling), while it is hard to isolate factors that determine the performance.

¹<https://www.bbc.co.uk/programmes/p00hbfjw>



Figure 1: Geographical locations of the four *Mak na konac* data parts. **HR1**: Croatian, Zagreb, Northern variety spoken in the administrative centre. **HR2**: Croatian, Split, Southern variety spoken in a major city. **SR1**: Serbian, Belgrade, Northern variety spoken in the administrative centre. **SR2**: Serbian, Niš, Southern variety spoken in a major city.

New ASR technology that employs large models trained on large amounts of data (thousands of hours of transcribed audio) seems to have reached the level of generalisation that makes multi-purpose ASR viable. In particular, the multilingual model Whisper (Radford et al., 2023) has become a universal reference, able to process various languages, including Croatian and Serbian, without any additional training. Since this model is released as open-weights, it can also be fine-tuned and thus, in theory, more precisely adapted for a particular language and variety.

The possibility of fine-tuning an already well-performing model to reduce its error rate on a given variety is clearly tempting, but the success is not guaranteed. The main problem is a mismatch between the size of the model and the size of the data available for any single variety, this even if we leave aside the problem of computing resources needed to employ a large model.

The goal of our study is to start an exploration towards a better understanding of what happens in various train-test scenarios involving similar but distinct languages and their variants. We focus

on the case of Croatian and Serbian as two languages that both belong to the same macro language, BCMS (for Bosnian, Croatian, Montenegrin and Serbian). The crucial question for macrolanguages is whether the data in one variety can be used to train or improve ASR models in all varieties. Through several recent projects, Croatian and Serbian have been equipped with the necessary resources to start investigating this problem, making them an interesting case study potentially relevant to many other similar languages.

2 Related Work

Regional variation is, to some degree, addressed in the case of Arabic and Swiss German, which are both famous for their regional diversity.

While Modern Standard Arabic (MSA) has been the main focus of research in the case of Arabic (Dhouib et al., 2022), dialect variation is starting to attract more attention in recent studies (Mubarak et al., 2021; Alharbi et al., 2024; Al-Fetyani et al., 2021; Al Ali and Aldarmaki, 2024; Djanibekov et al., 2025). What emerges from these studies is considerable variation in the ASR performance across dialects. While dialect-specific fine-tuning can help, this applies only if there is enough data (over 800 hours of transcribed speech). Also the impact of data availability does not seem to be straightforward. For instance, the performances tend to be better on the Syrian variety than on the Saudi or Egyptian one although the latter two are far better represented in the available data and models (Djanibekov et al., 2025).

Interesting dialectal patterns can be seen in the studies on Swiss German too: ASR performances are the worst for the Wallis variant while they are the best for the Grisons variant. This same outcome was achieved by two independent studies that evaluated very different ASR systems trained on different data sets. The first evaluation (Nigmatulina et al., 2020) was performed in 2020 with a Swiss-German Kaldi (Povey et al., 2011) recipe trained on the ArhciMob corpus (Samardžić et al., 2016; Scherrer et al., 2019). The second evaluation was performed in 2022 (Schraner et al., 2022) with a multilingually pre-trained XLS-R model (Babu et al., 2021) fine-tuned on other Swiss German data, including SDS-200 (Plüss et al., 2022) and a Swiss local parliaments corpus (Plüss et al., 2021), and tested on the STT4SG-350 data set (Plüss et al., 2023). Both of these variants are geographically

Data part	Source	Language	Region	Size
HR1	Radio Student	Croatian	Zagreb	5h
HR2	TV Dalmacija	Croatian	Split	5h
SR1	Pešćanik	Serbian	Belgrade	5h
SR2	Južne vesti	Serbian	Niš	5h

Table 1: Summary of the data key features.

peripheral, but one of them seems to be easier for ASR models than the other.

Until recently, the performance of speech-to-text models on Croatian and Serbian was rarely reported. A Kaldi recipe for Serbian (Popović et al., 2015) showed a large performance drop on the out-of-domain test set. In the past few years, however, considerable progress has been made in data development (Ljubešić et al., 2024a,c,b; Rupnik and Ljubešić, 2022a), which enabled model training and testing initially on Croatian (Ljubešić et al., 2022) and more recently on Serbian (Sagić, 2023). These data sets contain over 3000 hours of automatically aligned transcribed audio for Croatian and almost 1000 hours for Serbian.

The project *Mak na konac* (Samardžić et al., 2024) resulted in a new test set consisting of 20 hours of speech sampled from four media sources and manually transcribed. This data set targets specifically regional variation which makes it especially suitable for our analysis.

3 Data

The *Mak na konac* data set (MNK) (Samardžić et al., 2024) is composed of four parts of equal size (5 hours of audio each), two representing Croatian (HR) and two Serbian (SR) varieties. As shown in Figure 1, each data part represents one region. In addition to the administrative centres (Zagreb and Belgrade), both language varieties are represented with a Southern variety (Split and Niš). All data sources listed in Table 1 are interview-type TV or radio programmes featuring one host and one guest speaker, with the exception of Radio Student hosted by two presenters.

This is a multi-reference data set, where each segment of speech is aligned with multiple text reference options created to cover known variability

Model	Pre-training	Training / Fine-tuning
WhisperV	Multilingual	None
WhisperS	WhisperV	Serbian, around 70h: Flores ASR (multilingual Eastern Europe without Croatian) + Serbian Common Voice + JuzneVesti-SR
WhisperSJV	WhisperV	Serbian, around 120h: Flores ASR (multilingual Eastern Europe without Croatian) + Serbian Common Voice + Serbian "Juzne vesti" Serbian Common Voice + Unknown Serbian
Transducer CTC	None None	Croatian, 1816 hours, ParlaSpeech.v1 Croatian, 1816 hours, ParlaSpeech.v1
W2V2XLSR	Multilingual only audio: BABEL, Multilingual LibriSpeech, CommonVoice (no Croatian), VoxPopuli (no Serbian)	Croatian, 300 hours, ParlaSpeech.v1
W2V2Slavic	Multilingual only audio: VoxPopuli (no Serbian)	Croatian, 300 hours, ParlaSpeech.v1

Table 2: Summary of key features of the models tested in the MNK project.

in speech transcription. In some options, elements of speech (repetition, fillers) are included in the reference transcript, while they are excluded in others. Further options are created by a different treatment of numbers and abbreviations. As an example, Table 4 shows how such a reference would look like in English. Taking the best score out of all options when evaluating models has the effect of neutralising irrelevant variation and enables comparing models trained on different data. This is especially important when comparing solutions based on off-the-shelf models because the user has no control over the selection of pre-training data.

The audio part of the MNK test data is available online on Hugging Face.² The multi-reference transcript is not publicly available, to prevent model contamination, but researchers can obtain an evaluation report on a submitted output of a model.

4 Models

We analyse the performance of seven freely available models:

- WhisperV: multilingual Whisper large-v3, which is pre-trained on 1 million hours of weakly labelled data and 4 million hours of pseudo-labelled data, produced with its predecessor, Whisper-large-v2. It is capable of automatically determining the language of the input speech as well as translating input speech into a variety of languages. In this setting, the model is applied on the *Mak na konac* data in a **zero-shot** fashion.
- WhisperS and WhisperSJV: two variants of

the WhisperV model fine-tuned on transcribed Serbian audio. The first variant is fine-tuned on Mozilla Common Voice 13 and Google Fleurs and the ASR training data set for Serbian JuzneVesti-SR v1.0 (Rupnik and Ljubešić, 2022b). The second variant has 50+ hours of unspecified Serbian data added to the training set. Both of these models can be considered fine-tuned towards **Serbian** since Common Voice did not contain Croatian at the time of fine-tuning.

- Transducer and CTC are parts of NVIDIA’s NeMo toolkit. The main difference between the two variants is that Transducer takes previously generated letter as input at the next step, while CTC does not (it combines the acoustic and the language model in a more traditional way). In both of these settings, the models that were trained from scratch on **Croatian** parliamentary data set ParlaSpeech-HR (Ljubešić et al., 2022).
- W2V2XLSR and W2V2Slavic are models based on the wav2vec2 architecture. This means that they are first trained on large quantities of unlabelled audio alone then continued training on labelled data. The former is pre-trained on multilingual partially labelled text and speech datasets (XLS-R), while for the latter, the VoxPopuli Slavic labelled data (Wang et al., 2021) are used. For our study, it is important to note that the VoxPopuli data set is part of both settings and that it contains only Croatian without Serbian. Both of these models are finally fine-tuned on a subset of the **Croatian** ParlaSpeech-HR data set.

²https://huggingface.co/datasets/classla/mak_na_konac/viewer/default/SR1?row=4

Segment ID		Whisper V	Whisper S	Whisper SJV	Transducer	CTC	W2V2 XLSR	W2V2 Slavic
170918_13_14	r1	[0.286,	[0.429,	[0.429,	[0.571,	[0.571,	[1.000,	[0.857,
	r2	0.143,	0.286,	0.286,	0.429,	0.429,	0.857,	0.857,
	r3	0.125,	0.500,	0.250,	0.375,	0.375,	0.750,	0.714,
	r4	0.000]	0.375]	0.125]	0.250]	0.250]	0.625]	0.500]
170918_14_15	r1	[0.429,	[0.286,	[0.714,	[0.857,	[0.857,	[0.857,	[0.857,
	r2	0.400]	0.400]	0.300]	0.300]	0.300]	0.300]	0.300]
170918_15_16	r1	[0.111,	[0.000,	[0.556,	[0.667,	[0.556,	[0.556,	[0.556,
	r2	0.091]	0.182]	0.455]	0.364]	0.455]	0.364]	0.455]
170918_16_17	r1	[0.100,	[0.100,	[0.100,	[0.300,	[0.400,	[0.100,	[0.300,
	r2	0.100]	0.100]	0.100]	0.300]	0.400]	0.100]	0.300]
170918_17_18	r1	[0.222,	[0.111,	[0.222,	[0.556,	[0.667,	[0.444,	[0.444,
	r2	0.455,	0.364,	0.455,	0.455,	0.545,	0.636,	0.545,
	r3	0.111,	0.000,	0.111,	0.444,	0.556,	0.333,	0.333,
	r4	0.364,	0.273,	0.364,	0.364,	0.455,	0.545,	0.455,
	r5	0.100,	0.200,	0.100,	0.400,	0.500,	0.300,	0.300,
	r6	0.333,	0.417,	0.333,	0.333,	0.417,	0.500,	0.417,
	r7	0.000,	0.100,	0.000,	0.300,	0.400,	0.200,	0.200,
	r8	0.250]	0.333]	0.250]	0.250]	0.333]	0.417]	0.333]

Table 3: The first five entries in an evaluation report. Each entry contains multiple WER scores (one for each reference option r1–rn) for one audio segment. The best scores are **in bold**. The full evaluation report for all data parts contains around 11’000 of such entries.

r1	We’re gonna meet on the 20th of December.
r2	We’re gonna meet on the twentieth of December.
r3	We are going to meet on the 20th twentieth of December.
r4	We are going to meet on the twentieth of December.

Table 4: An English illustration of a multi-reference data set. Some versions (r1, r2) are more literal and some (r3, r4) more standard. In some versions (r1, r3) the number is written as a digit, while in others (r2, r4) they are spelled out.

These models can be divided into three groups depending on whether they are pre-trained and used in zero-shot testing (WhisperV), pre-trained and fine-tuned for a specific variant (WhisperS, WhisperSJV, W2V2Slavic and W2V2XLSR to some degree) or trained from scratch (Transducer and CTC). The latter option allows more control over the training data and fitting the models more closely to the target, but these models are smaller and might not capture all the nuances that can be accommodated by the pre-trained models that are bigger. On the other hand, the bigger models require much more data to be trained, which is why they cannot be trained for any single language (except English) specifically.

A few notes are due on fine-tuning. The only model that is not specifically fine-tuned for any Croatian or Serbian target data is WhisperV, al-

though it is obvious that some Croatian and Serbian data are included in its, highly multilingual, training set given mid-tier performance on these languages. Two models (WhisperS and WhisperSJV) are said to be fine-tuned towards Serbian, but some of the fine-tuning data might have already been included in the pre-trained model as well. The other models are trained on Croatian data (recordings of the Parliament sessions). Transducer and CTC are trained from scratch on Croatian data, while W2V2XLSR and W2V2Slavic are fine-tuned on a portion of the the same Croatian data. The latter two models are pre-trained on multilingual and Slavic data respectively. A summary of the models’ key features is given in Table 2

The models are initially tested in the *Mak na konac* project (Samardžić et al., 2024), which has also provided the multi-reference gold standard. The full evaluation reports created in the original evaluation are large tables that contain meta-data and CER (character error rate) scores in addition to the WER (word error rate) scores shown in Table 3. For the current study, we use only some parts of these reports, the WER scores. We extract the best WER score for each model for a given audio segment and calculate the average best score for each model and each data part. This allows us to observe regional variation in best WER scores per

model.

Our analysis is performed on the evaluation reports for the seven models listed in Table 2 separately for the four data parts described in Table 1. The first five lines of one of the analysed evaluation reports is shown in Table 3. For each audio segment (ID in the first column) and each tested model, we obtain an array of WER scores (here rounded up to three decimal places). For example, the model WhisperV obtains eight scores on the segment 170918_17_18 depending on the overlap between the model output and each reference option. The number of scores per audio segment and per model can go over 70 depending on the presence of numbers and abbreviations, but also on the length of the utterance.

5 Results and Discussion

Figure 2 shows the average best WER score for each model and each data part.

5.1 Zero-Shot Whisper Wins, for Now

It turns out that, somewhat surprisingly, the best performing model overall is WhisperV, followed by the two models trained on Croatian (Transducer and CTC), while the worst performance is obtained with the two W2V2 models fine-tuned on Croatian. It seems that the currently available data for both Croatian and Serbian do not reach the level that is needed to see the benefit of training and fine-tuning specific to the target variety, as it could be seen in Arabic. This might change very soon but, for now, zero-shot use of pre-trained Whisper gives the best results on both Croatian and Serbian.

5.2 Regional Patterns

The only settings that show regional variation are the two versions of Whisper fine-tuned on Serbian (WhisperS and WhisperSJV). In these cases, fine-tuning does not improve the results for Serbian compared to the zero-shot setting, but it does spoil the performance on Croatian. We also note that one part of the data used for fine-tuning comes from the same source as the SR2 data part. Although there is no data overlap, one would expect to see more benefits of fine-tuning on the same data source. In reality, the difference in favour of the zero-shot setting increases on SR2 compared to SR1, which remains puzzling. It looks like fine-tuning resulted in overfitting to the specific data set used for fine-tuning, so the model became more

sensitive to any variation.³ Nevertheless, the drop in the performance is bigger on Croatian.

5.3 Training From Scratch Better Than Fine-Tuning

The only case where zero-shot Whisper does not give the best performance are the two models trained from scratch on Croatian (Transducer and CTC) and tested on HR2. This said, these models show much more variation across data parts with a considerable performance drop on the other Croatian part (HR1), even bigger than on the two Serbian parts. Still, their performance remains superior to all fine-tuned settings on Croatian and better than the two W2V2 models on both languages. These results suggest that training from scratch is a better way for reaching good performance on a target variant at the expense of some generalisation. Fine-tuning large models on smaller data sets seems to result in strong overfitting preventing improvements even on similar data and reducing the transfer across variants.

5.4 Style Variation Rather Than Regional

Another outcome that is unexpected is overall better performance on Southern variants (HR2 and SR2) than on the ones spoken in the two administrative centres (HR1 and SR1). Given the biases in the population size and overall media presences in favour of the central variants, one would expect that they are better represented in the data available for training and fine-tuning models, which, in turn, should lead to better results. Still the pattern is the opposite. The difference is more pronounced in the case of Croatian (HR1 vs. HR2) than in the case of Serbian (SR1 vs. SR2). The data part that seems the hardest for all the models is HR1, while HR2 seems the easiest for all models except the two Whisper variants fine-tuned on Serbian (WhisperS and WhisperSJV). All models trained and fine-tuned on Croatian data perform worse on HR1 than on the two Serbian parts.

This naturally raises the question of what makes HR1 harder than HR2 for all the models regardless of the training/fine-tuning settings. The fact that the source of the data is a radio programme with two hosts and a rather informal conversational setting might be a part of the answer as this might introduce more dynamics in the interactions and more

³Note also that the performance reported on the Hugging Face repository (Sagić, 2023) is much better than what we observe, which might also be interpreted as a sign of overfitting.

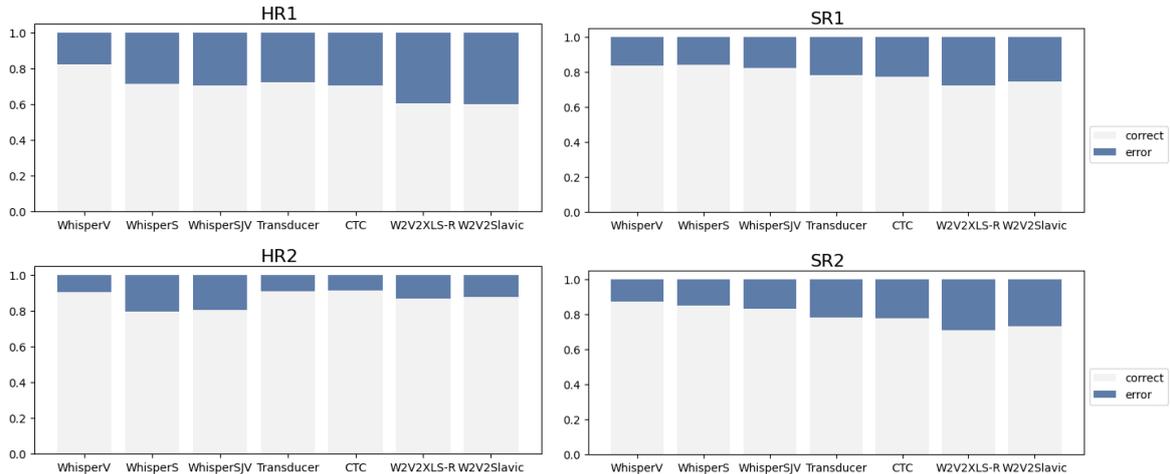


Figure 2: Average best WER scores on the four parts of the MNK data set. The dark bars on the top (error) show the WER, the light bars in the bottom show the complement (1-WER), which can be seen as word accuracy.

small speaker overlaps. Modern ASR models are rather robust to such overlaps, but they still might be impacted. Other explanations might be obtained from the metadata (e.g. the speakers’ demographics) and data analysis (e.g. lexical diversity), which remain outside of the scope of our current analysis.

One feature that can be extracted from our data is the level of variability of the WER scores. If we count the number of distinct WER scores assigned to the models, this can show how variable reference transcripts were in each data part. For example, the total number of scores assigned to WhisperV in Table 3 is 18 and the number of distinct scores is 14. Table 5 shows the counts of distinct scores for all data parts. Indeed, model performance seems correlated with the level of variability in the data: the number is the highest in HR1, followed by SR1, then SR2 and HR2. We see the same ranking in the overall model performance. This means that at least some differences in the model performance can be explained by the presence of elements of speech, numbers and abbreviations in the data sets, which can be associated with styles of speech rather than geographical regions.

6 Conclusion and Future Work

In this paper, we have studied different scenarios of training and fine-tuning speech-to-text models on Croatian and Serbian with the goal of understanding whether the two variants need to be separated to achieve better overall results.

Our findings suggest that regional variation does play a major role in model’s performance in the currently available data and model settings. Com-

Model	HR1	HR2	SR1	SR2
WhisperV	380	250	323	250
WhisperS	406	257	315	249
WhisperSJV	427	281	331	268
Transducer	397	266	331	280
CTC	412	272	362	282
W2V2XLS-R	429	273	349	303
W2V2Slavic	436	264	336	302
Average (rounded)	412	266	335	276

Table 5: The number of distinct WER scores per model and data part.

posing a large data set that contains both Croatian and Serbian seems to be the best way to achieve good scores on both languages. With additional Serbian data added to the Croatian set, training from scratch might give better scores than zero-shot use and fine-tuning multilingual pre-trained models.

The performance of fine-tuned models in the settings analysed in our study was clearly worse than the other two options (training from scratch and zero-shot testing). While this is a consistent pattern in our findings, the potential of fine-tuning a large pre-trained model is still to be clarified in future work. For this, we will need to evaluate models trained from scratch and fine-tuned on the same amount of data.

Limitations

To study the impact of regional variation on the performance of ASR models on Croatian and Serbian, we used a balanced data set representing four

variants. However, the model training and fine-tuning settings could not be controlled due to limited resources. As a consequence, our settings are not fully comparable. On one hand, we do not have results of the Whisper model fine-tuned on Croatian to compare it with the one fine-tuned on Serbian. On the other hand, we have no results showing what happens if we train from scratch on Serbian or on a combination of Croatian and Serbian. Finally, W2V2 models are fine-tuned only on Croatian. Observing all the settings in a fully comparable way would make our conclusions more sound, but the partial observations that were possible by reusing existing information already show interesting patterns that can be helpful for designing fully comparable settings in future studies.

Another limitation of our study is leaving aside the question of model response time, which is an important issue in speech processing. Most of the models that we analysed are likely to respond too slowly for a practical application. With the available data, we could not analyse this aspect in detail, but we could show that models trained from scratch can outperform large pre-trained models, which is an interesting point for future research taking into account practical aspects such as response time as well.

Finally, our study is performed on data and models that were available at the time of the original evaluation. In the meantime, new data sets were published increasing significantly the size of available data for both Croatian and Serbian. Also a few new models are emerging as good candidates for studies such as ours. We believe that, despite this limitation, our insights can inform future studies regarding the factors to be tested with new data and models.

Acknowledgments

References

- Maryam Khalifa Al Ali and Hanan Aldarmaki. 2024. [Mixat: A data set of bilingual emirati-English speech](#). In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 222–226, Torino, Italia. ELRA and ICCL.
- Mohammad Al-Fetyani, Muhammad Al-Barham, Gheith Abandah, Adham Alsharkawi, and Maha Dawas. 2021. [Masc: Massive arabic speech corpus](#).
- Sadeen Alharbi, Areeb Alowisheq, Zoltán Tüske, Kareem Darwish, Abdullah Alrajeh, Abdulmajeed Alrowithi, Aljawharah Bin Tamran, Asma Ibrahim,

Raghad Aloraini, Raneem Alnajim, Ranya Alkahtani, Renad Almuasaad, Sara Alrasheed, Shaykhah Alsubaie, and Yaser Alonaizan. 2024. [Sada: Saudi audio dataset for arabic](#). In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10286–10290.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhota, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. [XLS-R: self-supervised cross-lingual speech representation learning at scale](#). *CoRR*, abs/2111.09296.

Amira Dhouib, Achraf Othman, Oussama El Ghoul, Mohamed Koutheair Khribi, and Aisha Al Sinani. 2022. [Arabic automatic speech recognition: A systematic literature review](#). *Applied Sciences*, 12(17).

Amirbek Djanibekov, Hawau Olamide Toyin, Raghad Alshalan, Abdullah Alatir, and Hanan Aldarmaki. 2025. [Dialectal coverage and generalization in Arabic speech recognition](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29490–29502, Vienna, Austria. Association for Computational Linguistics.

Nikola Ljubešić, Danijel Koržinek, and Peter Rupnik. 2024a. [Parliamentary spoken corpus of Croatian ParlaSpeech-HR 2.0](#). Slovenian language resource repository CLARIN.SI.

Nikola Ljubešić, Danijel Koržinek, Peter Rupnik, and Ivo-Pavao Jazbec. 2022. [ParlaSpeech-HR - a freely available ASR dataset for Croatian bootstrapped from the ParlaMint corpus](#). In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, pages 111–116, Marseille, France. European Language Resources Association.

Nikola Ljubešić, Danijel Koržinek, Peter Rupnik, Ivo-Pavao Jazbec, Vuk Batanović, Lenka Bajčetić, and Bojan Evkoski. 2022. [ASR training dataset for croatian ParlaSpeech-HR v1.0](#). Slovenian language resource repository CLARIN.SI.

Nikola Ljubešić, Peter Rupnik, and Danijel Koržinek. 2024b. [Parliamentary spoken corpus of Serbian ParlaSpeech-RS 1.0](#). Slovenian language resource repository CLARIN.SI.

Nikola Ljubešić, Peter Rupnik, and Tea Perinčić. 2024c. [The "Mići Princ" text and speech dataset of Chakavian micro-dialects](#). Slovenian language resource repository CLARIN.SI.

Hamdy Mubarak, Amir Hussein, Shammur Absar Chowdhury, and Ahmed Ali. 2021. [QASR: QCRI aljazeera speech resource a large scale annotated Arabic speech corpus](#). 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 2274–2285, Online. Association for Computational Linguistics.

- Iuliia Nigmatulina, Tannon Kew, and Tanja Samardžić. 2020. [ASR for non-standardised languages with dialectal variation: the case of Swiss German](#). In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 15–24, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).
- Michel Plüss, Jan Deriu, Yanick Schraner, Claudio Paonessa, Julia Hartmann, Larissa Schmidt, Christian Scheller, Manuela Hürlimann, Tanja Samardžić, Manfred Vogel, and Mark Cieliebak. 2023. [STT4SG-350: A speech corpus for all Swiss German dialect regions](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1763–1772, Toronto, Canada. Association for Computational Linguistics.
- Michel Plüss, Manuela Hürlimann, Marc Cuny, Alla Stöckli, Nikolaos Kapotis, Julia Hartmann, Malgorzata Anna Ulasik, Christian Scheller, Yanick Schraner, Amit Jain, Jan Deriu, Mark Cieliebak, and Manfred Vogel. 2022. [SDS-200: A Swiss German speech to Standard German text corpus](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3250–3256, Marseille, France. European Language Resources Association.
- Michel Plüss, Lukas Neukom, Christian Scheller, and Manfred Vogel. 2021. [Swiss parliaments corpus, an automatically aligned swiss german speech to standard german text corpus](#). *Preprint*, arXiv:2010.02810.
- Branislav Popović, Stevan Ostrogonac, Edvin Pakoci, Nikša Jakovljević, and Vlado Delić. 2015. Deep neural network based continuous speech recognition for serbian using the kalditoolkit. In *Speech and Computer*, pages 186–192, Cham. Springer International Publishing.
- Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The kalditoolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.
- Peter Rupnik and Nikola Ljubešić. 2022a. [ASR training dataset for serbian JuzneVesti-SR v1.0](#). Slovenian language resource repository CLARIN.SI.
- Peter Rupnik and Nikola Ljubešić. 2022b. [ASR training dataset for Serbian JuzneVesti-SR v1.0](#). Slovenian language resource repository CLARIN.SI.
- Andrija Sagić. 2023. [Whisper-large-v3-sr-combined](#).
- Tanja Samardžić, Yves Scherrer, and Elvira Glaser. 2016. [ArchiMob - a corpus of spoken Swiss German](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4061–4066, Portorož, Slovenia. European Language Resources Association (ELRA).
- Tanja Samardžić, Peter Rupnik, Mirjana Starović, and Nikola Ljubešić. 2024. [Mak na konac: A multi-reference speech-to-text benchmark for croatian and serbian](#). Institute of Contemporary History.
- Yves Scherrer, Tanja Samardžić, and Elvira Glaser. 2019. [Digitising Swiss German: how to process and study a polycentric spoken language](#). *Language Resources & Evaluation*, 53:735–769.
- Yanick Schraner, Christian Scheller, Michel Plüss, and Manfred Vogel. 2022. [Swiss german speech to text system evaluation](#). *Preprint*, arXiv:2207.00412.
- Changhan Wang, Morgane Rivière, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Miguel Pino, and Emmanuel Dupoux. 2021. [Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation](#). *CoRR*, abs/2101.00390.

Syllable Structures Across Arabic Varieties

Abdelrahim Qaddoumi^{1,2} Jordan Kodner^{1,2} Salam Khalifa^{1,2,3}
Ellen Broselow¹ Owen Rambow^{1,2}

¹Institute for Advanced Computational Science; ²Department of Linguistics
Stony Brook University

³Computational Approaches to Modeling Language (CAMEL) Lab, NYU Abu Dhabi
{first.last}@stonybrook.edu

Abstract

This study compares the syllable structures of nine Arabic varieties from Wiktionary, using a computational syllabifier. It further investigates methods for learning syllable boundaries in unsyllabified words transcribed in the International Phonetic Alphabet (IPA). The syllabification algorithm is evaluated under three conditions: (i) **Default**, employing fixed rules; (ii) **Joint**, learning onsets and codas across all varieties collectively; and (iii) **Per-variety**, learning onsets and codas specific to each variety. Results indicate that the default configuration yields the highest accuracy, ranging from 97.05% to 100%. The per-variety approach achieves 90.64% to 100% accuracy, while the joint approach ranges from 84.63% to 94.74%. A cross-variety analysis using Jensen-Shannon divergence reveals three principal groupings: Egyptian, Hejazi, and Modern Standard Arabic are closely related; Levantine and Gulf varieties constitute a second cluster; and Juba Arabic, Maltese, and Moroccan emerge as outliers. A cleaned dataset encompassing all nine varieties is also provided.

1 Introduction

One of the fundamental elements of phonology is the syllable (Goldsmith, 2011), abstract constituents of the mental representation of sound structure (Al-Ani and May, 1973). While they are not always detectable from the audio signal, they have been invaluable for the study of phonological processes such as word-stress assignment (Broselow, 2017).

Syllables play a prominent role in early language acquisition. Infants as young as four days old perceive syllables and can discriminate words based on syllable length (Jusczyk and Derrah, 1987; Bijeljac-Babic et al., 1993). Syllables for the basic units for word segmentation, where errors overwhelmingly line up with syllable boundaries (Peters, 1983). Given this, syllables are commonly

Word	da . ras . ha #	dar . ras #	drUs #	dars
Structure	Cv . CvC . Cv #	CvC . CvC #	CCVC #	CvCC
Onset	d	r	h	d
Nucleus	a	a	a	a
Coda	∅	s	∅	r

Table 1: Syllabified examples, in Levantine Arabic, decomposed into onset, nucleus, and coda, with markers for syllable boundary ., and word boundary #. The syllable *structure* is represented with C for consonants, v for short vowels, and V for long vowels. The glosses for the words are <دَرَسَهَا> ‘he studied it [f.sg]’, <دَرَسَ> ‘he taught’, <دروس> ‘lessons’, and <دَرَس> ‘lesson’, respectively.

assumed as the basic unit in the developmental literature on word segmentation, both experimental Saffran et al. (1996) and computational (Lignos, 2011; Fourtassi et al., 2013). Syllable-based models for word segmentation have been shown to outperform phoneme-based models (Schrimpf and Jarosz, 2014).

Syllables have also played a similarly important role in word segmentation as an NLP task, forming the basic unit of segmentation in languages with scripts that lack explicit word spacing (Htay and Murthy, 2008; Chormai et al., 2020). In downstream applications, syllable boundaries have been shown to improve the performance of morphological analyzers (Khalifa et al., 2025) and to improve the generalization ability of a Chinese LLM in the presence of phonological ambiguities by introducing subtasks, such as converting orthography to a syllabified form written in IPA and handling tone. In addition, they use LLMs to generate synthetic data containing syllable and tone information. They show that using these subtasks and synthetic data for training their model yields performance gains (Ma et al., 2025).

However, what counts as an acceptable syllable is not identical across all languages, or even di-

lects of the same language. In this paper, we study syllable structure across Arabic varieties, which diverge considerably in what they accept as a syllable. For example, in Egyptian, it is rare to see syllable structures such as CCvC in loanwords as in <كسيون> [kum.sjɔn] ‘commission’. Moroccan has triple consonant CCC onset clusters, for example, <استعداد> [stʃa:d] to call for ‘help’. Maltese has single vowels as independent syllables, for example, <emozzjona> [ɛ.mɔt.tsjɔ:.na] ‘to move’.

Previous studies have examined empirical data on Arabic syllable structure. Hamdi et al. (2005) analyzed the syllable structures of Moroccan, Tunisian, and Lebanese varieties, showing that the frequency of syllable types differs across varieties. For example, Lebanese tends to favor syllables with long vowels, such as CV and CVC, while Moroccan prefers syllable structures with more consonants, like CCvC and CCvCC. Our analysis will include additional varieties and quantify their similarities based on syllable distributions.

Unsurprisingly given the importance of syllables for linguists and developmental researchers, There is a long history of research on automatic syllabification, though this has primarily focused on English. Marchand et al. (2007) compares different algorithms, including Fisher (1996)’s rule-based approach based on Kahn’s procedure (Kahn, 2015), as well as several data-driven methods: syllabification by analogy, a look-up procedure, and exemplar generalization. Their results show that data-driven approaches outperform rule-based ones. Wu and Yarowsky (2021) explores multilingual syllabification and stress prediction, using phonemic representations in IPA transcriptions.

In this paper, we automatically extract syllables from a range of Arabic varieties using Kodner (2016)’s SIMPLESYLLABIFY, a simple, language-independent syllabifier (see Section 3) that can run on both orthographic and phonemic representations. This paper makes the following contributions to the study of Arabic syllable structure:

1. We introduce a new IPA dataset derived from Wiktionary as a “silver standard” for Arabic syllabification (Section 4).
2. We use this dataset to evaluate the performance of SIMPLESYLLABIFY (Section 5).
3. We use it to perform a cross-variety comparison of syllable inventories (Section 6). We use targeted error analysis (Section 7) to separate algorithmic errors from data artifacts.

2 Linguistic Background

This section provides the linguistic background needed to follow the paper. Our work does not make claims about which linguistic theory is correct; we simply state the theoretical assumption used in this analysis. For terminological convenience, we use the terms *variety* and *varieties* to mean all of the following: Arabic dialects, Modern Standard Arabic (MSA), Juba, and Maltese.

In the following discussion, forms in square brackets ([]) represent transcriptions of surface forms (SF) using the International Phonetic Alphabet (IPA), while those between slashes (/ /) represent underlying representations (UR). Arbitrary consonants are represented with C, and short and long vowels with v and V, respectively. Specific short and long vowels are also represented in lower and upper case, for example, short [a] and long [A]. This is sufficient granularity for the present study of Arabic. Breaks between syllables are indicated with a period (.), and word boundaries are indicated by ‘#’. For example, in South Levantine, <درس> /darras/ ‘he taught’ has a geminate: [dar.ras] (CvC.CvC), while <درس> /dars/ ‘lesson’ has a complex cluster: [dars] (CvCC).

2.1 Syllables

In this paper, we divide syllables into three constituents, the onset, the nucleus, and the coda (Hayes, 2009). The nucleus is the core of the syllable, and usually takes the form of a vowel or a vocalic consonant. The onset is the segment preceding the nucleus, and the coda is the segment following the nucleus. Table 1 demonstrates examples of this definition, it showcases different shapes of syllables in different contexts.

2.2 Arabic Varieties

Arabic is often treated as a single language. However, it is actually a collection of regional varieties that differ in many linguistic dimensions, including their syllable inventories and phonological patterns. These regional varieties are the native varieties of Arabic speakers across the Arabic-speaking world and coexist alongside MSA. MSA is considered the “prestige” variety, as it is used in formal settings, while the regional spoken variety is used day to day. This special phenomenon of a single community using two dialects or varieties is described in the literature as diglossia (Ferguson, 1959).

In this paper, we analyze the following varieties: Egyptian (EGY), Gulf (GUL), Hejazi (HEJ), Juba (JUB), Maltese (MAL), Moroccan (MOR), North Levantine (NLE), South Levantine (SLE), and MSA. MSA is the standard written form across the Arabic-speaking world. The Egyptian Arabic variety is notable for its prominence in media across the Arab world, its wide intelligibility, and it is spoken by 119 million people. Moroccan Arabic, a Maghrebi variety, is spoken by 40 million people and is influenced by contact with Tamazight. Hejazi Arabic, spoken in western Saudi Arabia, and Gulf Arabic, spoken in the states bordering the Persian Gulf, are Arabian Peninsula varieties, each with about 11 million speakers. Levantine encompasses about 60 million speakers and is divided between South Levantine (Jordanian/Palestinian) and North Levantine (Lebanese/Syrian) (Eberhard et al., 2025).

The remaining two varieties are less frequently featured in comparative studies. Located in the peripheries of the Arabic-speaking world, they have been subject to extensive contact with non-Semitic languages. Juba Arabic is an Arabic-based creole spoken by 250,000 people in South Sudan. Maltese, which around 500,000 people speak in Malta (Rosner and Borg, 2022), is an official EU working language descended from medieval Siculo-Arabic. It is a relative of modern North African varieties but with substantial influence from Romance languages (Brincat, 2005). It is notable as one of the few Semitic languages to be written in a Latin script. Taken together, its orthography and institutional status make it atypical relative to Arabic regular diglossia.

2.3 Syllable Structure Variation

Due to diglossia between local varieties and MSA, data for any local variety is relatively scarce. Diglossia presents a research challenge for both NLP and linguistics. In NLP, Arabic regional varieties are low-resource, making it challenging to work with limited data. However, their syllable structures vary in predictable ways, providing ground for cross-variety transfer techniques. From a linguistic perspective, this is an ideal setting for quantitatively comparing features of linguistic variety, such as syllable structures.

Since Arabic varieties are closely related yet differ in the syllable structures that they employ, they constitute a useful comparison set in the phonological study of syllables. We illustrate some of the

differences here. All Arabic varieties in this paper, allow at least the following syllable structures: CV, and CVC, except for Juba Arabic, which lacks long vowels. For a complete literature summary, see Table 4.

JUB additionally permits CCv (Manfredi and Petrollino, 2013). Onsets are mandatory in Arabic, so underlyingly vowel-initial syllables are realized with a glottal stop [ʔ] (Broselow, 2017). This is generally true in MAL as well, making V marginal as a phonotactic template (Galea and Ussishkin, 2018).

While CC and CCC are possible onsets in MOR in analyses such as (Kiparsky, 2003), there is still debate about how to analyze syllables in MOR. For example, Shaw et al. (2009) use articulatory measures to argue for syllables consisting only of simplex-onsets. Thus, /skru/ ‘his ploughshares’ surfaces as [s.kru] “C.CCv” instead of [skru] “CCCv”. MOR has been shown to have complex-onset varieties such as CCv/CCvC, CCCv, and CC-CvC in the surface form (Hamdi et al., 2005).

SLE and NLE allow for CCvC and CCVC structures (Hamdi et al., 2005). Moreover, GLF allows complex-onset CCV. Standard Maltese permits syllables of the form (C)(C)(C)V(V)(C)(C), so we find a range of complex onsets and codas in the language (Galea and Ussishkin, 2018).

Finally, Arabic varieties impose various restrictions on syllable types depending on word position. Some Levantine varieties allow CC onsets only in word-initial position; EGY allows CVC, CvCC only in word-final position (Broselow, 2017). Despite not using syllable position in the syllabification, the high accuracy in our data (Table 3) suggests that position-conditioned syllables are either rare in the data or are handled correctly by the default setting.

Table 4 illustrates how syllable-type inventories can differ across varieties. Examples of the differences are: the availability of complex onsets (e.g., CCvC/CCVC), the distribution of super-heavy syllables in word-final position, permissible onsets, and edge-sensitive restrictions on heavy codas. These are precisely the values that matter for the syllabification algorithm. Examples of these cases are explained in further detail in Section 3

3 Syllabification Method

To obtain the actual syllabification, we used the SIMPLESYLLABIFY syllabification script by Kod-

ner (2016)¹ which identifies syllable nuclei and then inserts syllable boundaries according to the Maximum Onset Principle (Kahn, 2015), a widely understood generalization from phonology (Goldsmith, 2011). The SIMPLESYLLABIFY is intended to run on orthographic representations, but, as there are no standard orthographies for Arabic varieties, standard Arabic orthography usually omits short vowels, and IPA has wide acceptance among linguists, we use IPA-like transcriptions from Wiktionary, as described in Section 4.1.

The script requires a list of characters that represent syllable nuclei (usually vowels, but possibly characters representing syllabic sonorants).² Given this set of characters, the script proceeds by identifying all of the syllable nuclei in a word. It then considers each consonant cluster to the left of a nucleus as a potential onset. Maximum Onset Principle states that the largest string of onset consonants in this cluster that can licitly be incorporated into the onset should be, and the remainder is incorporated into the previous syllable’s coda. All word-initial consonants are added to the initial syllable’s onset, and all word-final consonants are added to the final syllable’s coda.

Since what counts as a licit onset differs from one variety to another, the script defaults to splitting consonant clusters in half, assigning half (plus one in the case of odd-sized clusters) into the onset, with the remainder added to the previous syllable’s coda. For example, vCCCCV is syllabified as vCC.CCV and vCCCV are syllabified as vC.CCV.

Alternatively, the script can learn a language-specific set of licit onsets from word-initial onsets in the dataset. However, this may hurt performance if the training set is small, leading to under-learning, or if the variety allows word-initial onsets that are banned word-medially. Unfortunately, these are both problems in our case, and learning from word-initial onsets dramatically lowers performance (See Table 3). A third option is for the user to provide an explicit list of valid onsets. However, specifying variety-specific onsets up-front undermines the spirit of our experiments. Since syllabification performance is over 96% for all varieties and above 99% for the majority, we find that

¹<https://github.com/jkodner05/syllabify>

²The following set of symbols were treated as vowels for the purposes of identifying nuclei: ʊ, a, ɜ, u, e, æ, o, ɑ, i, ɪ, ə, ɔ, ɐ, ə, ä, ε, ɯ, ɔ̃, œ, ɨ, ɒ, ʌ, ʊ, A, U, E, Æ, O, Q, I, T, ʒ, ɔ̃, V, Á, É, ʘ, Ö, Œ, Ǽ, ǻ. Lowercase symbols are the IPA vowels used in Wiktionary. We used the upper case symbols for our normalization of long vowels.

the default splitting approach is sufficient for our purposes.

It is theoretically possible to learn per-variety syllabification algorithms that account for variety-specific inventories of licit onsets and codas, the sonority of different consonants, and whether the variety allows rising or lowering sonority in onsets or codas. In practice, this is difficult to derive from Wiktionary data because it is limited and noisy. This is why we use a general deterministic algorithm.

4 Data

4.1 Wiktionary

Wiktionary³ is a free, collaborative online lexicon that provides structured entries for many languages, often including IPA transcriptions, audio, and glosses (Meyer and Gurevych, 2012). It has data for 22 Arabic varieties. From these, we selected the varieties for which there were at least 100 entries with syllabified IPA transcriptions: Egyptian (EGY, 584 words), Gulf (GUL, 636 words), Hejazi (HEJ, 2,919 words), Juba (JUB, 189 words), Maltese (MAL, 17,609 words), Moroccan (MOR, 1,955 words), Modern Standard Arabic (MSA, 16,718 words), North Levantine (NLE, 575 words), and South Levantine (SLE, 5,059 words).⁴ We extracted our datasets from a Wiktionary dump (Ylönen, 2022) and processed them into per-variety word lists.

4.1.1 Preprocessing

While Wiktionary is overall a reliable source, it is crowdsourced, and thus the data may be noisy (Sakunkoo and Sakunkoo, 2025). We noticed that IPA conventions vary across contributors, which we address with preprocessing. We remove duplicates and any entries with transcriptions that contain only one or two characters, since they are trivial to syllabify. We further standardize transcriptions as follows: Long vowels are capitalized in order to represent them as a single character, for example, < كاتب > [ka:.tib] ‘writer’, will be [kA.tib]. Geminates are represented by repeating the same character, for example, < كتّاب > [kat:ab] ‘he made x write’, will be represented as [kat.tab]. Finally, we represent the sibilant affricate segments “ts”, “dʒ”, and “tʃ” as single consonants “T”, “D”, “S”

³<https://www.wiktionary.org/>

⁴We omit Algerian, Andalusian, Baharna, Chadian, Cypriot, Hassaniya, Iraqi, Libyan, North Mesopotamian, Omani, Sudanese, Tunisian, and Yemeni.

respectively. This decision is supported by the fact that affricates are contour segments, which means that they have two phonetic qualities, but can be treated as a single segment phonologically (Hayes, 2009).

When an entry provides multiple IPA transcriptions, we retain each distinct transcription as a separate surface form for that variety, for example, <شوكولاتة> [ʃu.kU.lA.ta] or [ʃu.ku.lA.ta] ‘chocolate’. For the cases where one segment is presented as optional in a single transcription, we always include the optional segment, as in the following example, where <اسمر> /ʔ(ɪ)smar/ ‘dark’ is treated as [ɪs.mar].

Stress marks are removed. All superscripts such as ‘ˀ’ and subscripts such as ‘ˁ’ are removed as well, since these correspond to annotations or to phonological distinctions which do not affect Arabic syllabification (Association, 1999), for example, the tie bar ‘ˁ’ and pharyngeal diacritics ‘ˀ’ in <آية الله> /ʔA.ja.tu.lʔ.lʔAh/ ‘sign of God’. We remove [] and //, since they do not represent segments.

4.2 “Bronze”, “Silver”, and “Gold” Standard

We compare five tiers of postprocessing from the raw Wiktionary data to quantify (i) how well the rule-based syllabifier matches crowdsourced data and (ii) how much of the observed mismatch is due to data labeling errors instead of syllabifier errors. **Wiktionary** is raw data with the website’s syllable annotations preserved as-is. **Predicted** is the output of the syllabifier applied to Wiktionary IPA with original syllabification information discarded. **Bronze** excludes the three easy-to-identify mislabeling errors. The data excluded is described in the following subsection and reported in Table 2.

Silver retains **Bronze** entries and automatically corrects mislabeling as described below. **Gold** retains entries where Predicted matches Wiktionary, and manually corrects mislabeling by a native speaker.

4.2.1 Bronze Data

To generate the Bronze data, we remove the following labeling errors in the Wiktionary data. The first is Under-Segmentation (US) errors, where transcriptions clearly have an insufficient number of syllable boundaries N_{syll} compared to number of vowels N_{vowel} . Since there should be a boundary somewhere between every pair of vowels, we can exclude words where $(N_{\text{syll}} < N_{\text{vowel}} - 1)$ (Hayes, 2009). An annotation like [ka.tabɑ] can be ex-

cluded because it has a structure CvCvCv with three vowels, but only one boundary.

The second type is Geminate (G) errors, which occur when transcriptions contain geminate sounds in onsets rather than splitting them into a preceding coda and a following onset, as they should be treated in Arabic (Farwaneh, 2009). For example, [ka.ttab] can be excluded, since it should be annotated as [kat.tab]. The third type is the Single Consonant (C) errors, which occur when syllables are annotated without a nucleus (v or V). For example, [k.at.tab] and [ka.t.tab] are excluded on these grounds.

4.2.2 Silver Data

To generate the Silver data, we first include any words where the predicted syllables match the reference. For items with mismatches, we apply targeted heuristics to recover reliable, predictable errors. For the single-consonant syllable error, we reattach the single-C syllable in the reference. If it is word-initial, it is attached to the subsequent syllable [k.at.tab] → [kat.tab]; if it is word-medial, it will be attached to the previous syllable [ka.t.tab] → [kat.tab].

Geminate errors are corrected by splitting the geminate and attaching it to the previous syllable if it is word-medial [ka.ttab] → [kat.tab]. For under-segmentation errors, we look for words that do not contain triple-consonant sequences (CCC) in their structure, as these are generally more prone to error. If the word does not contain CCC, then we apply our syllabification to it.

4.3 Gold

The author-annotators checked a random sample of 100 words in three varieties (SLE, NLE, EGY). In these samples, we found almost no issues: 0 for EGY, 1 for NLE, and 1 for SLE. We could not manually correct additional varieties because we lacked native speakers. We use this **Gold** tier to check that the **Silver** automatically fixed data matches the manually labeled data. The number of mismatched data points between Gold and Silver is 2 for EGY, 5 for NLE, and 50 for SLE.

5 Evaluation

The evaluation is conducted on the bronze, silver, and gold datasets. These datasets are described in Section 4.2. Data points listed in Table 2 are excluded from the evaluation of bronze only.

Variety	US	US+G	G	C	US+C	G+C
EGY	186	32	3	2	2	0
GUL	184	39	13	0	0	0
HEJ	296	25	0	22	0	0
JUB	74	1	0	1	0	0
MAL	177	9	1	1,288	11	3
MOR	58	2	0	2	0	0
MSA	112	7	0	3	0	0
NLE	106	16	0	11	0	1
SLE	97	13	1	1	0	0

Table 2: Counts of excluded words from bronze by error type. Some words exhibited two error types simultaneously. *US* under-segmentation error. *G* geminate error. *C* single consonant error.

We provide two evaluation metrics: accuracy, computed as the percentage of words whose predicted syllabification matches the annotated reference syllabification exactly, and average edit distance between the predicted and reference syllabifications. For example, if the reference word is <kiten> [kɪ.tən] and the predicted is [kɪtən], then the edit distance is one (insertion of the syllable boundary marker); if the output is [kɪt.ən], the edit distance is 2 (deletion of the incorrect boundary and insertion of the correct one).

From the results summarized in Table 3, we see that across all tiers, the Default setting achieves almost perfect accuracy (97–100%) for every variety, while Joint achieves the lowest accuracy (84–97%). The Per-variety setting is usually similar to Default, but it underperforms for MAL and MOR. Since the Bronze and Silver tiers are automatically derived, performance on these sets could potentially be inflated in principle, so they should be compared with Gold, where the same ranking (Default/Per-variety much greater than Joint) remains.

6 Analysis

We examine the distribution of syllable structures across the different varieties. The goal is to see how syllable structure differs among them. To run the analysis, we generate syllable-shape distributions for five tiers: Predicted (syllabifier output), Wiktionary (raw), Bronze, Silver, and Gold.

We summarize the distribution of the syllables in Table 4. While the distributions captured from the predictions and from the silver annotations are very similar, they are not identical. To better understand the difference between them, we use Jensen-Shannon divergence (JSD), which quantifies the difference between probability distributions, and has been widely used in the analysis of symbolic

Data			Default		Joint		Per-variety	
Tier	Variety	# Words	% Acc	SED	% Acc	SED	% Acc	SED
B	EGY	359	99.44	2.0	90.81	2.3	99.44	2.0
B	GUL	398	100.00	0.0	94.74	2.1	100.00	0.0
B	HEJ	2,574	99.65	2.2	91.46	2.0	99.22	2.1
B	JUB	104	100.00	0.0	93.27	2.0	98.08	2.0
B	MAL	16,119	99.15	2.0	84.63	2.0	90.64	2.0
B	MOR	1,893	99.74	2.0	85.63	2.0	93.03	2.0
B	MSA	16,595	99.96	1.7	87.87	2.1	99.27	2.0
B	NLE	441	97.05	1.9	89.34	2.0	97.05	1.9
B	SLE	4,947	98.99	2.0	88.56	2.1	98.24	2.0
S	EGY	584	99.66	2.0	94.35	2.3	99.66	2.0
S	GUL	636	99.84	2.0	96.54	2.1	99.84	2.0
S	HEJ	2,919	99.45	2.1	92.22	2.0	99.08	2.0
S	JUB	189	100.00	0.0	96.30	2.0	98.94	2.0
S	MAL	17,609	98.98	1.9	85.01	2.1	90.89	2.0
S	MOR	1,955	99.64	2.0	85.98	2.0	93.15	2.0
S	MSA	16,718	99.73	1.1	87.73	2.1	99.05	1.8
S	NLE	575	97.04	1.9	90.78	2.0	97.04	1.9
S	SLE	5,059	97.98	1.5	87.78	2.0	97.25	1.6
G	EGY	584	100.00	0.0	94.52	2.3	100.00	0.0
G	NLE	575	97.74	1.9	91.48	2.0	97.74	1.9
G	SLE	5,059	98.10	2.0	87.94	2.1	97.43	2.0

Table 3: Variety-level evaluation across three settings and three tiers (B=Bronze, S=Silver, G=Gold). Default, Joint (pooled onsets + codas), and Per-variety (variety-specific onsets + codas) report word-level exact-match accuracy and mean edit distance (SED). Gold is available only for EGY, NLE, and SLE.

sequences (Grosse et al., 2002). JSD is related to the Kullback-Leibler divergence (KLD), so it shares the same mathematical properties, but it differs from KLD in two essential ways: it is symmetrical and smooth, unlike KLD (Fuglede and Topsoe, 2004).

6.1 Data Quality Validation

The primary pattern to highlight in Table 5 is that the Gold results validate the automatic correction stage for the three manually checked varieties. The Silver distributions for SLE, NLE, and EGY are indistinguishable from their corresponding Gold distributions, whereas the Wiktionary distributions differ significantly from Gold for varieties such as EGY and NLE.

Silver is a proxy for Gold These patterns indicate that a large portion of the mislabeled data in Wiktionary is automatically repairable, and that the Silver tier is a good proxy for the Gold tier. These results motivate the use of Silver for the following cross-variety analyses.

Cross-variety Differences We observe notable differences between the varieties, for example, in their relative proportions of Cv and CvC. MSA and JUB prefer the former, whereas MOR, NLE, and SLE favor the latter. The rest do not seem to have a clear preference between Cv or CvC. We also see that some syllable structures are nearly ab-

Varieties		Cv	CvC	CVC	CV	CvCC	CCvC	CCVC	CCV	vC	CCv	CVCC	v	Other
EGY	L.	✓	✓	F	✓	F	X	X	X	X	X	-	X	0.4
	P.	37.6	36.4	12.8	6.8	5.8	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.4
	B.	30.7	34.3	19.4	4.7	10.1	0.3	0.3	0.0	0.0	0.0	0.0	0.0	0.3
	S.	37.5	36.5	12.7	6.8	5.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7
	G.	37.6	36.4	12.8	6.8	5.8	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.4
GUL	L.	✓	✓	✓	✓	F	I	I	I	X	✓	-	X	0.4
	P.	33.9	34.3	13.6	10.1	3.1	0.7	1.9	0.9	0.0	0.5	0.4	0.3	0.4
	B.	34.6	30.3	16.7	8.4	4.9	0.4	3.0	0.5	0.0	0.3	0.7	0.0	0.1
	S.	33.9	34.2	13.6	10.0	3.1	0.7	1.8	0.9	0.0	0.5	0.4	0.3	0.5
HEJ	L.	✓	✓	✓	✓	F	*	*	*	X	*	F	X	0.8
	P.	34.5	37.9	13.5	10.0	3.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8
	B.	32.9	38.8	14.2	9.6	3.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8
	S.	34.5	37.8	13.5	10.0	3.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.9
JUB	L.	✓	✓	X	X	X	*	X	X	*	F	X	✓	0.2
	P.	61.7	29.1	0.0	0.0	0.5	0.0	0.0	0.0	2.6	1.0	0.0	5.0	0.2
	B.	60.9	35.0	0.0	0.0	0.5	0.0	0.0	0.0	2.3	0.9	0.0	0.0	0.5
	S.	61.7	29.1	0.0	0.0	0.5	0.0	0.0	0.0	2.6	1.0	0.0	5.0	0.2
MAL	L.	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	2.2
	P.	30.4	30.7	8.0	6.6	2.7	3.6	1.3	2.1	7.4	3.3	0.0	1.8	2.2
	B.	30.7	31.5	8.2	6.5	3.0	3.6	1.4	1.9	8.1	3.4	0.0	0.0	1.8
	S.	30.5	30.4	8.0	6.6	2.8	3.6	1.3	2.0	7.4	3.2	0.0	1.8	2.4
MOR	L.	✓	✓	*	*	✓	✓	*	*	X	✓	X	X	0.3
	P.	19.9	33.6	15.2	10.2	4.7	5.6	4.2	4.5	0.0	1.3	0.4	0.0	0.3
	B.	19.8	33.5	15.2	10.1	5.0	5.7	4.4	4.3	0.0	1.3	0.4	0.0	0.3
	S.	20.0	33.5	15.3	10.2	4.8	5.5	4.2	4.6	0.0	1.3	0.4	0.0	0.3
MSA	L.	✓	✓	✓	✓	✓	X	X	X	X	-	✓	X	0.5
	P.	38.5	30.7	12.4	13.5	4.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5
	B.	38.6	30.7	12.4	13.5	4.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4
	S.	38.6	30.7	12.4	13.5	4.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5
NLE	L.	✓	✓	✓	✓	✓	*	I	*	X	-	✓	X	0.9
	P.	28.3	33.5	12.6	11.9	4.5	2.5	2.4	2.5	0.0	0.5	0.5	0.0	0.9
	B.	24.1	33.4	15.6	11.8	6.9	2.0	2.7	1.6	0.0	0.0	0.7	0.0	1.2
	S.	28.3	32.5	12.9	12.0	5.1	2.5	2.0	2.4	0.0	0.4	0.5	0.0	1.5
	G.	28.2	32.7	12.9	12.1	5.3	2.4	1.9	2.4	0.0	0.6	0.5	0.0	1.1
SLE	L.	✓	✓	✓	✓	✓	*	*	*	X	*	-	X	0.1
	P.	30.8	45.3	8.7	8.4	2.9	2.1	0.8	0.8	0.0	0.3	0.0	0.0	0.1
	B.	31.1	45.1	8.9	8.5	3.4	1.4	0.7	0.7	0.0	0.0	0.0	0.0	0.2
	S.	31.0	44.9	8.8	8.4	3.4	1.4	0.7	0.7	0.0	0.0	0.0	0.0	0.7
	G.	30.9	45.1	8.8	8.4	3.8	1.4	0.7	0.7	0.0	0.0	0.0	0.0	0.2

Table 4: Distribution of syllable shapes reported in linguistics literature (*L.*), extracted from syllabifier predictions (*P.*) and dataset tiers (*B./S./G.*) for each variety. *Other* sums all remaining rare types of syllable shapes. ‘F’ means word-final, ‘I’ means word-initial, ‘*’ marginal could be for certain conditions such as sonority or certain segments or loanwords, X means it is illicit by the literature, ✓ means it exists based on the literature, and ‘-’ means the literature does not mention it. Egyptian syllables (McCarthy, 1979; Aquil, 2013), Gulf (Qafisheh, 1977; Al-Qenaie et al., 2011), Hejazi (Alzaidi et al., 2019; Bokhari, 2020), Juba (Miller, 2006), Maltese (Galea and Ussishkin, 2018), for Moroccan (Boudlal, 2001; Benhallam, 1989; Shaw et al., 2009; Heath, 2020), MSA (Halpern, 2009; Ryding, 2005), North Levantine (Kelly, 2021), and South Levantine (Rakhieh, 2009).

sent in some varieties; for example, CCV, and CCv syllables are prominent only in MOR, and MAL, and CVC and CV syllables are absent from JUB. The results are consistent with prior descriptions of these varieties. MAL, MOR, NLE, and SLE show more complex (CC) onsets than other varieties.

Tier Differences For each tier in Table 5, we compute the distribution of syllable shapes. Then we measure pairwise similarity using JSD. Low JSD values mean distributions are more similar.

From Table 5 we see that Wiktionary’s distribution can differ substantially from the predicted distribution. This disagreement is variety-dependent, ranging from near-zero (e.g., MSA) to large values (e.g., EGY, JUB, GUL). This range in val-

ues suggests that Wiktionary’s syllabification is broadly consistent with the syllabifier’s predictions for some varieties, whereas others’ are not. These differences may be due to annotation practice (e.g., treatment of glides, epenthetic glottal stops, or geminates) or to systematic entry noise.

The other pattern we noticed is that Predicted is very close to Silver (≈ 0) for all varieties. The result means that the Silver tier is distributionally almost identical to the syllabifier’s output, indicating that the automatic fixes are correcting Wiktionary’s inconsistencies without changing the syllabifier’s overall syllable structure profile, where we see a nontrivial Bronze–Silver difference in EGY, GUL, JUB, MAL, and NLE.

Lang	W-P	W-B	W-S	W-G	P-B	P-S	P-G	B-S	B-G	S-G
EGY	0.096	0.090	0.093	0.096	0.010	0.001	0.000	0.012	0.010	0.001
GUL	0.077	0.078	0.076	—	0.006	0.000	—	0.006	—	—
HEJ	0.015	0.014	0.013	—	0.000	0.000	—	0.000	—	—
JUB	0.084	0.104	0.084	—	0.021	0.000	—	0.021	—	—
MAL	0.001	0.009	0.001	—	0.008	0.000	—	0.008	—	—
MOR	0.006	0.006	0.005	—	0.000	0.000	—	0.000	—	—
MSA	0.000	0.000	0.000	—	0.000	0.000	—	0.000	—	—
NLE	0.047	0.039	0.039	0.043	0.006	0.001	0.001	0.005	0.005	0.000
SLE	0.005	0.002	0.000	0.002	0.002	0.003	0.002	0.001	0.000	0.001

Table 5: JSD across dataset tiers (rounded to 3 decimals). P=Predicted (syllabifier output), W=Wiktionary (raw), B=Bronze (matches+exclude known mislabels), S=Silver (matches+automatic fixes), G=Gold (matches+manual fixes; available only for SLE, NLE, EGY). **Metric.** Each entry is the Jensen–Shannon divergence (JSD) between syllable distributions.

	EGY	GUL	HEJ	JUB	MAL	MOR	MSA	NLE	SLE	Mean
EGY	—	0.020	0.004	0.128	0.081	0.074	0.008	0.036	0.017	0.046
GUL	0.020	—	0.017	0.144	0.054	0.034	0.019	0.011	0.013	0.039
HEJ	0.004	0.017	—	0.140	0.081	0.069	0.005	0.032	0.015	0.045
JUB	0.128	0.144	0.140	—	0.126	0.233	0.142	0.185	0.140	0.155
MAL	0.081	0.054	0.081	0.126	—	0.063	0.084	0.054	0.061	0.076
MOR	0.074	0.034	0.069	0.233	0.063	—	0.074	0.015	0.045	0.076
MSA	0.008	0.019	0.005	0.142	0.084	0.074	—	0.033	0.023	0.049
NLE	0.036	0.011	0.032	0.185	0.054	0.015	0.033	—	0.019	0.048
SLE	0.017	0.013	0.015	0.140	0.061	0.045	0.023	0.019	—	0.042

Table 6: Cross-variety distance matrix (Silver): JSD over syllable distributions. Entries are Jensen–Shannon divergences (JSD) between varieties’ Silver tier syllable distributions. The **Mean** column reports each variety’s mean JSD to the other varieties (excluding self).

6.2 Cross-variety Analysis

Table 6 reports Jensen–Shannon divergence (JSD) values of the syllable distributions of Silver-tier varieties. JSD is a symmetric measure of distributional dissimilarity, with lower values indicating greater similarity between distributions.

Variety Clusters The general conclusion from the data is that the distances are mostly minimal outside of JUB, MAL, and MOR. The results can be split into three clusters. The first one includes HEJ-EGY, HEJ-MSA, and EGY-MSA. The second cluster of varieties includes GUL, NLE, and SLE; these three varieties are closely related. The last cluster, with its most significant divergence, includes JUB, MAL, and MOR; these varieties are the furthest from the others.

Centrality (mean column) The Mean column provides an overview of each variety’s centrality. GUL is identified as the most central variety (Mean 0.039), whereas JUB is the most distant from the mean (0.155), followed by MAL and MOR both (0.076). JUB exhibits the lowest similarity to all

other varieties. MAL is the second least similar to most varieties, except in the cases of JUB and MOR, where NLE is the second least similar to JUB, and MSA is the second least similar to MOR.

These results align with the linguistic literature, as the following varieties are generally considered distinct from other Arabic varieties. For example, JUB is creole; Italian and other Romance languages heavily influence MAL, and Tamazight heavily influences MOR and has a very unique consonant cluster compared to other varieties.

7 Error Analysis

SIMPLESYLLABIFY’S errors are primarily of one type, which is the consistent division of word-internal CCC clusters into C.CC. As discussed in Section 3, this is the default behavior of the algorithm when no set of licit onsets is learned or pre-specified. Nevertheless, performance is very high, indicating that this is rarely a problem. For example, in NLE <بريتان> [birt.?An] ‘oranges’ was predicted as [bir.t?An] and in Moroccan, <أربعة> [?arb.ʕa] ‘four’ was predicted as [?ar.bʕa].

To further refine the syllabifications, we could employ a more complex syllabification algorithm grounded in additional phonology that uses sonority hierarchies, in addition to the Maximum Onset Principle, to adjudicate between syllabifications without explicitly specifying how a variety should split consonant clusters (Clements, 1990). However, sonority hierarchies can also vary cross-linguistically, meaning that segments tend to be arranged according to the Sonority Sequencing Principle, generally with rising sonority in onsets and falling sonority in codas. A simple Sonority Scale from most sonorous to least: vowel < glide < liquid < nasal < obstruent (fricative, stop) (Selkirk, 1984; Clements, 1990).

Second type of error is when there is a glide followed by a consonant, such as in <دوية> [du.wajb.ba] ‘a small animal’, the syllabifier generates [du.waj.bba]. The third error type occurs when there are diphthongs, which are two different vowels that share the same syllable nucleus; for example, the reference for JUB has <kweys> [kweis] ‘good’, and the predicted is [kwe.is]. While SIMPLESYLLABIFY has the means to capture these syllables, it requires additional language-specific parameters to be specified.

8 Conclusion and Future Work

We study the problem of syllabification of multiple Arabic varieties from their unsyllabified surface forms written in IPA. We use Maltese, Egyptian, South Levantine, North Levantine, Moroccan, Hejazi, Gulf, and Modern Standard Arabic to evaluate the syllabification algorithm on a wide geographic range and the Arabic continuum. We compare the algorithm’s results to the syllable structures described in the literature. Our method shows that applying a non-probabilistic rule-based algorithm is highly successful at syllabifying all Arabic varieties. Then we compare results from using a single syllabifier across all varieties versus one per variety, showing that a general syllabification algorithm performs better for these Arabic varieties. Another thing is how practical this syllabification algorithm is for cleaning data by comparing where it diverges. This allows us to understand the syllable structures in these varieties better.

Future work may involve extending this analysis to a broader range of languages to determine whether the results are consistent with those of the current study. Another avenue for exploration is

incorporating sonority ranking into the syllabifier to assess potential improvements in performance. Sonority ranking could be learned using a Bottom-Up Factor Inference Algorithm (BUFIA) (Chandlee et al., 2019), a grammar inference algorithm that infers the most general set of forbidden constraints from positive data. Additionally, methods for learning onsets and codas directly from data, without relying on a sonority hierarchy, warrant investigation.

The identification of different Arabic varieties can be investigated based on their syllable structures and distributions. Furthermore, incorporating this information into neural networks may facilitate data generation for low-resource varieties by augmenting Modern Standard Arabic data, thereby enabling transfer learning.

Acknowledgments

We thank Jeffrey Heinz for his helpful discussion and feedback. We also thank the anonymous reviewers for their feedback. Qaddoumi and Rambow gratefully acknowledge support from the Institute for Advanced Computational Science at Stony Brook University.

Limitations

One of the main limitations of this distributional analysis is that we are looking at the frequency of syllable structures across types rather than the actual token frequency. For example, it can be that the types with CV are less used in real life than those with a different syllable structure because many of the Arabic varieties delete certain short vowels in open syllables. Unfortunately, this is not something we can solve for now, as there is no real frequency for tokens for Arabic varieties for this type of data.

References

- Salman Al-Ani and David May. 1973. The phonological structure of the syllable. In Salman Al-Ani, editor, *Readings in Arabic Linguistics*, pages 113–125. Indiana University Linguistics Club, Bloomington.
- Shamlan Al-Qenaie and 1 others. 2011. *Kuwaiti Arabic: A socio-phonological perspective*. Ph.D. thesis, Durham University.
- Muhammad Swaileh Alzaidi, Yi Xu, and Anqi Xu. 2019. Prosodic encoding of focus in hijazi arabic. *Speech Communication*, 106:127–149.

- Rajaa Aquil. 2013. Cairne arabic syllable structure though different phonological theories. *Open Journal of Modern Linguistics*, 3(3):259–267.
- International Phonetic Association. 1999. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press.
- Abderrafi Benhallam. 1989. Moroccan arabic syllable structure. *Revue Langues et Littératures*, 8:177–191.
- Ranka Bijeljic-Babic, Josiane Bertoncini, and Jacques Mehler. 1993. How do 4-day-old infants categorize multisyllabic utterances? *Developmental psychology*, 29(4):711.
- Hassan Abdulrashid Bokhari. 2020. *A comprehensive analysis of coda clusters in Hijazi Arabic: An optimality-theoretic perspective*. Indiana University.
- Abdelaziz Boudlal. 2001. *Constraint interaction in the phonology and morphology of Casablanca Moroccan Arabic*. Ph.D. thesis, Rutgers University.
- Joseph M Brincat. 2005. Maltese—an unusual formula. *MED Magazine*, 27.
- Ellen Broselow. 2017. Syllable structure in the dialects of arabic. *The Routledge handbook of Arabic linguistics*, pages 32–47.
- Jane Chandlee, Remi Eyraud, Jeffrey Heinz, Adam Jardine, and Jonathan Rawski. 2019. **Learning with partially ordered representations**. In *Proceedings of the 16th Meeting on the Mathematics of Language*, pages 91–101, Toronto, Canada. Association for Computational Linguistics.
- Pattarawat Chormai, Ponrawee Prasertsom, Jin Cheevaprawatdomrong, and Attapol Rutherford. 2020. **Syllable-based neural Thai word segmentation**. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4619–4637, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- George N Clements. 1990. The role of the sonority cycle in core syllabification. *Papers in laboratory phonology*, 1:283–333.
- David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2025. *Ethnologue: Languages of the World*, 28 edition. SIL International, Dallas, Texas. Online edition.
- Samira Farwaneh. 2009. Toward a typology of arabic dialects: The role of final consonantality. *Journal of Arabic and Islamic studies*, 9:82–109.
- Charles F Ferguson. 1959. Diglossia. *Word*, 15(2):325–340.
- William Fisher. 1996. **Tsylb syllabification package**. FTP archive.
- Abdellah Fourtassi, Benjamin Börschinger, Mark Johnson, and Emmanuel Dupoux. 2013. **Why is English so easy to segment?** In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*, pages 1–10, Sofia, Bulgaria. Association for Computational Linguistics.
- Bent Fuglede and Flemming Topsoe. 2004. Jensen-shannon divergence and hilbert space embedding. In *International symposium on Information theory, 2004. ISIT 2004. Proceedings.*, page 31. IEEE.
- Luke Galea and Adam Ussishkin. 2018. *Onset clusters, syllable structure and syllabification in Maltese*. Language Science Press.
- John Goldsmith. 2011. The syllable. *The handbook of phonological theory*, pages 164–196.
- Ivo Grosse, Pedro Bernaola-Galván, Pedro Carpena, Ramón Román-Roldán, Jose Oliver, and H Eugene Stanley. 2002. Analysis of symbolic sequences using the jensen-shannon divergence. *Physical Review E*, 65(4):041905.
- Jack Halpern. 2009. Word stress and vowel neutralization in modern standard arabic. In *2nd International Conference on Arabic Language Resources and Tools*, pages 1–7. Cairo.
- Rym Hamdi, Salem Ghazali, and Melissa Barkat-Defradas. 2005. Syllable structure in spoken arabic: a comparative investigation. In *Eurospeech—9th European Conference on Speech Communication and Technology*.
- Bruce Hayes. 2009. *Introductory Phonology*. Number 23 in Blackwell Textbooks in Linguistics. Wiley-Blackwell.
- Jeffrey Heath. 2020. Moroccan arabic. *Arabic and contact-induced change*, 1:213.
- Hla Hla Htay and Kavi Narayana Murthy. 2008. **Myanmar word segmentation using syllable level longest matching**. In *Proceedings of the 6th Workshop on Asian Language Resources*.
- Peter W Jusczyk and Carolyn Derrah. 1987. Representation of speech sounds by young infants. *Developmental Psychology*, 23(5):648.
- Daniel Kahn. 2015. *Syllable-based generalizations in English phonology*. Routledge.
- Niamh Kelly. 2021. Syllable weight, vowel length and focus in lebanese arabic. *Glossa: a journal of general linguistics*, 6(1).
- Salam Khalifa, Abdelrahim Qaddoumi, Jordan Kodner, and Owen Rambow. 2025. Learning cross-dialectal morphophonology with syllable structure constraints. *VarDial 2025*, page 157.
- Paul Kiparsky. 2003. Syllables and moras in arabic. *The syllable in optimality theory*, 147:182.

- Jordan Kodner. 2016. [Simple Syllabify](#).
- Constantine Lignos. 2011. [Modeling infant word segmentation](#). In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 29–38, Portland, Oregon, USA. Association for Computational Linguistics.
- Jianfei Ma, Zhaoxin Feng, Emmanuele Chersoni, Huacheng Song, and Ziqi Zhang. 2025. Phonothink: Improving large language models’ reasoning on chinese phonological ambiguities. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19018–19033.
- Stefano Manfredi and Sara Petrollino. 2013. Juba Arabic. *The survey of pidgin and creole languages*, 3:54–65.
- Yannick Marchand, Connie R Adsett, and Robert I Damper. 2007. Evaluating automatic syllabification algorithms for english. In *Proceedings of the 6th ISCA Workshop on Speech Synthesis*.
- John J McCarthy. 1979. On stress and syllabification. *Linguistic inquiry*, 10(3):443–465.
- Christian M Meyer and Iryna Gurevych. 2012. *Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography*. na.
- Catherine Miller. 2006. Juba arabic. In Kees Versteegh, editor, *Encyclopedia of Arabic Language and Linguistics*, volume 2, pages 517–525. Brill, Leiden.
- Ann M Peters. 1983. *The units of language acquisition*. Cambridge University Press.
- Hamdi A Qafisheh. 1977. *A Short Reference Grammar of Gulf Arabic*. ERIC.
- Belal A. Rakhieh. 2009. *The Phonology of Ma’ani Arabic: Stratal or Parallel OT*. Ph.D. thesis, University of Essex.
- Mike Rosner and Claudia Borg. 2022. [D1.25: Report on the Maltese language](#). Deliverable D1.25 (Public), dated 28-02-2022.
- Karin C Ryding. 2005. *A reference grammar of modern standard Arabic*. Cambridge university press.
- Jenny R Saffran, Richard N Aslin, and Elissa L Newport. 1996. Statistical learning by 8-month-old infants. *science*, 274(5294):1926–1928.
- Jonathan Sakunkoo and Annabella Sakunkoo. 2025. Lost and found: Computational quality assurance of crowdsourced knowledge on morphological defectivity in wiktionary. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 998–1003.
- Natalie Schrimpf and Gaja Jarosz. 2014. [Comparing models of phonotactics for word segmentation](#). In *Proceedings of the 2014 Joint Meeting of SIGMORPHON and SIGFSM*, pages 19–28, Baltimore, Maryland. Association for Computational Linguistics.
- Elisabeth Selkirk. 1984. On the major class features and syllable theory. *Language sound structure*.
- Jason Shaw, Adamantios I Gafos, Philip Hoole, and Chakir Zeroual. 2009. Syllabification in Moroccan Arabic: evidence from patterns of temporal stability in articulation. *Phonology*, 26(1):187–215.
- Winston Wu and David Yarowsky. 2021. On pronunciations in wiktionary: Extraction and experiments on multilingual syllabification and stress prediction. In *Proceedings of the 14th Workshop on Building and Using Comparable Corpora (BUCC 2021)*, pages 68–74.
- Tatu Ylönen. 2022. Wiktextextract: Wiktionary as machine-readable structured data. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 1317–1325, Marseille, France.

Curriculum Learning and Pseudo-Labeling Improve the Generalization of Multi-Label Arabic Dialect Identification Models

Ali Mekky* Mohamed El Zeftawy* Lara Hassan* Amr Keleg Preslav Nakov

Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE

{firstname.lastname}@mbzuai.ac.ae

Abstract

Being modeled as a single-label classification task for a long time, recent work has argued that Arabic Dialect Identification (ADI) should be framed as a multi-label classification task. However, ADI remains constrained by the availability of single-label datasets, with no large-scale multi-label resources available for training. By analyzing models trained on single-label ADI data, we show that the main difficulty in repurposing such datasets for Multi-Label Arabic Dialect Identification (MLADI) lies in the selection of negative samples, as many sentences treated as negative could be acceptable in multiple dialects. To address these issues, we construct a multi-label dataset by generating automatic multi-label annotations using GPT-4o and binary dialect acceptability classifiers, with aggregation guided by the Arabic Level of Dialectness (ALDi). Afterward, we train a BERT-based multi-label classifier using curriculum learning strategies aligned with dialectal complexity and label cardinality. On the MLADI leaderboard, our best-performing LAHJATBERT model achieves a macro F1 of 0.69, compared to 0.55 for the strongest previously reported system. Code and data are available at <https://mohamedalaa9.github.io/lahjatbert/>.

1 Introduction

Arabic has a wide range of diverse dialects spoken across the Arab World. While Modern Standard Arabic (MSA) is used in official communication, education, and media, everyday conversations typically happen in local dialects (Habash, 2010). Dialects vary between countries, and can even vary within same-country cities and local communities, creating a complex linguistic landscape (Zaidan and Callison-Burch, 2014; Althobaiti, 2020).

Arabic Dialect Identification (ADI) is the task that aims to identify the dialect of a sentence. ADI

* Equal contribution.

Dialects	Sentence
Jordan, Palestine	اتفقنا نظل جنب بعض ع الحلوه والمره (We agreed to stick by each other through thick and thin.)
Algeria, Egypt, Jordan, Palestine, Sudan, Syria, Tunisia, Yemen	احسبو حسابي معاكم (Make sure to count me in with you.)
Algeria, Palestine, Sudan, Yemen	بسمتك يا زين تسوي ألف بسمه (Your smile, oh beautiful one, is worth a thousand other smiles.)
All Arabic Dialects	اللهم الجنه لمن ذهب ارواحهم إليك (O Allah, grant Paradise to those whose souls have returned to You.)

Table 1: Examples illustrating dialect overlap in Arabic, sampled from the manually annotated NADI2024 Sub-task 1 development set. **Note:** the dataset has labels for eight country-level dialects only.

has been modeled as a single-label classification problem, with systems trained and evaluated on resources such as AOC (Zaidan and Callison-Burch, 2011), MADAR (Bouamor et al., 2019), QADI (Abdelali et al., 2021), and NADI benchmarks (Abdulmageed et al., 2020, 2021b, 2022, 2023). However, as illustrated in Table 1, a single utterance may simultaneously sound natural to speakers from multiple countries, making it inherently multi-dialectal and difficult for traditional single-label classification methods to produce accurate results (Keleg and Magdy, 2023; Olsen et al., 2023). Even moderately long sentences can be acceptable in multiple dialects (Keleg et al., 2025).

To better capture this phenomenon, the field has begun transitioning to *Multi-Label Arabic Dialect Identification* (MLADI), which allows an utterance to be tagged as acceptable in multiple dialects.¹ This transition has been driven by the evolution of the NADI shared tasks, a recurring benchmark se-

¹Multi-label Dialect Identification is being explored for other languages such as French and Spanish (Bernier-colborne et al., 2023; Zampieri et al., 2024; Lopetegui et al., 2025).

ries for Arabic Dialect Identification. While earlier editions of NADI (2020–2023) framed ADI as a single-label classification problem, assigning each sentence to a single country-level dialect (Abdul-Mageed et al., 2020, 2021b, 2023), the NADI2024 shared task introduced multi-label annotations to explicitly account for dialectal overlap (Abdul-Mageed et al., 2024). However, in NADI2024, multi-label annotations are provided only for the development and test sets, while the training data remains single-labeled, creating a mismatch between the nature of the task and the structure of the training data. As a result, MLADI introduces a weakly-supervised classification problem, where models must learn to predict multiple dialects based on datasets that provide only one ground-truth label per instance.

In this work, we examine the mismatch between the multi-label nature of Arabic dialect usage and the single-label structure of existing ADI datasets by analyzing the behavior of binary dialect-specific acceptability classifiers trained on single-label NADI datasets (Abdul-Mageed et al., 2020, 2021b, 2023). Using the classifiers’ training dynamics (Swayamdipta et al., 2020), we show that a substantial portion of samples treated as negative supervision is in fact judged acceptable by native speakers. This observation highlights the difficulty of accurately selecting negative samples when repurposing single-label, geo-located datasets for multi-label dialect identification and motivates explicit multi-label supervision.

Building on these insights, we construct a multi-label training set by aggregating pseudo-labels from two heterogeneous sources: GPT-4o and a set of 18 binary dialect acceptability classifiers. Using this dataset, we train a BERT-based multi-label classifier, which we refer to as LAHJATBERT. We train it under three settings: without curriculum learning, with an ALDi-aware curriculum, and with a label-cardinality-based curriculum (Bengio et al., 2009). The two curriculum strategies progressively expose the model to sentences of increasing dialectal ambiguity, allowing it to learn from simpler instances before confronting harder ones.

Our contributions are threefold:

1. We provide an in-depth analysis of the limitations of reusing single-label ADI datasets for multi-label dialect acceptability.
2. We construct a pseudo-labeled multi-label dataset for MLADI by combining the predic-

tions of GPT-4o and binary dialect classifiers.²

3. We introduce LAHJATBERT, a family of multi-label BERT-based models trained on the constructed dataset, and investigate curriculum-based training variants aligned with the multi-label structure of MLADI. The best-performing variant achieves 69.04% macro-F1 on the MLADI leaderboard (Keleg et al., 2025), surpassing the top NADI2024 system and outperforming larger multilingual and Arabic-specific language models.

2 MLADI Task’s Setup and Previous Attempts

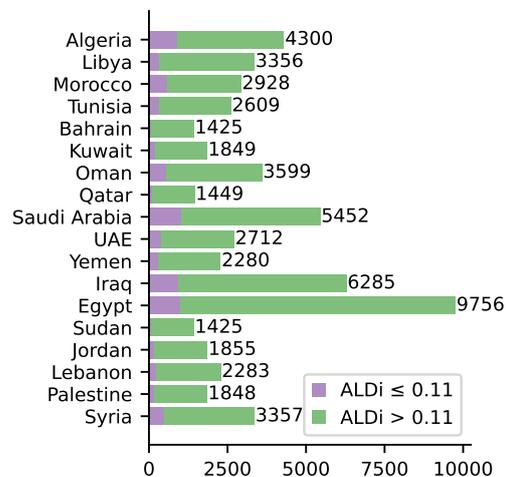


Figure 1: Number of samples in each dialect after combining the NADI 2020, 2021, 2023 datasets. Samples with automatically estimated *Arabic Level of Dialectness* (ALDi; Keleg et al., 2023) ≤ 0.11 are expected to be in MSA. The majority of the MSA samples are expected to be acceptable in all dialects.

The MLADI dataset only provides a development set of 120 samples and a test set of 1000 samples. Each sample of the development set is manually labeled by annotators from 8 countries, while the test set has 11 country-level acceptability labels. To build multi-label ADI systems, the task requires using the following three single-labeled NADI datasets: NADI 2020 (Abdul-Mageed et al., 2020), NADI 2021 (Abdul-Mageed et al., 2021b), and NADI 2023 (Abdul-Mageed et al., 2023). All these datasets provide tweets annotated with their estimated geo-location country label covering 18 countries: Algeria, Bahrain, Egypt, Iraq, Jordan,

²Following the NADI shared-task license, we release only tweet IDs and derived labels, not the underlying tweet text.

Kuwait, Lebanon, Libya, Morocco, Oman, Palestine, Qatar, Saudi Arabia, Sudan, Syria, Tunisia, the UAE, and Yemen. Since these datasets rely on the user geo-location rather than manual linguistic annotation, posts authored by users whose geo-location differs from their country of origin can be mislabeled (Abdul-Mageed et al., 2024). NADI 2023 contains 1,000 tweets for each of 18 Arabic country-level dialects. By contrast, NADI 2020 and NADI 2021 are more imbalanced, with dialects such as Bahraini and Qatari underrepresented relative to more frequent varieties like Egyptian and Iraqi. The combined distribution of samples across dialects is shown in Figure 1.

Previous MLADI Attempts. Kanjirangat et al. (2024) used a nearest-neighbour approach to predict multiple labels for each sample, by encoding the training data samples and the test set into the same embedding space. Karoui et al. (2024) tackled the weak supervision problem by applying a similarity-based label expansion strategy. Their method, SIMMT, heuristically assigns additional labels to each sample based on vocabulary similarity between dialects, followed by multi-label fine-tuning of transformer models. This approach demonstrated that moving from single-label to multi-label supervision can improve performance on the MLADI task. Our work replaces heuristic label expansion with a pseudo-labeling framework that integrates complementary signals from multiple models to produce multi-label annotations. This yields a richer supervision signal than surface-level similarity measures.

3 Difficulties of Using Existing Datasets for Dialect Acceptability Classification

An intuitive idea to build multi-label ADI systems is to repurpose single-label ADI datasets for training multiple independent binary classifiers, each of which assesses the acceptability of a sentence in a specific dialect. For a specific country-level dialect (*Cntry*), one might take samples geolocated to *Cntry* as **positive** (i.e., *acceptable*) samples for the country’s classifier, and samples geolocated to other countries as **negatives**. Previous attempts have shown that this technique does not achieve the best classification performance (Karoui et al., 2024; Kanjirangat et al., 2024). However, they have not analyzed the reasons for the failure of this technique, which we investigate here.

3.1 Other Countries’ Samples Are Not Always Negative Samples

Definition For a dialect acceptability model, the negative class represents *sentences with any linguistic feature* (e.g., a morpheme or a lexical item) that is **not acceptable** in the considered dialect. The positive class represents the remaining sentences acceptable in this dialect or in MSA.

The majority of the samples geo-located to the considered dialect are expected to be positive samples. The fact that some of these samples could also be acceptable in other dialects does not impact their categorization as positive samples, according to the aforementioned definition. To the contrary, considering samples geo-located to other dialects as negative samples is problematic. More specifically, a subset of these samples is expected to also be acceptable in the dialect considered. Consequently, this subset of negative sentences should be reassigned to the positive class. We next show how model training dynamics could guide the identification of wrongly assigned negative samples.³

3.2 Training Dynamics and Multi-label Samples

Swayamdipta et al. (2020) used three metrics tracked during the model training process to categorize each sample’s difficulty to be learned. For a training sample x_i with a label y_i , the confidence that the model assigns to the target y_i is tracked at the end of each training epoch (or every number ‘ N ’ of training batches). The mean and standard deviation of the tracked confidence scores assigned to the sample’s target label are termed the **Confidence** and **Variability** metrics, respectively. **Correctness** represents the percentage of epochs (or N batches) for which the model assigns a higher probability to the sample’s target label than all the other labels. Based on these metrics, the training dataset is split into three main categories—(1) *easy to learn* samples: high-**Confidence** and low-**Variability**, (2) *ambiguous* samples: moderate-**Confidence** and high-**Variability**, and (3) *hard to learn*: low-**Confidence** and low-**Variability**.

Intuition A dialect acceptability classifier would struggle to learn negative samples that are also acceptable in the considered dialect, as these samples

³Abdul-Mageed et al. (2024) found that NADI’s geo-location method has a moderate to high accuracy of ensuring a sentence’s acceptability in a specific country-level dialect.

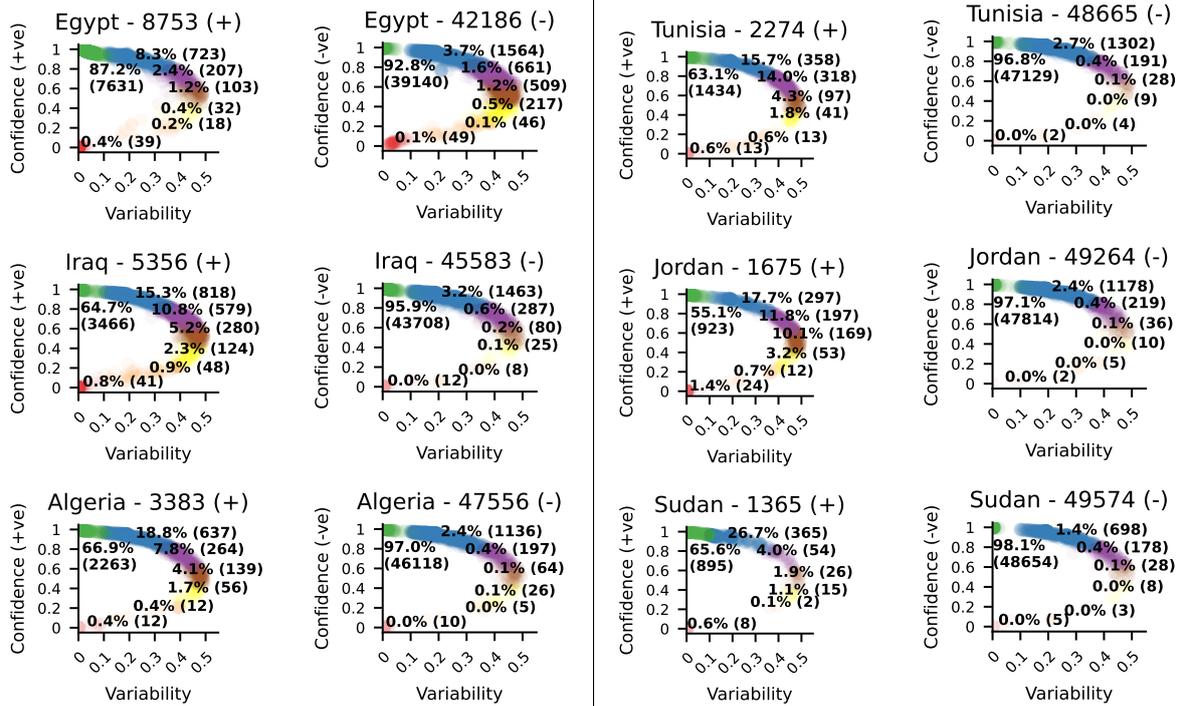


Figure 2: The training dynamics for 6 binary acceptability classifiers, characterized by the mean confidence in the label across different steps/stages of the model training (y-axis), and the standard deviation of these confidence values (x-axis). Each pair shows the training dynamics’ metrics for the non-MSA positive (left) and negative (right) samples of a single classifier, with the respective number of samples shown above each subplot. **Note:** Sample’s correctness ranges are ■: 0 ■: [0, 0.2[■: [0.2, 0.4[■: [0.4, 0.6[■: [0.6, 0.8[■: [0.8, 1[■: 1

should belong to the positive class. These samples are expected to be *ambiguous* or even *hard to learn*.

Methodology We train 18 dialect acceptability classifiers, one for each country represented in the NADI datasets. For each country’s classifier, samples of the NADI datasets geo-located to this country, and MSA samples (ones with ALDi < 0.11) are considered as positive (acceptable) samples, with the remaining samples considered as negative (unacceptable) samples. Confidence scores are tracked every 300 steps ($\approx \frac{1}{5}$ th of an epoch) for 5 epochs. However, the first epoch’s confidence scores are ignored, as the model’s training dynamics could be unstable during the early learning stages (Swayamdipta et al., 2020).

Findings Figure 2 shows the training dynamics for six different acceptability classifiers, for the positive (non-MSA) samples and the negative samples. First, a few samples (<50) for each set have a correctness value of 0. Moreover, the larger the imbalance between the number of positive and negative samples becomes, the smaller the percentage of negative samples with non-perfect correctness is.

For instance, 7.2% (n=3,046) of Egypt’s negative samples have non-perfect correctness scores compared to only 1.9% (n=920) for Sudan. These negative samples with non-perfect correctness scores are potentially wrongly assigned samples.

3.3 Usability of Training Dynamics in Flagging Wrongly Assigned Samples

To understand the effectiveness of training dynamics in identifying wrongly assigned samples, we engaged native speakers to manually evaluate the acceptability of sentences with varying correctness scores. Specifically, we recruited one annotator from the following countries: Egypt (EG), Iraq (IQ), Algeria (DZ), Tunisia (TN), Jordan (JO), and Sudan (SD).⁴

For each country, we sample 10 examples from each of the seven correctness ranges: 0, [0, 0.2[, [0.2, 0.4[, [0.4, 0.6[, [0.6, 0.8[, [0.8, 1[, 1. In total, 140 samples are annotated for each dialect (70 of which are positive and the remaining 70 are negative), unless one of the bins had fewer

⁴Notably, the six recruited annotators previously participated in the annotation of NADI 2024’s evaluation sets.

than 10 samples. The annotators follow the same guidelines of NADI 2024’s first subtask (Abdul-Mageed et al., 2024), whereby they answer: *Is it possible that the tweet is authored by someone who speaks one of your country’s dialects? Options: Yes, Not Sure/Maybe, or No.*

Results Inspecting the number of acceptable samples in each bin in Table 2, a large number of negative samples with *correctness* < 1 are rated as acceptable, indicating the effectiveness of using non-perfect correctness scores to flag wrongly assigned negative samples. Conversely, a majority of the positive samples are rated as acceptable for the different correctness bins, even for correctness scores that are less than 0.6, which could be attributed to the class imbalance toward negative samples.

Label	Correctness							
	0	0,0.2]	0.2,0.4]	0.4,0.6]	0.6,0.8]	0.8,1]	1	
EG	+ve	2	6	6	6	7	9	10
	-ve	10	10	10	9	8	4	2
IQ	+ve	4	2	1	6	4	5	7
	-ve	9	6/8	10	9	7	5	2
DZ	+ve	3	4	5	5	8	7	9
	-ve	9	5/5	8	9	9	8	4
TN	+ve	1	3	1	0	1	5	7
	-ve	2/2	4/4	8/9	7	5	4	1
JO	+ve	3	7	6	5	8	10	9
	-ve	2/2	5/5	10	10	9	9	2
SD	+ve	5/8	1/2	6	7	10	8	10
	-ve	5/5	3/3	8/8	7	8	10	1

Table 2: The number of acceptable samples for the different correctness bins. For each row, each bin contains 10 total samples, except for a few bins marked by the *acceptable/total* format. **Note:** Values marked in blue indicate that >50% of the bin’s samples are acceptable, while values marked in red indicate that <50% of the bin’s samples are acceptable. The high-correctness bins of the positive class are expected to have a large number of acceptable samples. Conversely, the high-correctness bins of the negative class are expected to have a small number of acceptable samples, which is not the case. Country codes follow the ISO 3166-1 alpha-2 standard.

Moving Forward Our analysis shows that a large proportion of negative samples with non-perfect correctness scores should be reassigned to the positive class. However, manual annotation is still required to assess these automatically flagged samples. Moreover, the class imbalance seems to have an impact on the model’s dynamics. Future work could consider using a *human and model in the*

loop setup (Vidgen et al., 2021). More specifically, the initial model’s training dynamics are used to automatically identify ambiguous samples, which are then manually reassigned to the correct class. Afterward, another model is trained from scratch on the dataset after reassignment, with the new model’s dynamics used to automatically identify new ambiguous samples.

4 Multi-Label ADI Dataset Creation

It seems inevitable that building multi-label ADI models requires the presence of multi-label ADI training datasets, especially to reduce the number of false negatives for each dialect (i.e., samples wrongly assumed unacceptable in that dialect). However, building large enough multi-label ADI datasets is expensive, since annotating just 1,120 samples by speakers of 9 countries could cost as much as \$1,700 (Abdul-Mageed et al., 2024). Hence, we propose two pseudo-labeling methods for building multi-label ADI datasets:

(1) Binary Dialect Classifiers. We again build 18 independent acceptability classifiers, one per country-level dialect *dia*, each predicting whether a sentence *x* is acceptable in dialect *dia*. These classifiers are trained only on the balanced NADI 2023 dataset to avoid skewed pseudo-labels. For a sentence *x_i*, we generate 18 acceptability pseudo-labels using the 18 classifiers as:

$$\hat{y}_{x_i}^{\text{BIN}} = (\text{Accept}_{dia_1}(x_i), \dots, \text{Accept}_{dia_{18}}(x_i)).$$

Here, we notably rely on the Arabic Level of Dialectness (ALDi; Keleg et al., 2023) score as a global signal characterizing the degree of dialectness of each sentence ($a_i = \text{ALDi}(x_i) \in [0, 1]$). Keleg et al. (2025) showed that ALDi moderately correlates with the number of dialects in which a sentence is acceptable. The higher the ALDi score of a sentence, the less the number of dialects in which the sentence is acceptable.

For the **positive** samples of *dia*’s acceptability classifier $\text{Accept}_{dia}(x_i)$, we consider all sentences from the NADI datasets that are geolocated to *dia*, in addition to MSA sentences with $a_i < 0.11$, which are broadly acceptable across dialects, irrespective of their geolocation. Since treating all sentences geolocated to other regions as negative examples leads to systematic errors (as shown in Section 3), we prioritize the precision of selecting true **negative samples**. To this end, we only select sentences that are (1) geolocated to *non-neighbouring* dialect regions, (2) with high dialect-

Pseudo-Labeling Method	P_{macro}	R_{macro}	$F1_{\text{macro}}$	Acc.
(1) Binary Dialect Classifiers	78.2	51.4	60.4	76.2
(2) GPT-4o Pseudo-Labels	73.5	66.3	67.8	77.4
Hybrid ALDi-Based Labels (§4.2)	77.5	62.8	68.5	79.3

Table 3: Macro-averaged Precision, Recall, F1, and Accuracy of the two methods in addition to the aggregation method (§4.2) on the NADI2024 development set.

ness ($a_i > 0.77$), where linguistic overlap with dialect *dia* is less likely.⁵

(2) GPT-Based Multi-Label Annotation. For each sentence, we obtain multi-label annotations from GPT-4o by prompting the model to independently assess the acceptability of the sentence in each of the 18 dialects. This yields a binary relevance vector $\hat{y}_i^{\text{GPT}} \in \{0, 1\}^{18}$. The full prompting template is provided in Appendix E.

4.1 Quality of Supervision Signals

To evaluate the two labeling methods, we use the development set of the NADI2024 shared task (Abdul-Mageed et al., 2024), which has 120 sentences annotated for 8 country-level dialects. This provides a useful benchmark for evaluating the performance of the two methods, as shown in Table 3.

Our conservative way of selecting the samples for the binary classifiers results in a high macro precision of 78.2 and a moderate macro recall of 51.4. In contrast, GPT-4o achieves substantially higher recall and a better overall F1 score. This hints that GPT-4o is more willing to assign multiple dialect labels, particularly in cases where dialectal overlap is present.

4.2 Pseudo-Labels Aggregation

To better understand the results in Table 3, we analyze performance across ALDi ranges in Table 4. For MSA samples ($[0, 0.11)$), both GPT-4o and the Binary Classifiers achieve perfect precision. For highly dialectal samples ($[0.77, 1]$), the Binary Classifiers attain higher precision, indicating more reliable negative supervision when strong dialectal cues are present.

In contrast, GPT-4o consistently achieves higher recall and F1 in the intermediate ALDi ranges ($[0.11, 0.77]$), where dialectal overlap is more common. This behavior is consistent with the conservative construction of the Binary Classifiers, whose

⁵See Appendices B, C and D for more implementation details.

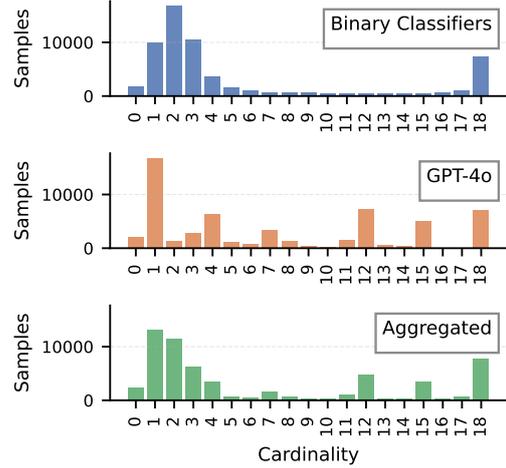


Figure 3: Number of samples for each label cardinality according to the three pseudo-labeling methods.

negative samples are restricted to highly dialectal cases ($a_i > 0.77$), improving precision at the extremes but limiting coverage in the mid-range.

ALDi Bin	Supervision Source	P_{macro}	R_{macro}	$F1_{\text{macro}}$
$[0, 0.11)$ (n=7)	(1) Binary Classifiers	100.00	94.64	97.12
	(2) GPT-4o	100.00	90.48	94.84
$[0.11, 0.44)$ (n=16)	(1) Binary Classifiers	76.99	46.81	55.99
	(2) GPT-4o	84.72	61.36	69.72
$[0.44, 0.77)$ (n=48)	(1) Binary Classifiers	63.28	36.72	44.38
	(2) GPT-4o	63.86	56.94	57.15
$[0.77, 1.0]$ (n=49)	(1) Binary Classifiers	77.60	51.38	59.03
	(2) GPT-4o	62.36	68.14	62.62

Table 4: Performance of the two pseudo-labeling methods on subsets of the NADI2024 development set.

Consequently, we aggregate the predictions of the two pseudo-labeling methods by using the Binary Classifiers at the ALDi extremes ($a_i < 0.11$ or $a_i > 0.77$) and GPT-4o in the intermediate range ($0.11 \leq a_i \leq 0.77$). This aggregation combines the strengths of both methods and achieves the best macro F1 and accuracy on the evaluation set (last row of Table 3), reflecting an improved precision-recall trade-off.

5 Multi-Label Dialect Classification Models

Using the pseudo-labeled multi-label ADI dataset described in Section 4, we fine-tune MARBERT (Abdul-Mageed et al., 2021a) as a multi-label classifier over the 18 Arabic dialects (country-level ADI). We refer to the resulting family of models fine-tuned under different settings, as LAHJAT-

BERT. We train using binary cross-entropy with logits loss, treating each label independently. During fine-tuning, we freeze the bottom 8 transformer layers of MARBERT and update only the top 4 layers and the classification head. We do not fully fine-tune all MARBERT layers since fine-tuning only a fraction of the final layers recovers most downstream effectiveness (Lee et al., 2019). Also, we discard zero-cardinality samples (i.e., instances for which the pseudo-labeling step assigns no dialect), since keeping them would treat the samples as negative for all 18 dialects and inject systematic noise. We report the complete training and inference hyperparameters in Appendix A.

An analysis of the dataset (Figure 3) reveals a strong skew in the cardinality distribution: most samples contain only one or two active dialects. This bias might encourage the model to predict low-cardinality outputs, substantially reducing recall for sentences acceptable in multiple dialects. To mitigate the dataset’s inherent low-cardinality bias, we adopt two strategies for curriculum learning (CL). In both strategies, the complexity of the training examples is increased in a controlled manner during the model training process, to achieve better model generalization (Bengio et al., 2009). The exact ordering of both the cardinality buckets and the ALDi buckets is motivated in Appendix F.

5.1 Cardinality-Based Curriculum Learning

In this strategy, the samples’ label cardinality is used as a proxy for their respective complexities. For a sentence x_i with a pseudo-labeled target vector $\mathbf{y}_i \in \{0, 1\}^{18}$, the label cardinality $c(x_i) = \|\mathbf{y}_i\|_0$ represents the number of dialects in which the sentence is acceptable. This strategy is proposed to mitigate the skewness of the training dataset samples toward lower cardinalities.

We partition the training set into cardinality buckets B_c , where B_c contains samples with label cardinality c . We then define a curriculum ordering over these buckets using the mean loss of each cardinality. Let $\pi(e)$ denote the cardinality index of the bucket selected at curriculum stage e . At stage e , the model is trained on (1) all examples from $B_{\pi(e)}$, and (2) an equal number of randomly sampled examples from each previously introduced bucket $B_{\pi(1)}, \dots, B_{\pi(e-1)}$.

The model first trains exclusively on the bucket introduced at the first curriculum stage. In the next stage, the bucket selected next is introduced

together with an equal number of samples drawn from each previously introduced bucket. This progression continues until the full cardinality range in the dataset is incorporated. The overall training schedule is illustrated in Figure 4, which visualizes how buckets are gradually introduced while maintaining balanced exposure across earlier stages.

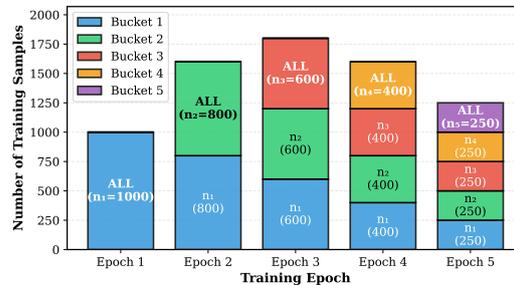


Figure 4: Illustration of the cardinality-based curriculum schedule, showing the progressive introduction of higher difficulty cardinality samples. The numerical values are illustrative and do not reflect the actual dataset.

5.2 ALDi-Based Curriculum Learning

ALDi provides another intuitive proxy for the samples’ complexities. Samples with intermediate ALDi scores are expected to be harder than MSA samples with low scores and high-score samples that show clear cues of a specific dialect.

We divide the continuous ALDi range into four contiguous intervals: $I_1 = [0, 0.11)$, $I_2 = [0.11, 0.44)$, $I_3 = [0.44, 0.77)$, $I_4 = [0.77, 1]$. For each interval I_k , we construct a bucket B_k that contains all training examples whose ALDi score falls in that range: $B_k = \{x_i \mid a(x_i) \in I_k\}$; where $a(x_i) \in [0, 1]$ denotes sentence x_i ’s ALDi score.

We define a curriculum ordering over the ALDi buckets based on average training loss. Let $\pi(e)$ denote the index of the ALDi bucket selected at epoch e , with buckets ordered from lowest to highest loss. At epoch e , the model is trained on (1) all examples from the current bucket $B_{\pi(e)}$, and (2) a random subset of examples from each previously introduced bucket $B_{\pi(1)}, \dots, B_{\pi(e-1)}$, following the same sampling strategy as the cardinality-based CL.

6 Results

We evaluate our multi-label dataset construction and training strategies on the NADI2024 development set, which provides multi-label annotations for **8 country-level dialects**. Since our training

and supervision span **18 dialects**, we report metrics only on the overlapping label set.

6.1 Multi-Label Model Performance

We evaluate how a single multi-label classifier can learn from different supervision signals by training the same model architecture on each dataset without curriculum learning. This setup isolates the effect of the supervision labels themselves.

As shown in Table 5 (Block II), the model trained on GPT-4o labels achieves higher macro Recall but lower Precision, indicating a tendency to over-predict dialect labels. In contrast, training on the hybrid labels yields higher Precision and Accuracy while maintaining comparable macro F1. This behavior is consistent with the design of the hybrid labels, which combines the conservative behavior of the binary classifiers at ALDi extremes with the richer GPT-based labels in the ALDi mid-range.

Model	P_{macro}	R_{macro}	$F1_{\text{macro}}$	Acc.
(I) Baseline				
NADI 2024 Baseline	71.2	30.9	39.7	69.3
(II) Multi-Label Model Performance				
Trained on GPT-4o labels	69.6	65.7	66.2	75.5
Trained on hybrid labels	73.7	63.7	67.4	77.5
(III) Curriculum Learning on Hybrid Dataset				
Hybrid + Cardinality-Based CL	69.0	80.6	72.7	77.5
Hybrid + ALDi-Based CL	71.4	71.3	70.3	78.2

Table 5: Macro-averaged Precision, Recall, F1, and Accuracy on the NADI2024 development set. Results are grouped into (i) multi-label model performance without CL, and (ii) the effect of CL.

6.2 Impact of Curriculum Learning

We next examine the effect of curriculum learning by comparing curriculum-based training with the baseline setting without a curriculum. As shown in Table 5 (Block III), both curriculum learning strategies improve macro F1 relative to training without curriculum learning, while consistently increasing recall and reducing precision.

In our setup, the curriculum order is derived from a loss-based criterion (Appendix §F), where examples introduced at the latest stages tend to be more ambiguous, corresponding to intermediate label cardinalities and intermediate ALDi scores.

One possible interpretation of the observed recall increase is that exposure at the latest curriculum stages to higher-loss, more ambiguous examples encourages the model to activate a larger set of

labels per instance. These examples are characterized by intermediate ALDi scores and multiple valid dialect labels per instance. This broader label coverage may help recover more relevant dialect labels, leading to higher recall. At the same time, predicting more labels per instance can make the model less conservative, which is reflected in the accompanying reduction in precision.

This interaction also helps explain why curriculum learning is most effective when applied to the hybrid supervision. Since the hybrid annotations exhibit higher precision than GPT-4o labels alone, they help limit the precision loss associated with increased recall.

6.3 Generalization on the MLADI Test Set

We evaluate whether the improvements on the development set generalize to the MLADI test set. Table 6 reports the performance of the three LAHJATBERT variants compared with the NADI 2024 baseline and previously reported systems.

The NADI 2024 baseline is a single-label dialect identification model, but it is converted into a multi-label predictor at test time. For each sentence, it computes a softmax distribution over the 18 dialect classes, then selects the most likely dialects until their cumulative probability reaches a fixed value (we use the Top-90% setting, i.e., $P=0.9$), and returns those dialects as the predicted label set (Abdul-Mageed et al., 2024).

All LAHJATBERT variants outperform the shared-task baseline and prior approaches in terms of macro F1, indicating that the gains obtained from the constructed supervision and training strategies extend beyond the development setting. Among the LAHJATBERT variants, the ALDi-based curriculum achieves the highest macro F1, while the cardinality-based curriculum yields the highest recall, mirroring the precision-recall trade-offs observed on the development set.⁶

Model	P_{macro}	R_{macro}	$F1_{\text{macro}}$	Acc.
LAHJATBERT (no curriculum)	69.0	69.7	68.0	79.1
LAHJATBERT + Cardinality CL	59.3	81.0	66.6	73.9
LAHJATBERT + ALDi CL	65.0	76.4	69.0	78.2
Aya-32B (Dang et al., 2024)	49.5	64.5	54.5	65.6
Elyadata (Karoui et al., 2024)	50.2	56.9	52.4	67.0
NADI 2024 Baseline	64.8	39.9	47.0	72.3

Table 6: Macro-averaged Precision, Recall, F1, and Accuracy on the MLADI test set (Keleg et al., 2025), which contains 1,000 sentences annotated for 11 dialects.

⁶We analyze the models’ predictions in the Appendix (§G).

7 Conclusion and Future Work

This work examines the limitations of reusing single-label Arabic dialect identification (ADI) datasets for multi-label ADI (MLADI). We specifically highlight that the difficulty of building binary dialect acceptability classifiers lies in the selection of negative samples. While training dynamics can help in automatically flagging wrongly-assigned negative samples, manual verification is still required to assess the acceptability of these flagged samples. Using single-label ADI data, we construct a pseudo-labeled multi-label dataset by aggregating predictions from GPT-4o and binary dialect acceptability classifiers, and introduce LAHJATBERT, a family of BERT-based multi-label models that outperform existing MLADI systems.

Future work could incorporate targeted human annotation for samples identified as ambiguous by the analysis, enabling iterative refinement of multi-label supervision. Finally, our CL approach substantially improves recall. Future work could examine whether similar gains extend to other multi-label tasks beyond dialect identification.

Limitations

Our evaluation is limited by the label coverage of available benchmarks. Although our training and supervision signals span 18 dialects, the NADI2024 development set provides multi-label annotations for only 8 country-level dialects, so development metrics are computed only on the overlapping label subset. Likewise, the test set covers only 11 dialects, restricting conclusions about generalization to the full 18-dialect label space.

A second limitation originates from the fact that NADI relies on user geo-location as a proxy for dialect rather than direct linguistic annotation, which can be noisy when posting location and actual dialect do not align.

Ethics and Broader Impact

This work uses anonymized Arabic tweets from the publicly released NADI shared-task datasets, collected via the Twitter API under the platform’s terms. Following NADI licensing restrictions, we release only tweet IDs and our automatically generated multi-label annotations (not the underlying tweet text), along with code and trained models to enable reproducibility. The research is methodological, and while dialect identification can support

beneficial language technologies and sociolinguistic analysis, it also carries risks of misuse (e.g., profiling or surveillance), which we explicitly do not endorse and is outside the scope of this work.

References

- Ahmed Abdelali, Hamdy Mubarak, Younes Samih, Sabit Hassan, and Kareem Darwish. 2021. [QADI: Arabic dialect identification in the wild](#). In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 1–10, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.
- Muhammad Abdul-Mageed, AbdelRahim Elmadany, Chiyu Zhang, El Moatez Billah Nagoudi, Houda Bouamor, and Nizar Habash. 2023. [NADI 2023: The fourth nuanced Arabic dialect identification shared task](#). In *Proceedings of ArabicNLP 2023*, pages 600–613, Singapore (Hybrid). Association for Computational Linguistics.
- Muhammad Abdul-Mageed, AbdelRahim Elmadany, et al. 2021a. [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.
- Muhammad Abdul-Mageed, Amr Keleg, AbdelRahim Elmadany, Chiyu Zhang, Injy Hamed, Walid Magdy, Houda Bouamor, and Nizar Habash. 2024. [NADI 2024: The fifth nuanced Arabic dialect identification shared task](#). In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 709–728, Bangkok, Thailand. Association for Computational Linguistics.
- Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020. [NADI 2020: The first nuanced Arabic dialect identification shared task](#). In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 97–110, Barcelona, Spain (Online). Association for Computational Linguistics.
- Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2021b. [NADI 2021: The second nuanced Arabic dialect identification shared task](#). In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 244–259, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.
- Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2022. [NADI 2022: The third nuanced Arabic dialect identification shared task](#). In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 85–97, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

- Maha J. Althobaiti. 2020. [Automatic arabic dialect identification systems for written texts: A survey](#).
- Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2022. [Hierarchical aggregation of dialectal data for Arabic dialect identification](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4586–4596, Marseille, France. European Language Resources Association.
- Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. [Curriculum learning](#). In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09*, page 41–48, New York, NY, USA. Association for Computing Machinery.
- Gabriel Bernier-colborne, Cyril Goutte, and Serge Leger. 2023. [Dialect and variant identification as a multi-label classification task: A proposal based on near-duplicate analysis](#). In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 142–151, Dubrovnik, Croatia. Association for Computational Linguistics.
- Houda Bouamor, Sabit Hassan, and Nizar Habash. 2019. [The MADAR shared task on Arabic fine-grained dialect identification](#). In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 199–207, Florence, Italy. Association for Computational Linguistics.
- John Dang, Shivalika Singh, Daniel D’souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. [Aya expand: Combining research breakthroughs for a new multilingual frontier](#).
- Nizar Y Habash. 2010. *Introduction to Arabic Natural Language Processing*. Morgan & Claypool Publishers.
- Vani Kanjirangat, Tanja Samardzic, Ljiljana Dolamic, and Fabio Rinaldi. 2024. [NLP_DI at NADI 2024 shared task: Multi-label Arabic dialect classifications with an unsupervised cross-encoder](#). In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 742–747, Bangkok, Thailand. Association for Computational Linguistics.
- Amira Karoui, Farah Gharbi, Rami Kammoun, Imen Laouirine, and Fethi Bougares. 2024. [ELYADATA at NADI 2024 shared task: Arabic dialect identification with similarity-induced mono-to-multi label transformation](#). In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 758–763, Bangkok, Thailand. Association for Computational Linguistics.
- Amr Keleg, Sharon Goldwater, and Walid Magdy. 2023. [ALDi: Quantifying the Arabic level of dialectness of text](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10597–10611, Singapore. Association for Computational Linguistics.
- Amr Keleg, Sharon Goldwater, and Walid Magdy. 2025. [Revisiting common assumptions about Arabic dialects in NLP](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3327, Vienna, Austria. Association for Computational Linguistics.
- Amr Keleg and Walid Magdy. 2023. [Arabic dialect identification under scrutiny: Limitations of single-label classification](#). In *Proceedings of ArabicNLP 2023*, pages 385–398, Singapore (Hybrid). Association for Computational Linguistics.
- Jaejun Lee, Raphael Tang, and Jimmy Lin. 2019. [What would elsa do? freezing layers during transformer fine-tuning](#).
- Javier A. Lopetegui, Arij Riabi, and Djamé Seddah. 2025. [Common ground, diverse roots: The difficulty of classifying common examples in Spanish varieties](#). In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 168–181, Abu Dhabi, UAE. Association for Computational Linguistics.
- Helene Olsen, Samia Touileb, and Erik Velldal. 2023. [Arabic dialect identification: An in-depth error analysis on the MADAR parallel corpus](#). In *Proceedings of ArabicNLP 2023*, pages 370–384, Singapore (Hybrid). Association for Computational Linguistics.
- Mohammad Salameh, Houda Bouamor, and Nizar Habash. 2018. [Fine-grained Arabic dialect identification](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1332–1344, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. [Dataset cartography: Mapping and diagnosing datasets with training dynamics](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.
- Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. [Learning from the worst: Dynamically generated datasets to improve online hate detection](#). 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 1667–1682, Online. Association for Computational Linguistics.

Omar F. Zaidan and Chris Callison-Burch. 2011. [The Arabic online commentary dataset: an annotated dataset of informal Arabic with high dialectal content](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 37–41, Portland, Oregon, USA. Association for Computational Linguistics.

Omar F. Zaidan and Chris Callison-Burch. 2014. [Arabic dialect identification](#). *Computational Linguistics*, 40(1):171–202.

Marcos Zampieri, Kai North, Tommi Jauhiainen, Mariano Felice, Neha Kumari, Nishant Nair, and Yash Mahesh Bangera. 2024. [Language variety identification with true labels](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10100–10109, Torino, Italia. ELRA and ICCL.

A Experimental Details

We provide the full training and inference hyperparameters for the multi-label setup described in [section 5](#). We fine-tune a BERT-based model for multi-label country-level Arabic dialect identification with 18 binary labels, where the classification head outputs one logit per country. Training minimizes binary cross-entropy with logits, which implements a combination of sigmoid and binary cross-entropy and treats each label independently. We split the data into 90% training and 10% validation with a fixed random seed of 42. During fine-tuning, we freeze the bottom 8 transformer layers of MARBERT and update only the top 4 layers and the classification head. We set both the hidden-state and attention dropout probabilities to 0.3. We train for 3 epochs with a batch size of 24 for both training and evaluation, evaluating once per epoch. The best checkpoint is selected based on validation micro F1. At inference time, we apply a sigmoid to the logits and use a threshold of 0.3 to obtain binary label assignments. We chose this value by maximizing validation micro F1 on the held-out split, the default threshold of 0.5 yields lower validation macro F1, so we use 0.3 for all reported results.

B ALDi Threshold Selection

We adopt two ALDi thresholds, $a_i = 0.11$ and $a_i = 0.77$, to distinguish sentences with minimal dialectal evidence from those with strong dialectal

evidence. These thresholds are grounded in the annotation scheme and definition of the Arabic Level of Dialectness (ALDi) score introduced by [Keleg et al. \(2023\)](#).

ALDi Annotation Scheme. Following the ALDi annotation protocol described by [Keleg et al. \(2023\)](#), each sentence is annotated by three native Arabic speakers. Annotators assign one of four ordinal labels reflecting the degree of dialectness: *MSA* (0), *Little* ($\frac{1}{3}$), *Mixed* ($\frac{2}{3}$), and *Most* (1). The ALDi score for a sentence is defined as the mean of the three annotations. Consequently, ALDi values lie in $[0, 1]$ and take discrete steps of $\frac{1}{3}$.

Derivation of the Thresholds. Under this definition, the smallest non-zero ALDi value is

$$\frac{0 + 0 + \frac{1}{3}}{3} = \frac{1}{9} \approx 0.11,$$

which corresponds to the weakest possible evidence of dialectal content, where two annotators label the sentence as *MSA* and one assigns a *Little* dialect label. Conversely, an ALDi score of

$$\frac{1 + 1 + \frac{1}{3}}{3} = \frac{7}{9} \approx 0.77$$

corresponds to strong dialectal evidence, where two annotators assign the highest dialectness label (*Most*) and the third assigns at least *Little*. These two values therefore mark natural boundaries between minimal dialectal signal ($a_i < 0.11$), strong dialectal signal ($a_i > 0.77$), and an intermediate range characterized by mixed or graded annotator judgments.

ALDi Score Distribution. ALDi scores are concentrated near the *MSA* ($a_i < 0.11$) and highly dialectal ($a_i > 0.77$) ranges, with fewer samples in the mid-range ($0.11 \leq a_i \leq 0.77$), as shown in [Figure B1](#).

C Non-Neighbouring Dialect Regions

When constructing negative samples for the binary dialect classifiers ([section 4](#)), we restrict negatives to sentences geolocated in non-neighbouring countries. Neighbouring countries are defined as those sharing a land border with the target country, and negative samples are drawn exclusively from non-bordering countries.

By restricting negative samples to non-bordering countries, we avoid cases where sentences may be linguistically compatible with the target dialect, resulting in more reliable negative supervision.

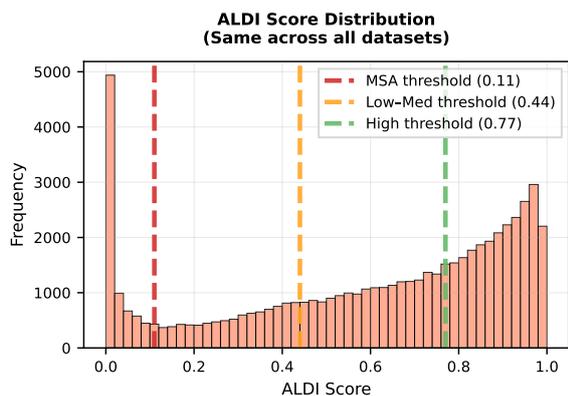


Figure B1: Distribution of ALDi scores in the dataset. Vertical dashed lines indicate the thresholds distinguishing MSA text ($a_i < 0.11$), low–medium dialectness ($0.11\text{--}0.44$), medium dialectness ($0.44\text{--}0.77$), and highly dialectal text ($a_i > 0.77$).

D Dataset Diagnostics

This appendix presents diagnostics of the constructed multi-label dataset, focusing on how label cardinality varies with the Arabic Level of Dialectness (ALDi) score for the supervision sources considered in this work.

D.1 Label Cardinality Across ALDi Ranges

Figure D2 shows the distribution of label cardinality across ALDi categories for each labeling method, as well as for the aggregated dataset.

Binary Dialect Classifiers. The binary classifiers exhibit a strong dependence on the ALDi score. For sentences with low ALDi values (MSA text), they frequently activate many dialect labels, resulting in high cardinality. For sentences with high ALDi values (strongly dialectal text), they tend to activate very few labels. In the intermediate ALDi ranges ($0.11 \leq a_i \leq 0.77$), where explicit negative supervision is absent, classifier outputs concentrate at either high or low cardinalities, and intermediate label counts are rarely observed.

GPT-4o. GPT-4o displays a smoother relationship between ALDi and label cardinality. While cardinality generally decreases as ALDi increases, GPT-4o assigns moderate numbers of labels more frequently in the Low–Med and Medium ALDi ranges. This pattern is consistent with graded representations of dialectal overlap and contrasts with the polarized outputs of the binary classifiers in the same ALDi regions.

Aggregation Effects. The aggregated dataset (right panel of Figure D2) combines these two behaviors. At low and high ALDi values, where binary classifier predictions are stable, the aggregated cardinality closely matches their outputs. In the intermediate ALDi ranges, where binary classifier predictions are concentrated at extreme cardinalities, the aggregation relies on GPT-4o, resulting in intermediate label counts and a smoother dependence of cardinality on ALDi.

D.2 Overall Label Cardinality Distribution

Figure 3 shows the overall distribution of label cardinality for the supervision sources used in this work, independent of ALDi conditioning.

E Prompt Template for GPT-Based Annotation

Instruction: You are a native Arabic speaker and highly qualified linguist with expert-level understanding of regional Arabic dialects. Given the sentence provided, evaluate its dialectal characteristics independently for each of the following dialects: Iraq, Egypt, Morocco, Libya, UAE, Saudi Arabia, Bahrain, Syria, Lebanon, Oman, Palestine, Algeria, Jordan, Tunisia, Kuwait, Yemen, Sudan, and Qatar. Return findings in JSON format:

```
{
  "Iraq": 0/1,
  "Egypt": 0/1,
  ...
  "Qatar": 0/1
}
```

Input sentence: {tweet}

Figure E3: GPT-based annotation prompt template.

F Curriculum Learning Ordering

To obtain a principled ordering for both the cardinality-based curriculum learning and the ALDi-based curriculum learning, we first train a baseline MARBERT model on the aggregated multi-label dataset without any curriculum learning. After fine-tuning, we freeze this baseline model and run inference on the full training set, recording the per-example loss.

We then aggregate these losses in two ways: (i) by label cardinality, computing the mean loss for each cardinality value, and (ii) by ALDi bin, also computing the mean loss but for each ALDi interval. The resulting mean-loss profiles are visualized

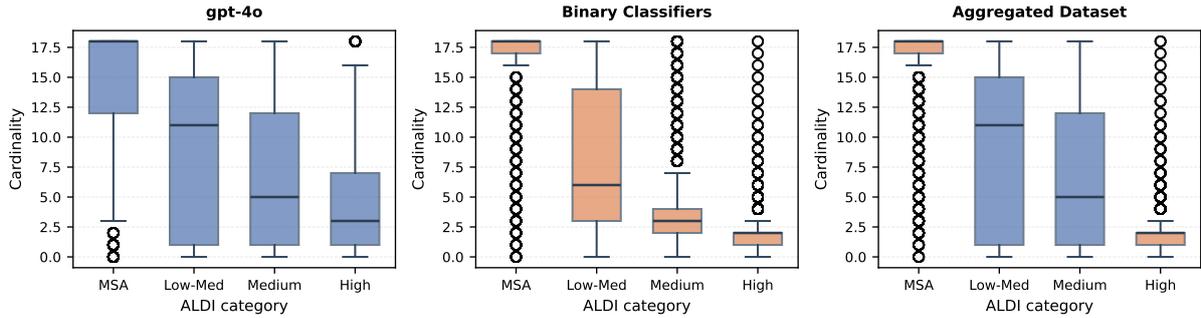


Figure D2: Cardinality across ALDi categories for the three labeling methods. Due to instabilities, we rely on GPT-4o for ambiguous cases and on the binary classifiers when the dialectness is clearly low (MSA) or high. This leads to more consistent and linguistically plausible multi-label patterns.

in Figure F4 (mean loss per cardinality) and Figure F5 (mean loss per ALDi bin).

We treat the mean loss of a bucket as a proxy for its difficulty, and order the stages from lower-loss (easier) buckets to higher-loss (harder) buckets. This difficulty-based ordering directly determines the progression of stages in both the cardinality-based and ALDi-based curriculum learning schedules described in section 5.

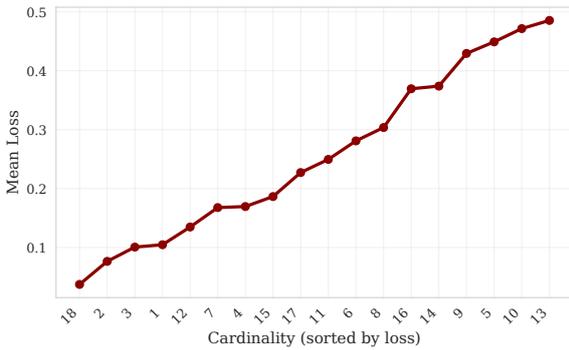


Figure F4: Sorted mean loss per cardinality, to measure the difficulty for the baseline model in predicting different cardinalities.

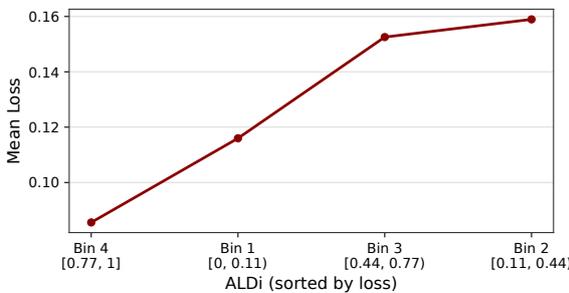


Figure F5: Sorted mean loss per ALDi bin, to measure the difficulty for the baseline model in predicting different cardinalities.

G Model’s Predictions on MADAR’s Samples

To analyze the behavior of the three newly introduced models, we study their predictions on 200 sentences from the MADAR-26 corpus (Salameh et al., 2018), for each of the 6 anchor dialects identified by the corpus creators. Additionally, we contrast their predictions to two baseline models: a single-label DI model **DID-Country**, and the multi-label DI **NADI2024 shared task’s baseline**.

Comparing our three LAHJATBERT models to the NADI2024 baseline, it is clear that the baseline is more susceptible to predicting the sentence to be acceptable in countries from different regions than our newly introduced model. For instance, out of the 200 sentences in Rabat’s dialect (spoken in Morocco), the baseline model unintuitively predicts that 190 of them are acceptable in Egypt, in comparison to less than 20 of them predicted to be acceptable by the three LAHJATBERT models. Hence, assuming that the labels within the top-p ($p=90\%$) of a single-label DI model as acceptable dialects for the input sentence is not an optimal strategy for multi-label dialect identification.

For the three LAHJATBERT models, the Cardinality-based one seems to frequently predict labels from other regions, which explains its lower precision yet higher recall than the two other models. The ALDi-based model seem to be achieving the best compromise between the precision and the recall, as indicated by the fact it achieves the highest overall macro-F1 score.

Dialect	DID-Country (Salameh et al., 2018)	NADI2024 Baseline (Abdul-Mageed et al., 2024)	LAHJATBERT	LAHJATBERT + ALDi CL	LAHJATBERT + Cardinality CL
BEI (200)	LB 129 , <u>SY 32</u> , <u>JO 15</u> , <u>PS 12</u> , EG 4, SA 3, IQ 2, OM 2, MA 1	LB 200 , <u>SY 200</u> , <u>JO 199</u> , <u>PS 199</u> , EG 180, IQ 174, TN 160, SD 127, BH 86, MA 78, KW 68, DZ 57, SA 56, QA 53, LY 41, OM 36, YE 25, AE 19	<u>SY 165</u> , LB 149 , <u>PS 131</u> , <u>JO 121</u> , IQ 36, SA 35, AE 34, KW 32, BH 31, OM 31, QA 29, YE 27, EG 23, LY 18, SD 17, TN 10, DZ 9, MA 9	<u>SY 174</u> , LB 146 , <u>PS 135</u> , <u>JO 129</u> , OM 35, SA 33, AE 33, IQ 31, KW 31, BH 30, QA 30, YE 26, EG 25, SD 19, LY 18, TN 9, DZ 8, MA 7	<u>SY 188</u> , LB 184 , <u>PS 181</u> , <u>JO 173</u> , SA 45, AE 44, BH 42, OM 42, KW 41, QA 41, YE 41, IQ 35, EG 31, SD 23, LY 22, MA 9, TN 9, DZ 7
CAI (200)	EG 147 , SA 12, SY 12, SD 9, JO 7, DZ 5, LY 4, TN 4, MA 3, YE 3, IQ 2, PS 2, QA 2, KW 1, LB 1, BH 1, OM 1, AE 1	EG 200 , YE 193, SA 192, SD 192, PS 188, JO 187, LY 182, LB 178, SY 167, MA 159, OM 142, TN 130, IQ 73, KW 54, QA 51, AE 40, DZ 20	EG 190 , <u>PS 131</u> , <u>SD 115</u> , LB 54, SA 54, JO 52, SY 52, LY 51, KW 46, IQ 44, BH 44, QA 44, AE 43, OM 42, YE 42, TN 14, MA 12, DZ 10	EG 188 , <u>PS 152</u> , <u>SD 132</u> , LY 49, SA 49, JO 46, LB 45, IQ 44, SY 44, KW 43, BH 41, OM 41, QA 41, AE 41, YE 41, TN 11, MA 9, DZ 8	PS 193, EG 189 , <u>SD 165</u> , SA 135, LB 103, JO 99, SY 93, QA 83, OM 81, AE 80, BH 79, YE 79, KW 74, LY 73, IQ 60, TN 14, MA 12, DZ 10
DOH (200)	QA 144 , SA 17, IQ 7, JO 5, SY 5, <u>OM 4</u> , MA 3, SD 3, TN 3, PS 3, LY 2, EG 2, YE 2	<u>SA 200</u> , QA 199 , <u>OM 198</u> , LY 194, JO 191, SY 181, IQ 163, SD 156, YE 137, <u>AE 122</u> , MA 112, <u>KW 86</u> , LB 80, TN 77, PS 76, <u>BH 64</u> , EG 59, DZ 14	<u>SA 173</u> , <u>KW 154</u> , <u>BH 142</u> , <u>AE 136</u> , QA 133 , <u>OM 128</u> , IQ 115, YE 111, PS 93, JO 91, SY 91, LB 89, SD 55, LY 54, EG 52, MA 24, TN 24, DZ 21	<u>SA 179</u> , <u>KW 165</u> , <u>BH 151</u> , QA 147 , <u>AE 142</u> , <u>OM 134</u> , YE 117, IQ 111, PS 105, LB 102, JO 100, SY 100, EG 53, SD 49, LY 46, MA 17, TN 17, DZ 14	<u>SA 197</u> , <u>BH 194</u> , QA 193 , <u>KW 190</u> , <u>AE 190</u> , <u>OM 175</u> , YE 149, IQ 132, PS 115, JO 111, LB 96, SY 96, SD 53, EG 50, LY 46, MA 14, TN 13, DZ 11
RAB (200)	MA 176 , <u>DZ 6</u> , <u>TN 4</u> , JO 3, LY 3, SA 3, PS 2, OM 1, SY 1, QA 1	MA 198 , <u>DZ 196</u> , LY 194, EG 190, QA 188, LB 180, AE 165, SA 159, SY 154, SD 151, <u>TN 151</u> , PS 147, KW 143, BH 125, IQ 70, JO 41, OM 32, YE 11	<u>DZ 172</u> , MA 171 , <u>TN 84</u> , <u>LY 27</u> , EG 16, PS 15, IQ 13, SY 13, JO 11, LB 11, SA 11, OM 10, AE 10, KW 9, BH 8, QA 8, YE 8, SD 7	<u>DZ 181</u> , MA 175 , <u>TN 68</u> , <u>LY 23</u> , PS 14, EG 10, IQ 8, SY 8, JO 7, SA 7, LB 6, AE 6, QA 6, KW 5, BH 5, OM 5, YE 4, SD 1	MA 178 , <u>DZ 174</u> , TN 130, <u>LY 52</u> , PS 32, SY 30, JO 29, LB 27, AE 25, KW 24, OM 24, SA 24, QA 24, YE 24, IQ 21, BH 20, EG 19, SD 16
TUN (200)	TN 167 , <u>DZ 6</u> , SA 5, JO 4, IQ 3, OM 3, PS 3, <u>LY 2</u> , <u>MA 2</u> , SY 2, QA 2, YE 1	<u>LY 198</u> , TN 197 , LB 192, IQ 191, <u>DZ 190</u> , EG 189, <u>MA 183</u> , OM 181, SD 180, QA 160, YE 150, KW 104, PS 99, AE 60, SY 42, SA 40, JO 28, BH 14	TN 159 , <u>LY 149</u> , <u>DZ 143</u> , <u>MA 94</u> , IQ 24, SY 24, PS 24, LB 23, JO 21, EG 19, KW 17, SA 17, AE 16, YE 15, BH 14, OM 14, SD 14, QA 14	TN 163 , <u>DZ 155</u> , <u>LY 153</u> , <u>MA 57</u> , PS 28, IQ 25, JO 23, LB 23, SY 23, EG 21, KW 20, SA 20, AE 20, OM 19, YE 18, BH 17, QA 16, SD 15	<u>LY 170</u> , TN 162 , <u>DZ 159</u> , MA 103, PS 48, SY 44, JO 42, LB 41, IQ 35, SA 33, KW 32, OM 31, AE 31, QA 30, YE 30, BH 29, EG 28, SD 21
MSA (200)	OM 162 , SA 162, SD 152 , <u>LY 151</u> , <u>SY 151</u> , DZ 150 , <u>IQ 149</u> , <u>JO 149</u> , EG 149, <u>PS 148</u> , KW 147 , <u>LB 147</u> , BH 147 , <u>MA 147</u> , QA 147 , <u>TN 147</u> , AE 147 , <u>YE 147</u>	OM 200 , SA 200, SD 200 , IQ 195 , <u>YE 189</u> , <u>JO 165</u> , AE 151 , <u>QA 145</u> , <u>LY 144</u> , EG 143 , <u>DZ 111</u> , <u>KW 100</u> , BH 91 , TN 75, MA 57 , SY 56 , <u>PS 30</u> , <u>LB 22</u>	IQ 196 , <u>JO 195</u> , LB 195 , PS 195 , <u>SA 194</u> , <u>SY 194</u> , AE 194 , <u>KW 193</u> , LY 193 , <u>EG 193</u> , OM 192 , <u>YE 192</u> , BH 191 , <u>SD 191</u> , QA 190 , <u>TN 188</u> , MA 187 , <u>DZ 183</u>	IQ 198 , <u>EG 198</u> , <u>LY 197</u> , PS 197 , <u>SA 197</u> , <u>JO 196</u> , KW 196 , <u>LB 196</u> , SD 196 , <u>SY 196</u> , <u>AE 196</u> , YE 196 , <u>BH 195</u> , OM 195 , <u>QA 195</u> , MA 189 , <u>TN 189</u> , DZ 188	IQ 200 , <u>JO 200</u> , KW 200 , <u>BH 200</u> , OM 200 , <u>PS 200</u> , QA 200 , <u>SA 200</u> , AE 200 , <u>YE 200</u> , LB 199 , <u>SY 199</u> , <u>LY 197</u> , SD 197 , <u>EG 196</u> , MA 190 , <u>TN 190</u> , DZ 188

Table G1: The number of times a country label is predicted by the model for the 200 sentences of the six anchor dialects of the MADAR CORPUS-26 corpus (Salameh et al., 2018): Beirut (BEI), Cairo (CAI), Doha (DOH), Rabat (RAB), Tunis (TUN), and MSA. Each dialect’s representative country is in **bold**, and the country’s same-region countries are underlined, following the regional grouping of Baimukan et al. (2022). **Note #1**: the DID-country model is a single-label DI model trained on the MADAR corpus. Its province-level predictions are mapped into country-level ones, with MSA predictions mapped to all the 18 country-level dialects we consider. **Note #2**: Country abbreviations - **AE**: UAE, **BH**: Bahrain, **DZ**: Algeria, **EG**: Egypt, **IQ**: Iraq, **JO**: Jordan, **KW**: Kuwait, **LB**: Lebanon, **LY**: Libya, **MA**: Morocco, **OM**: Oman, **PS**: Palestine, **QA**: Qatar, **SA**: Saudi Arabia, **SD**: Sudan, **SY**: Syria, **TN**: Tunisia, **YE**: Yemen.

OpenLID-v3: Improving the Precision of Closely Related Language Identification – An Experience Report

Mariia Fedorova Nikolay Arefyev Maja Buljan Jindřich Helcl
Stephan Oepen Egil Rønningstad Yves Scherrer

Language Technology Group
Department of Informatics
University of Oslo
mariiaf@ifi.uio.no

Abstract

Language identification (LID) is an essential step in building high-quality multilingual datasets from web data. Existing LID tools (such as OpenLID or GlotLID) often struggle to identify closely related languages and to distinguish valid natural language from noise, which contaminates language-specific subsets, especially for low-resource languages. In this work we extend the OpenLID classifier by adding more training data, merging problematic language variant clusters, and introducing a special label for marking noise. We call this extended system OpenLID-v3 and evaluate it against GlotLID on multiple benchmarks. During development, we focus on three groups of closely related languages (Bosnian, Croatian, and Serbian; Romance varieties of Northern Italy and Southern France; and Scandinavian languages) and contribute new evaluation datasets where existing ones are inadequate. We find that ensemble approaches improve precision but also substantially reduce coverage for low-resource languages.

1 Introduction

Growing interest in large-scale LLM pre-training data for languages other than English puts a spotlight on robust and broad-coverage language identification (LID). Common pre-training datasets are typically distilled from massive collections of web documents, which are characterized by immense diversity in, for example, genres and domains, degrees of (in)formality, juxtaposition of language and non-language content, presence of machine-generated content, code switching, and other sources of variation. For instance, the two largest and linguistically broadest pre-training datasets, FineWeb 2 (Penedo et al., 2025) and HPLT 3.0 (Oepen et al., 2025), apply LID as a document-level classification task. Specifically, FineWeb builds on the third-party GlotLID classifier (Kargaran et al., 2023), which supports

some 2,000 distinct languages, whereas HPLT has developed a custom classifier dubbed OpenLID (Burchell et al., 2023) covering around 200 languages; see Section 2 for further background. In this work, we seek to shed more light on LID performance and challenges in the realm of noisy web documents, with a particular emphasis on selected groups of closely related languages.

The contributions of this paper are as follows:

- We train a new version of the fully open-source OpenLID system for our experiments and publicly release it as OpenLID-v3.¹ This version covers 194 languages plus ‘not-a-language’ class.
- We evaluate OpenLID-v3 on mainstream LID benchmarks such as FLORES+ and UDHR and show that they are not sufficient for evaluating the quality of similar LID.² In addition, we employ several existing benchmarks for similar languages and create new ones for the BCMS macrolanguage, and for Norwegian Bokmål and Nynorsk.
- We also report negative results on our efforts with two-step coarse-to-fine classification approach, to be found in Appendix F.

2 Related Work

Accurate language identification (LID) is essential for building high-quality multilingual datasets, since documents or segments assigned to an incorrect language can severely contaminate the language-specific subsets, especially for low-resource languages.

¹<https://github.com/hplt-project/openlid>

²<https://github.com/hplt-project/openlid-v3-evaluation>

2.1 LID Tools and Methods

Numerous LID systems have been developed over the years, including `langid.py` (Lui and Baldwin, 2012), Google’s Compact Language Detector 2 (CLD2) and its neural network successor CLD3, the HeLI method (Jauhiainen et al., 2016), LanideNN (Kocmi and Bojar, 2017), or AfroLID (Adebara et al., 2022).

In recent years, classifiers based on the `fastText` model (Joulin et al., 2017) have become the de facto standard in large-scale corpus processing due to their efficiency and strong performance across many languages. `fastText` calculates a hidden representation of a text by summing the embeddings of words and character n-grams from this text. This representation is passed through a linear layer for classification. This approach takes both word- and character-level features into account, making it especially efficient for languages with rich morphology.

In this work, we focus on two `fastText`-based systems: **GlottLID** (Kargaran et al., 2023) and **OpenLID** (Burchell et al., 2023), both designed for multilingual scenarios and widely adopted in massive-scale data curation pipelines such as `FineWeb` (Penedo et al., 2024, 2025) or `HPLT` (de Gibert et al., 2024; Burchell et al., 2025; Oepen et al., 2025). `OpenLID` and `GlottLID` differ primarily in their language coverage and training data selection. `GlottLID` focuses on maximizing the language coverage and supports over 2,000 languages, though it incorporates training data from sources with more restrictive licenses. This extensive coverage results in a larger model size, but makes it suitable for use cases involving low-resource languages. In contrast, `OpenLID` prioritizes fully open-source training data with permissive licenses. The original version (v1) covering 201 languages was later updated (v2) to a reduced coverage of 189 languages. The reduction excluded three problematic languages and consolidated certain language varieties under their macrolanguage labels for compatibility with the `FLORES+` benchmark.

2.2 Broad-Coverage Evaluation Benchmarks

The most widely adopted multilingual language identification benchmarks are `FLORES+` and `UDHR`. **FLORES+**³ is a parallel corpus based on `FLORES-200` (NLLB Team et al., 2022), originally

³https://huggingface.co/datasets/openlanguagedata/flores_plus

developed for machine translation evaluation. It features two disjoint publicly available data splits, ‘dev’ and ‘devtest’, and currently covers 225 language varieties. In our experiments, we use the former for validation and the latter for evaluation. The **UDHR** dataset consists of translations of the Universal Declaration of Human Rights across 418 languages,⁴ which provides formal, declarative text in a single domain. The `UDHR` dataset features a test split only.

Additionally, **FastSpell** (Bañón et al., 2024) provides a benchmark specifically designed to discriminate between closely related languages in web documents, which addresses some of the limitations that stem from using relatively clean parallel corpora.

Finally, Oepen et al. (2025) present a manual inspection effort conducted in the context of `HPLT` 3.0 data collection.⁵ In that work, human annotators checked the prediction of `OpenLID-v2` for web documents. They could label documents as containing unnatural language, porn, web artifacts, and incorrect LID. While the data hasn’t been annotated for the correct language, it is possible to construct a LID evaluation dataset from the subset of correctly identified samples. Statistics of this resulting dataset, dubbed **HPLT-LID**, and evaluation results on it are to be found in the Appendix A. We also performed re-annotation of incorrectly classified examples for some languages.

2.3 Discriminating Between Similar Languages

The `VarDial` workshop has a long tradition of shared tasks that focus on language identification in challenging settings, such as the discrimination of closely related languages (e.g. Zampieri et al., 2014; Gaman et al., 2020; Aepli et al., 2023; Chifu et al., 2024). For example, shared tasks have focused on the Bosnian-Croatian-Montenegrin-Serbian macrolanguage or on regional languages of Italy. We leverage datasets made available in this context for a more fine-grained analysis of `OpenLID-v3` (see Section 4).

2.4 Evaluation Metrics

Caswell et al. (2020) argue that the commonly used metric precision (and its derivative, F1-score) are

⁴We use the `udhr-lid` version of the dataset available at <https://huggingface.co/datasets/cis-lmu/udhr-lid>

⁵https://github.com/hplt-project/release3_inspection/tree/main/annot_round1

misleading when evaluating models on the standard LID datasets, as the models face incomparably larger language imbalance in the real web crawls compared to any reasonable labeled dataset. For example, their baseline model achieving a median F1-score of 98% on benchmark data produced a set of monolingual corpora with a median precision of 5% only when evaluated manually on real data. On the other hand, recall and false positive rate (FPR) are not susceptible to class imbalance, and thus are recommended for model comparison.

Since most of our benchmarks are single-label, we report FPR, precision, recall and F1-score for them. For multilabel benchmarks, which are available for some related languages, where short samples may be valid in more than one language, we also report loose (is subset) and exact match metrics (cf. Fedorova et al., 2025).⁶ These metrics (we refer to them as multilabel classification metrics throughout the paper) are still less strict than our default ones, because they do not depend on the exact number of false positives or negatives in a set of predicted labels.

3 OpenLID-v3

This work is motivated by our experience with OpenLID-v2 in the context of its application to a large-scale web dataset HPLT 3.0, and the results of its manual inspection (Oepen et al., 2025). We identified a number of issues with OpenLID-v2, which led us to provide an updated version, OpenLID-v3, with a slightly different language inventory. We focus on improving OpenLID-v2 (rather than GlotLID, for example) because of the permissive license of all its training data. The identified issues are as follows:

- Support for the Latin language, which was present in OpenLID-v1, but removed in v2.
- OpenLID-v2 contains only Serbian in Cyrillic script, while Latin script is also widely used in Serbian non-governmental public and private communication. As a result, OpenLID-v2 erroneously classifies Serbian in Latin script as Bosnian or Croatian; according to the manual inspection, half of HPLT 3.0 Bosnian turned out to be Serbian written in Latin script.
- Since OpenLID-v2’s class inventory is limited to 200 languages, any text that is not natural

language (e.g., code, broken encoding – we label these instances as not-a-language, using `zxx_Zxxx` throughout the paper) or a natural language outside of the 200 the model was trained for (we refer to such instances as other), are still predicted to belong to one of the existing classes. In HPLT 3.0, the problem of not-a-language classes was overcome by rule-based filtering after LID, and the problem of the other class was solved by thresholding the predictions by 0.5 softmax scores. However, some classes can still accumulate a large number of not-a-language and other documents. We call this the *trash bin phenomenon*; in HPLT 3.0, Ligurian was found to be such a ‘trash bin’ language (Oepen et al., 2025).

- Some highly similar languages showed high confusion when tested on the FLORES+ development split and UDHR (Arabic dialects, Persian languages, Bambara and Dyula, Dzongkha and Tibetan, and other languages listed in Table 10).

Finally, the amount of identified data for certain languages in the HPLT dataset is very small compared to its number of speakers, e.g. Bengali and Tamil. See Table 10 for the full list of such languages.

When developing OpenLID-v3, we approached these issues in the following ways:

- We merged 8 Arabic dialects into one Arabic macrolanguage: `ara_Arab`; and two Persian varieties: `pes_Arab` and `prs_Arab` into the Farsi macrolanguage `fas_Arab`. This follows the OpenLID-v1 approach. Likewise, we merged Bambara and Dyula, which are mutually intelligible (Amuzu and Singler, 2014).
- We introduced the not-a-language class `zxx_Zxxx`, using the GlotLID training data labeled as `und_*` and `zxx_*`.
- We extended the training data for several languages, including Latin, and Serbian written in Latin script. We relied on the subsets of the GlotLID training data which were reported by its authors to not be noisy. We also removed one possibly noisy OpenLID-v2 training subset and added the most recent Wikipedia dumps, where adding GlotLID training data showed insufficient improvements during validation. The affected lan-

⁶<https://github.com/litglo/slide>

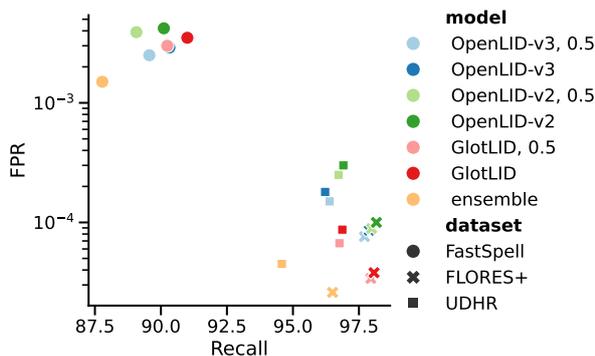


Figure 1: Model comparison on three LID benchmarks, FPR (lower is better) vs. recall (higher is better). For the full set of metrics see Table 9. FLORES+ refers to the devtest split, and FastSpell refers to our version, which excludes Nynorsk.

languages and data sources used are summarized in Table 10 in the Appendix.

- We also experimented with ensembling OpenLID-v3 and GlotLID by top-1 and top-3 agreement. Top-3 agreement worsened the results, which is expected for a single-label model, where only one class gets a relatively high softmax score, and all other classes are a long tail of small numbers.

The list of language labels supported by OpenLID-v3 is given in Table 15 in Appendix G.

3.1 Results

Figure 1 compares the results of our OpenLID-v3 model with the previous version (OpenLID-v2) as well as with GlotLID, focusing on the popular LID benchmarks: FLORES+, UDHR, and FastSpell. Following HPLT 3.0, we experimented with applying the same softmax thresholding of 0.5 to all three models.

OpenLID-v3 is on par with both OpenLID-v2 and GlotLID; the difference with and without thresholding softmax scores is not large among models, but the results are slightly better for OpenLID-v3 with thresholding, and for GlotLID without. This different behavior may be explained by the fact that GlotLID’s ‘trash bins’ fall outside of the classes we evaluate on. In our further experiments, we use OpenLID-v3 with thresholding, and GlotLID without, unless otherwise specified, including the ensembling approach.

Top-1 ensembling resulted in the lowest FPR across all datasets. Based on our experiments, this

approach was used to produce HPLT 4.0 datasets.⁷

4 Case Studies on Related languages

Results of both OpenLID and GlotLID are high on multilingual benchmarks. However, lower precision for similar languages may be hidden when averaging their scores with those of languages that are easy to distinguish. For this reason, we perform additional evaluation for groups of closely related languages. We choose three groups of such languages, within those with updated training data, based on benchmark availability and our language expertise.

4.1 Bosnian, Croatian, Montenegrin and Serbian

Bosnian, Croatian and Serbian are, along with Montenegrin, South Slavic languages spoken in the West Balkans; part of a group of historically and geographically close languages that are, for the most part, mutually intelligible to speakers of any one of them. They differ by variations in vocabulary and grammatical features – most notably, the orthography of the Slavic *jat* (ě) (Mešanović-Meša, 2011; Karavdić, 2017). Montenegrin is only present in GlotLID class inventory; for this reason, we do not focus on it. The analysis of its GlotLID predictions is to be found in Appendix C.

Test data. We use three datasets for the BCMS language group.⁸ First, BCMS **Twitter user** dataset (Ljubešić and Rupnik, 2022b), also featured in the BENCHiC BCMS benchmark (Rupnik et al., 2023). We use the version with multi-label annotations (Miletić and Miletić, 2024), comprising roughly 123 instances (in the test set), where each instance corresponds to one user and contains many texts written by that user. Second, we use the BCS portion of **ParlaSent** (Mochtak et al., 2023), comprising roughly 18,000 sentences from transcriptions of parliamentary debates, annotated with the speaker’s country of origin. Third, we make use of the **HPLT-LID** data. Since Serbian in Latin script was absent in OpenLID-v2, pre-annotation was only performed for Bosnian and Croatian. In total, 804 sentences were pre-annotated (402 predicted to be Bosnian and 402 predicted to be Croatian). Out of them, 114 Bosnian samples and 13 Croatian were pre-annotated to be false positives.

⁷<https://hplt-project.org/datasets/v4.0>

⁸Exact scores on these benchmarks are to be found in the Appendix C.

Model	Bosnian			Croatian			Serbian Latin		
	FPR	Prec.	Rec.	FPR	Prec.	Rec.	FPR	Prec.	Rec.
<i>Twitter users</i>	n=21			n=22			n=94		
OpenLID-v3	0.23	37.21	76.19	0.026	80.00	54.54	0	100.00	30.85
OpenLID-v2	0.8	16.96	90.48	0.017	84.62	50.00	0	0	0
GlottLID	0.42	16.95	47.62	0.26	36.17	77.27	0.12	82.14	24.49
Ensemble	0	0	0	0	0	0	0	0	0
<i>ParlaSent</i>	n=153			n=1,387			n=1,060		
OpenLID-v3	0.22	11.59	47.71	0.13	87.49	79.16	0.019	95.00	53.77
OpenLID-v2	0.44	0.08	57.51	0.26	78.05	80.25	0	0	0
GlottLID	0.26	9.59	43.79	0.242	81.02	90.48	0.006	97.10	31.60
Ensemble	0.14	13.15	34.64	0.11	89.25	77.21	0.003	98.38	28.77
<i>HPLT-LID (reannotated)</i>	n=5			n=7			n=110		
OpenLID-v3	0.15	5.26	40.00	0.03	42.86	85.71	0	100.00	50.00
OpenLID-v2	0.9	1.85	40.00	0.09	21.43	42.86	0	0	0
GlottLID	0.4	0.01	0.2	0.05	31.58	85.71	0	100.00	3.64
Ensemble	0.146	2.70	20.00	0.03	46.15	85.71	0	100.00	3.64

Table 1: Performance comparison across three benchmark datasets showing false positive rate (FPR), precision (Prec.), and recall (Rec.) for Bosnian, Croatian, and Serbian in Latin script. OpenLID-v3 uses softmax threshold 0.5. Ensemble represents top-1 agreement between OpenLID-v3 and GlotLID. Best values per metric and language are in **bold**.

We performed re-annotation of these false positive samples to their correct labels; the majority of them was found to be Serbian. The annotation was done by a native speaker of Croatian with linguistic background. One sample was detected to be valid both in Croatian and Serbian and annotated as multilabel; we excluded it from metric calculation.

4.1.1 Quantitative Evaluation

Table 1 presents FPR, precision and recall per language on Twitter user, ParlaSent and reannotated BCS part of HPLT-LID.

None of the models scored high on all three datasets. As HPLT 4.0 developers, we were mostly interested in results on Twitter, since its texts were the closest to the noisy web data, containing hyperlinks, emodjis etc. OpenLID-v3 turned out to be the best for Bosnian and Serbian on it; slightly higher precision of Croatian obtained from OpenLID-v2 comes at cost of lack of Serbian class label, which is not what we aim at. Importantly, GlotLID and OpenLID-v3 always disagree on Twitter data, which emphasizes that ensembling should be done cautiously for particular language groups. For the future work on improving predictions for the Twitter dataset, there is strong need for multilabel training data, silver approach from (Fedorova et al., 2025) might be helpful for future work.

Scores on ParlaSent were the highest, especially

Model	Loose	Exact	bos	hrv	srp
	Acc.	Acc.	F1	F1	F1
OpenLID-v3	46.34	40.65	57.14	72.73	49.15
GlottLID	40.65	33.33	28.17	55.74	40.00

Table 2: Loose accuracy, exact match accuracy and loose F1s per languages on multilabel BCS test data (Twitter). Ensemble@1 never agrees (always other, all metrics zero).

in precision of Serbian, which proves it to be the ‘easiest’ dataset.

Results of evaluation on reannotated HPLT-LID were similar to those on Twitter: OpenLID-v3 performed the best, and there were a high disagreement between it and GlotLID. However, the confusion of Bosnian and Serbian was still high.

Table 2 shows multilabel classification metrics on Twitter data. By these metrics, OpenLID-v3 is a clear winner over GlotLID.

4.1.2 Common Errors

We performed manual analysis of BCS predictions from all models and observed the following seven most frequent error patterns:

NE confusion Named entities are a common source of confusion. The presence of a different country name (e.g. Serbian news article about Croa-

(gold/predicted label)	b/h	b/s	b/x	h/b	h/s	h/x	s/b	s/h	s/x	% of total
NE confusion	12.0	-	2.0	4.0	-	2.0	10.0	6.0	8.0	6.1
lexical overlap	34.0	-	20.0	26.0	-	2.0	44.0	12.0	6.0	20.5
historic forms	-	-	-	36.0	-	2.0	-	-	-	5.8
<i>da</i> confusion	-	66.6	4.0	-	100.0	-	-	-	-	1.7
ungrammatical syntax	10.0	-	32.0	10.0	-	64.0	6.0	14.0	28.0	22.5
total ambiguity	26.0	-	12.0	6.0	-	20.0	-	34.0	26.0	19.9
misabeled minority rep	-	33.3	4.0	2.0	-	-	-	-	-	1.2
(other/unknown)	18.0	-	4.0	16.0	-	10.0	<i>40.0</i>	<i>34.0</i>	<i>32.0</i>	22.2

Table 3: V3-ensemble evaluation on ParlaSent; percentage of error occurrences, per sample of 50 mislabeled documents for each mismatched pair of labels. Exceptions are b/s, b/x, and h/s, which have a total of 2, 40, and 3 datapoints each, respectively. The bottom row, *other*, is a catch-all for unambiguous documents with clear language markers, which were nevertheless mislabeled.

tia), an individual or institution that adheres to the other language’s naming convention, or even word type confusion (“*Tome se niko ne raduje*” (“Nobody’s looking forward to that”, compared with the common Slovenian name *Tome*) identified as Slovenian), leads to mislabelling, due to lack of other language markers.

Lexical overlap For a human reader, the current orthography of *jat* is a strong discriminator between Serbian and Bosnian or Croatian, the former preferring the *e* form in the standard language (e.g. *videti, mleko*), and the other two using *je/ije* (*vidjeti, mlijeko*). However, documents are frequently mislabeled between Bosnian and Serbian, despite the presence of clear *e/ije* indicators, if the document also contains roots and noun/adjective inflections that are common to both Bosnian and Serbian, whereas Croatian uses a different surface form of the word (e.g. *obavezno* (Bosnian, Serbian) vs. *obvezatno* (Croatian)). The frequency of these error occurrences (*lexical overlap* in 3) seems to indicate that common roots and shared lexemes are a stronger signal than *jat* orthography.

Historic forms Similarly, all models frequently mislabel ParlaSent documents in Croatian as either Bosnian or Serbian, when the speakers slip into historic lexeme forms that were part of the shared Serbo-Croatian language, but are not present in current standard Croatian. Older speakers and colloquial language still frequently use *historic forms*, which again leads to mislabeling despite the unambiguous grammatical markers.

***da* confusion** Another theoretically strong indicator is the difference between the future tense in Croatian (modal verb + infinitive, e.g. (*ho*)ću *glasati, glasat ću*) and Serbian (modal verb + *da* + present simple, e.g. (*ho*)ću *da glasam*); both forms

are accepted in Bosnian. Frequently, Croatian documents are mislabelled as Serbian when there are multiple occurrences of the conjunction *da* (meaning *that*; not to be confused with the particle *da* (*yes*)). E.g., “*Ne sumnjam da je ovaj zakon nekima donio dobro i ne sumnjam da oni hvale ovaj zakon.*” (“I do not doubt **that** this legislation brought good to some, and I do not doubt **that** they praise it.”) – Croatian mislabelled as Serbian – vs. “*Neću da glasam za taj zakon.*” (“**I will not / don’t want to vote** for this legislation.”) – true Serbian.

Ungrammatical syntax Ungrammatical syntax comprises documents in non-standard language; mainly long run-on sentences, with a lack of punctuation, incorrect syntax, frequent interjections, and generally language that is highly colloquial in structure and word choice. While this mislabeling occurs across most pairings, these documents are most frequently labeled as *not-a-language* (zxx_Zxxx), particularly when there is a lack of any distinguishing grammatical or lexical language markers.

Finally, some datapoints are mislabeled due to extralinguistic factors.

Total ambiguity Some documents have no clear language markers between BCS. Most of these documents are mislabeled across models, and would arguably be hard even for a human annotator and native speaker to tease apart, due to a lack of unambiguous linguistic features.

Mislabeled minority representative Mislabeled minority representative is specific to the ParlaSent dataset, where statements are labeled according to the national parliament in which the discussion occurred. In a handful of cases, it is clear that the speaker is speaking a minority language, and the

model has identified it correctly, though it doesn’t match the parliament country of origin.

We performed a quantitative study of these error types on the ParlaSent predictions obtained from the best model (OpenLID-v3 top-1 ensemble with GlotLID). The results are presented in the Table 3. The last row (*other/unknown*) quantifies unambiguous documents with clear language markers, which were nevertheless mislabeled by the model. This is a particularly frequent occurrence for Latin-script Serbian.

4.2 Romance Languages of Italy and France

The HPLT 3.0 compilation efforts reported language identification issues connected to Ligurian. 31% of the samples annotated as Occitan in the UDHR dataset were predicted as Ligurian by OpenLID-v2. Upon closer inspection, it turned out that the Occitan part of UDHR contained seven translations – three truly Occitan ones and four Francoprovençal ones (see Table 12 in the Appendix). Although Ligurian, Occitan and Francoprovençal are all Romance languages and are spoken in adjacent areas (Southern France, Northwestern Italy and Western Switzerland), they are generally considered to belong to different genealogical sub-groups (Ledgeway and Maiden, 2016; Ramponi, 2024). Since Francoprovençal was missing from OpenLID-v2’s class inventory, its instances were erroneously predicted as Ligurian.

This suggests that (a) multilingual benchmarks cannot be fully relied on when it comes to closely related, unstandardized languages, and (b) LID tools struggle to identify low-resource languages especially if higher-resourced closely related varieties are present in their training data.

We further investigate the performance of OpenLID-v3 on these Romance languages. As we were unable to find a test set that covered Francoprovençal, Ligurian and Occitan,⁹ we use the ITDI dataset instead, which covers several languages and dialects of Italy (Aepli et al., 2022). The metrics were only calculated for Ligurian, Venetian and Friulian, as only these three languages are predicted by both OpenLID-v3 and GlotLID.

The results are presented in Table 4. While GlotLID is better than OpenLID-v3 both in precision and recall, ensembling them achieves the best precision and the lowest FPR. For both mod-

⁹While there exists an effort on Occitan LID (Miletic and Scherrer, 2022), we could not use it in our experiments, as it contains silver labels only.

Model	FPR	F1	Precision	Recall
OpenLID-v3	0.01	80.88	96.51	74.22
GlotLID	0.01	82.37	97.60	75.15
ensemble	0.004	76.68	98.47	68.95

Table 4: Comparison of OpenLID-v3 and GlotLID on the ITDI test set restricted to Ligurian, Venetian and Friulian (4,744 samples).

els, Venetian confusion is the highest with Italian, but OpenLID-v3 predicts fewer Italian false positives and more not-a-language ones (which is never the case for GlotLID). Friulian is mostly predicted correctly, more often with OpenLID-v3 (1,284 true positives) than with GlotLID (1,275). Ligurian is also more often predicted correctly by both models, with Sicilian and Friulian as the top confusion instances. For OpenLID-v3, the third most frequently predicted class is not-a-language. Ensembling the two models removes 45% of Venetian samples, while the number of Ligurian and Friulian ones remains roughly the same.

In sum, this case study emphasizes the importance of fine-grained benchmarks for related Romance languages. GlotLID might be a good choice as a standalone model; however, the most precise results are obtained when it is ensembled with OpenLID-v3, although at the price of removing samples from other easily confused languages such as Venetian.

4.3 Scandinavian Languages

Lastly, we focus on the North Germanic languages of Mainland Scandinavia, i.e. Norwegian Bokmål, Norwegian Nynorsk, Danish and Swedish. In addition to FLORES+ and UDHR, which cover all four languages, we experiment with two language-group-specific datasets: First, we use **SLIDE** (Fedorova et al., 2025), a multilabel dataset based on clean instances from Universal Dependencies (Nivre et al., 2020) (6,950 samples in total), containing many other languages. Second, we use **Nordic DSL** (Haas and Derczynski, 2021), a single-label dataset containing noisy sentences sourced from Wikipedia and featuring Faroese and Icelandic, in addition to the four aforementioned languages.

We did not use the **FastSpell** dataset for the calculation of metrics, as upon manual examination, we noticed a large share of its Nynorsk samples to be valid in both Bokmål and Nynorsk. We re-

Model	Norwegian Bokmål			Danish			Norwegian Nynorsk			Swedish			Other		
	FPR	Prec.	Rec.	FPR	Prec.	Rec.	FPR	Prec.	Rec.	FPR	Prec.	Rec.	FPR	Prec.	Rec.
<i>SLIDE</i>	n=2,098			n=677			n=1,628			n=1,250			n=1,745		
OpenLID-v3	0.04	88.73	87.04	0.02	81.81	83.75	0.03	89.66	84.64	0.008	96.11	96.80	0.03	89.68	95.13
OpenLID-v2	0.05	87.33	88.04	0.01	85.27	80.35	0.03	90.47	84.58	0.008	95.88	96.8	0.03	89.63	95.58
GlotLID	0.047	88.32	90.47	0.02	80.33	86.26	0.008	96.57	81.33	0.009	95.83	97.44	0.03	92.07	99.20
Ensemble	0.027	92.56	84.84	0.01	88.85	81.24	0.006	97.37	79.55	0.003	98.29	96.8	0.1	75.86	99.77
<i>Nordic DSL</i>	n=14,960			n=14,960			n=14,960			n=14,960			n=14,960		
OpenLID-v3	0.007	96.04	83.22	0.012	94.10	93.56	0.017	91.98	96.87	0.002	99.07	93.33	0.039	92.72	99.44
OpenLID-v2	0.02	91.85	85.37	0.007	96.12	89.78	0.02	91.97	96.75	0.002	98.77	92.73	0.04	92.82	99.56
GlotLID	0.01	94.58	94.50	0.006	97.09	93.77	0.006	97.20	96.47	0.004	98.16	93.69	0.025	95.16	99.36
Ensemble	0.005	97.19	82.78	0.003	98.17	91.82	0.004	98.13	95.92	0.001	99.35	92.01	0.079	86.32	99.74
<i>FLORES+</i>	n=1,012			n=1,012			n=1,012			n=1,012			n=212,520		
OpenLID-v3	1e-4	97.61	96.84	3e-5	99.31	99.21	7e-5	98.42	98.22	0	1.0	99.90	3e-3	99.99	1.0
OpenLID-v2	1e-4	96.84	96.84	2e-5	99.60	98.12	9e-5	98.12	98.22	0	1.0	99.90	0.004	99.99	99.99
GlotLID	1e-4	97.10	99.41	1e-5	99.70	99.51	1e-5	99.67	97.53	0	1.0	1.0	0	1.0	1.0
Ensemble	7e-5	98.39	96.74	6e-6	99.80	98.91	1e-5	99.70	97.23	0	1.0	99.90	1e-2	99.98	1.0
<i>UDHR</i>	n=62			n=61			n=58			n=61			n=27,515		
OpenLID-v3	4e-5	98.39	98.39	4e-5	98.36	98.36	8e-5	96.61	98.28	7e-5	96.83	1.0	8e-3	99.99	99.98
OpenLID-v2	7e-5	96.88	1.0	7e-5	96.77	98.36	4e-5	98.28	98.28	1e-4	95.31	1.0	0.004	99.99	99.97
GlotLID	4e-5	98.41	1.0	1e-4	93.75	98.36	4e-5	98.31	1.0	4e-5	98.39	1.0	0	1.0	99.98
Ensemble	4e-5	98.39	98.39	0	1.0	98.36	0	1.0	98.28	0	1.0	1.0	8e-3	99.99	1.0

Table 5: Performance comparison across four benchmark datasets showing the false positive rate (FPR), precision (Prec.), and recall (Rec.) for Scandinavian and other languages. *OpenLID-v3* uses softmax threshold 0.5. *Ensemble* represents top-1 agreement between *OpenLID-v3* and *GlotLID*. Best values per metric and language are in **bold**.

annotated this subset and evaluated it separately.

Table 5 summarizes the results. We find that GlotLID obtains the best recall, while OpenLID shows better false positive rate and precision. Precision and FPR are further improved by ensembling GlotLID and OpenLID-v3. OpenLID-v3 performs on par or slightly better than OpenLID-v2 on all benchmarks. Furthermore, the confusion matrices (available in the repository¹⁰) show that the two Norwegian varieties, Bokmål and Nynorsk, are most easily confused. Appendix E provides additional results: Table 14 shows the effect of applying softmax threshold across all models and datasets, and Table 13 shows the effect of ensembling across datasets. Both tables report multilabel classification metrics.

HPLT-LID The HPLT 3.0 manual inspection data (Oepen et al., 2025) only covers one Scandinavian language, namely Norwegian Bokmål. 402 documents were annotated as being correctly predicted Bokmål or not. We process this annotation in the same way as we did for BCMS annotation, excluding porn, unnatural language and artifacts. Documents with correctly identified language formed the dataset of documents guaranteed

¹⁰https://github.com/ltgoslo/slide/tree/main/src/eval_logs/vardial_2026

to be Bokmål, with 304 samples in total. 5 documents were annotated to be in a language other than Bokmål. We manually reviewed predictions of these 5 false positive Bokmål documents. Both OpenLID-v3 and GlotLID still predict them to be Bokmål. 3 of them are non-fluent Bokmål documents with many grammatical mistakes (articles of the wrong gender, articles and verbs omitted), as if written by foreign speakers, and 2 include code-switching, with Bokmål still being the predominant language.

FastSpell. A native speaker of Norwegian relabeled the FastSpell Nynorsk subset in a multilabel way, finding 40% of samples to be (also) valid Bokmål. The annotator also reported that many word sequences were not full sentences and some texts were just menu choices from a web page. The evaluation of OpenLID-v3 and GlotLID on the resulting dataset is presented in Table 6. While both models performed poorly in Bokmål detection, OpenLID-v3 performed slightly better for Nynorsk and produced fewer Danish and Swedish false positives, but more other false positives than GlotLID.

The manual analysis of the predictions showed that errors are often caused by named entities valid in any language using Latin script (“Georg Johannes Toft”), dates (“Levering 15. juli 2018”)

Model	NB F1	NN F1	DA FP	SV FP	Other FP
OpenLID-v3	71.90	92.19	16	8	52
GlottLID	79.07	91.14	27	19	16

Table 6: GlotLID (no threshold) and OpenLID-v3 (softmax threshold 0.5) on the re-annotated FastSpell Nynorsk data. F1 is for loose F1, FP is for false positives. 406 Nynorsk samples, 163 Bokmål samples.

and foreign words (“tomter i Parque natural cabo de gata”). These findings correspond to the error analysis from Fedorova et al. (2025) and further prove that shorter sequences, especially from the web, are the hardest for LID models.

In conclusion, we recommend OpenLID-v3 for general Scandinavian language identification if precision is prioritized, and GlotLID if recall is prioritized. For discriminating between Bokmål and Nynorsk, it is recommended to use a LID model trained specifically to separate these languages.

5 Conclusion

We have evaluated the performance of three LID models (two OpenLID versions and GlotLID) on four multilingual benchmarks (two based on ‘clean’ text and two based on web documents). Our new model, OpenLID-v3, performs on par or better in precision compared to its predecessors OpenLID-v2 and GlotLID, while the best precision is achieved by top-1 ensembling with GlotLID. We have also shown that reliable evaluation of LID models including similar languages should be done on benchmarks specific to these languages, even if overall metrics on large multilingual benchmarks are high. Since web data may contain short texts valid in more than one language, there is a need for more multilabel training data and benchmarks.

Limitations

The first limitation that we acknowledge is data mismatch between the evaluation and the intended use – we would ideally evaluate on web text. However, large-scale web LID data of sufficient quality were not available at the time of conducting our experiment. We are aware of the initiative by Suarez et al. (2026), but leave evaluation on their dataset for future work.

We have done our best to control for data contamination, where there was a known intersection of training and test data. However, this was not possible for tests on Nordic DSL, as this dataset,

while consisting of Wikipedia data, was heavily preprocessed by its authors.

A common limitation of all models under evaluation is the unavailability of fully parallel datasets for all languages of interest, which might cause models to overfit to some concepts and named entities that are more common in certain languages.

Ethical Considerations

All the new data annotations were done by the authors voluntarily and without any monetary compensation.

We did not check any of OpenLID-v3’s training data for inappropriate or biased content. We believe it does not do much harm, since the model is not generative. However, we can imagine some model predictions to be biased, particularly if used not as a classifier, but as a source of semantic representations.

A large amount of the current research is focused on collecting data for training large instruction-tuned generative models, capable of outputting grammatically and orthographically correct standard language, because of the high demand for such models in the public sector. This may result in loss of data written in low-resource varieties of mainstream languages, along with possibly valuable cultural knowledge.

Acknowledgements

This project was supported by the European Union Horizon Europe project no. 101070350 (HPLT). The computations were performed on resources provided by Sigma2 – the National Infrastructure for High-Performance Computing and Data Storage in Norway.

References

- Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Inciarte. 2022. *AfroLID: A neural language identification tool for African languages*. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1958–1981, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Noëmi Aepli, Antonios Anastasopoulos, Adrian Chifu, William Domingues, Fahim Faisal, Mihaela Găman, Radu Tudor Ionescu, and Yves Scherrer. 2022. Findings of the VarDial evaluation campaign 2022. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, Gyeongju,

- Republic of Korea. International Committee on Computational Linguistics (ICCL).
- Noëmi Aepli, Çağrı Çöltekin, Rob Van Der Goot, Tommi Jauhiainen, Mourhaf Kazzaz, Nikola Ljubešić, Kai North, Barbara Plank, Yves Scherrer, and Marcos Zampieri. 2023. [Findings of the VarDial evaluation campaign 2023](#). In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 251–261, Dubrovnik, Croatia. Association for Computational Linguistics.
- Milind Agarwal, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2023. [LIMIT: Language identification, misidentification, and translation using hierarchical models in 350+ languages](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14496–14519, Singapore. Association for Computational Linguistics.
- Evershed Amuzu and John Singler. 2014. [Codeswitching in west africa](#). *International Journal of Bilingualism*, 18:329–345.
- Marta Bañón, Gema Ramírez-Sánchez, Jaume Zaragoza-Bernabeu, and Sergio Ortiz Rojas. 2024. [FastSpell: The LangId magic spell](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7133–7140, Torino, Italia. ELRA and ICCL.
- Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. [An open dataset and model for language identification](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 865–879, Toronto, Canada. Association for Computational Linguistics.
- Laurie Burchell, Ona de Gibert, Nikolay Arefyev, Mikko Aulamo, Marta Bañón, Pinzhen Chen, Mariia Fedorova, Liane Guillou, Barry Haddow, Jan Hajič, Jindřich Helcl, Erik Henriksson, Mateusz Klimaszewski, Ville Komulainen, Andrey Kutuzov, Joonas Kytöniemi, Veronika Laippala, Petter Mæhlum, Bhavivya Malik, and 16 others. 2025. [An expanded massive multilingual dataset for high-performance language technologies \(HPLT\)](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17452–17485, Vienna, Austria. Association for Computational Linguistics.
- Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. 2020. [Language ID in the wild: Unexpected challenges on the path to a thousand-language web text corpus](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6588–6608, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Adrian-Gabriel Chifu, Goran Glavaš, Radu Tudor Ionescu, Nikola Ljubešić, Aleksandra Miletić, Filip Miletić, Yves Scherrer, and Ivan Vulić. 2024. [VarDial evaluation campaign 2024: Commonsense reasoning in dialects and multi-label similar language identification](#). In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 1–15, Mexico City, Mexico. Association for Computational Linguistics.
- Ona de Gibert, Graeme Nail, Nikolay Arefyev, Marta Bañón, Jelmer van der Linde, Shaoxiong Ji, Jaume Zaragoza-Bernabeu, Mikko Aulamo, Gema Ramírez-Sánchez, Andrey Kutuzov, Sampo Pyysalo, Stephan Oepen, and Jörg Tiedemann. 2024. [A new massive multilingual dataset for high-performance language technologies](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1116–1128, Torino, Italia. ELRA and ICCL.
- Mariia Fedorova, Jonas Sebulon Frydenberg, Victoria Handford, Victoria Ovedie Chruickshank Langø, Solveig Helene Willoch, Marthe Løken Midtgaard, Yves Scherrer, Petter Mæhlum, and David Samuel. 2025. [Multi-label Scandinavian language identification \(SLIDE\)](#). In *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025)*, pages 179–189, Tallinn, Estonia. University of Tartu Library, Estonia.
- Mihaela Gaman, Dirk Hovy, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Christoph Purschke, Yves Scherrer, and Marcos Zampieri. 2020. [A report on the VarDial evaluation campaign 2020](#). In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–14, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).
- Cyril Goutte, Serge Léger, and Marine Carpuat. 2014. [The NRC system for discriminating similar languages](#). In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 139–145, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.
- René Haas and Leon Derczynski. 2021. [Discriminating between similar Nordic languages](#). In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 67–75, Kiyv, Ukraine. Association for Computational Linguistics.
- Tommi Jauhiainen, Krister Lindén, and Heidi Jauhiainen. 2016. [HeLI, a word-based backoff method for language identification](#). In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 153–162, Osaka, Japan. The COLING 2016 Organizing Committee.
- Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2019. Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research*, 65:675–782.

- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. [Bag of tricks for efficient text classification](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.
- Zenaida Karavdić. 2017. Kako prepoznati bosanski, crnogorski, hrvatski i srpski standardni jezik? *BOSNIACA-časopis Nacionalne i univerzitetske biblioteke Bosne i Hercegovine*, (22):34–46.
- Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. [GlotLID: Language identification for low-resource languages](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218, Singapore. Association for Computational Linguistics.
- Tom Kocmi and Ondřej Bojar. 2017. [LanideNN: Multilingual language identification on character window](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 927–936, Valencia, Spain. Association for Computational Linguistics.
- Adam Ledgeway and Martin Maiden. 2016. *The Oxford Guide to the Romance Languages*. Oxford University Press.
- Nikola Ljubešić and Peter Rupnik. 2022a. [The news dataset for discriminating between bosnian, croatian and serbian SETimes.HBS 1.0](#). Slovenian language resource repository CLARIN.SI.
- Nikola Ljubešić and Peter Rupnik. 2022b. [The twitter user dataset for discriminating between bosnian, croatian, montenegrin and serbian twitter-HBS 1.0](#). Slovenian language resource repository CLARIN.SI.
- Marco Lui and Timothy Baldwin. 2012. [langid.py: An off-the-shelf language identification tool](#). In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.
- Emira Mešanović-Meša. 2011. *Kontrastivna analiza bosanskog, hrvatskog i srpskog jezika u zakonima Federacije Bosne i Hercegovine*. Slavistički komitet BiH.
- Aleksandra Miletić and Filip Miletić. 2024. A gold standard with silver linings: Scaling up annotation for distinguishing bosnian, croatian, montenegrin and serbian. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval)@ LREC-COLING 2024*, pages 36–46.
- Aleksandra Miletic and Yves Scherrer. 2022. [OcWikiDisc: a corpus of Wikipedia talk pages in Occitan](#). In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 70–79, Gyeongju, Republic of Korea. Association for Computational Linguistics.
- Michal Mochtak, Peter Rupnik, and Nikola Ljubešić. 2023. [The parlament multilingual training dataset for sentiment identification in parliamentary proceedings](#). (arXiv:2309.09783). ArXiv:2309.09783 [cs].
- 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.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefner, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. [No language left behind: Scaling human-centered machine translation](#). *Preprint*, arXiv:2207.04672.
- Stephan Oepen, Nikolay Arefev, Mikko Aulamo, Marta Bañón, Maja Buljan, Laurie Burchell, Lucas Charpentier, Pinzhen Chen, Mariia Fedorova, Ona de Gibert, Barry Haddow, Jan Hajič, Jindřich Helcl, Andrey Kutuzov, Veronika Laippala, Zihao Li, Risto Luukkonen, Bhavitvya Malik, Vladislav Mikhailov, and 13 others. 2025. [HPLT 3.0: Very large-scale multilingual resources for LLM and MT. mono- and bilingual data, multilingual evaluation, and pre-trained models](#). *Preprint*, arXiv:2511.01066.
- Guilherme Penedo, Hynek Kydlíček, Loubna Ben alal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. [The fineweb datasets: Decanting the web for the finest text data at scale](#). In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. [Fineweb2: One pipeline to scale them all – adapting pre-training data processing to every language](#). *Preprint*, arXiv:2506.20920.
- Alan Ramponi. 2024. [Language varieties of italy: Technology challenges and opportunities](#). *Transactions of the Association for Computational Linguistics*, 12:19–38.
- Daniel Romić. 2024. [Heritage bosnian, croatian, and serbian spoken by second generation speakers in germany he-bcs-ge](#).
- Peter Rupnik, Taja Kuzman, and Nikola Ljubešić. 2023. [BENCHiĆ-lang: A benchmark for discriminating between Bosnian, Croatian, Montenegrin and Serbian](#). In *Tenth Workshop on NLP for Similar Languages*,

Varieties and Dialects (VarDial 2023), pages 113–120, Dubrovnik, Croatia. Association for Computational Linguistics.

Pedro Ortiz Suarez, Laurie Burchell, Catherine Arnett, Rafael Mosquera-Gómez, Sara Hincapie-Monsalve, Thom Vaughan, Damian Stewart, Malte Ostendorff, Idris Abdulmumin, Vukosi Marivate, Shamsuddeen Hassan Muhammad, Atnafu Lambebo Tonja, Hend Al-Khalifa, Nadia Ghezaiel Hammouda, Verah Otiende, Tack Hwa Wong, Jakhongir Saydaliev, Melika Nobakhtian, Muhammad Ravi Shulthan Habibi, and 78 others. 2026. [Commonlid: Re-evaluating state-of-the-art language identification performance on web data](#). *Preprint*, arXiv:2601.18026.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. [A report on the DSL shared task 2014](#). In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 58–67, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

A Multilingual Benchmarks

In all evaluations on all datasets, metrics were only calculated on subset of languages, present in all models being compared. For GlotLID, we always used its the most recent version at the moment of writing, GlotLID-v3.

Table 7 shows statistics of the HPLT-LID dataset. Only those samples were accepted, which were not unnatural language, porn, web artifacts or incorrect LID. One must notice that, while this dataset contains texts from the web, it can not be used as a benchmark of noisy web data, because the texts were heavily cleaned and preprocessed before their annotation was performed. We believe however that HPLT-LID may be valuable for the community, because there exist not many LID benchmarks for some languages presented there..

Table 8 presents evaluation of all models (except OpenLID-v2, because the dataset was created with its help) under consideration on HPLT-LID. This is just another proof that there were no regression in performance of OpenLID-v3 compared to OpenLID-v2.

B Additional Data Sources

Table 10 shows the difference between OpenLID-v2 and OpenLID-v3 in terms of training data sources.

Languages outside OpenLID-v2 class inventory. We discussed adding natural language other class (languages present in glotlid-corpus other than our

Language	# docs	Language	# docs
ast_Latn	139	hrv_Latn	188
bos_Latn	49	ita_Latn	171
cat_Latn	196	jpn_Jpan	87
cmn_Hans	183	nob_Latn	304
ces_Latn	183	pes_Arab	209
deu_Latn	179	por_Latn	111
ell_Grek	613	rus_Cyrl	362
eng_Latn	126	spa_Latn	300
fra_Latn	139	slk_Latn	127
fin_Latn	335	srp_Cyrl	15
glg_Latn	182	yor_Latn	168
hin_Deva	136	total	4,502

Table 7: HPLT 3.0 manual inspection dataset statistics.

Model	FPR	F1	Precision	Recall
OpenLID-v3, 0.5	4e-5	99.84	99.91	99.78
OpenLID-v3	6e-5	99.84	99.85	99.84
GlottLID, 0.5	4e-5	99.51	99.89	99.16
GlottLID	4e-5	99.55	99.89	99.24
ensemble	1e-5	99.53	99.96	99.14

Table 8: Evaluation on HPLT-LID (n=3,873), with and without softmax threshold.

190+) for collecting ‘garbage’ documents and out-of-train-set languages. However, after looking closer into the data, just sampling from about 1,900 those languages and labeling them all as "other" did not seem a good idea. About 1,500 of those language were low-resource ones where only Bibles were available, reported as non-reliable data by GlotLID’s authors¹¹.

After removing non-reliable sources, 366 languages were left. The smallest language in OpenLID-v2 dataset is Yiddish with 923 samples only. Out of 366, we have 150 languages with more samples than that. Some of them were quite large like Low German (nds_Latn) with 117,676 samples (many languages in OpenLID-v2 dataset have less), national like Romansh (roh_Latn) or have more than a million speakers like Chechnyan (che_Cyrl). For the future OpenLID improvement, we recommend extending the model with these languages as separate labels instead of pushing them all into "other".

Not all languages with many samples are expected to be common in crawls: some of them

¹¹<https://github.com/cisnlp/GlotLID/blob/main/sources.md>

Model	FPR	F1	Precision	Recall
<i>FastSpell</i>	n=6,809			
OpenLID-v3, 0.5	2.5e-3	91.08	95.72	89.56
OpenLID-v3	2.9e-3	91.22	95.23	90.32
OpenLID-v2, 0.5	3.9e-3	90.04	94.55	89.07
OpenLID-v2	4.2e-3	90.37	94.05	90.11
GlottLID, 0.5	3.0e-3	91.32	95.19	90.24
GlottLID	3.5e-3	91.39	94.63	91.00
ensemble	1.5e-3	90.92	97.11	87.78
<i>FLORES+</i>	n=155,848			
OpenLID-v3, 0.5	7.6e-5	97.72	97.86	97.71
OpenLID-v3	8.6e-5	97.74	97.70	97.89
OpenLID-v2, 0.5	8.9e-5	98.09	98.62	97.98
OpenLID-v2	1.0e-4	98.11	98.46	98.16
GlottLID, 0.5	3.4e-5	98.34	99.52	97.95
GlottLID	3.8e-5	98.38	99.46	98.07
ensemble	2.6e-5	97.01	98.33	96.5
<i>UDHR</i>	n=10,283			
OpenLID-v3, 0.5	1.5e-4	96.14	97.13	96.39
OpenLID-v3,	1.8e-4	96.21	96.87	96.22
OpenLID-v2, 0.5	2.5e-4	96.27	97.26	96.73
OpenLID-v2	3.0e-4	96.21	96.93	96.91
GlottLID, 0.5	6.7e-5	97.39	98.96	96.76
GlottLID	8.7e-5	97.37	98.78	96.87
ensemble	4.5e-5	95.59	98.07	94.58

Table 9: Multilingual benchmarks. F1, precision and recall in %. ‘0.5’ stands for thresholding softmax scores at 0.5. Ensembling experiments are done with thresholded OpenLID-v3 and GlottLID without thresholding. FLORES+ refers to the devtest split, and FastSpell refers to our version that excludes Nynorsk.

are not alive like Classical Chinese (lzh_Hani) or constructed like Interlingua (ina_Latn). Since OpenLID already featured Esperanto and Latin, we believe other such language are also worth being edited.

Also we found more languages present in both Cyrillic and Latin script, like Serbian: Tatar, for example.

Training data for zxx_Zxxx. und_* are random sequences generated from different scripts.¹² zxx_* are non-linguistic noise in different scripts collected from the web.¹³ (We experimented with adding them as separated classes first; while our validation data had no such class, some languages outside our class inventory were classified as one of this classes; the choice between the two was random according to manual inspection.)

C Bosnian, Croatian, Montenegrin and Serbian

Other benchmarks We are aware of BCS news dataset (SETimes) (Ljubešić and Rupnik, 2022a), comprising roughly 9,000 parallel documents written in each of the languages, from a now-inactive news site that published articles in the languages of South-Eastern Europe. However, SETimes data were present among both OpenLID and GlottLID’s training data, but their splitting into sentences seem to be done in different ways. While we succeeded in deduplicating the test data from GlottLID’s training instances, we scored suspiciously high with OpenLID (more than 90% F1), which was never observed on other benchmarks. This is a clear sign that our deduplication has not worked. Thus, we only report GlottLID results here and leave proper evaluation on SETimes for future work. It is an important benchmark, because it comprises multi-sentence documents; average document length for each combination of gold and assigned label is given in Table 11. Generally, the common wisdom that longer documents lead to better classification holds here as well, mainly because longer documents give more opportunities for clear grammatical markers and discriminators to arise. Short sentences are much more likely to be mislabeled, across all models and datasets. Note that Serbian (srp, bottom row) doesn’t follow the pattern of

¹²https://github.com/cisnlp/GlottLID/blob/main/assets/train_gen_und.py

¹³<https://huggingface.co/datasets/cis-lmu/glottlid-corpus/tree/main/zxx>

Language	Sources	Justification
ast_Latn	commonvoice, leipzigwiki, tatoeba	high confusion with other
bam_Latn	ud, all OpenLID-v2 dyu data	high confusion with dyu_Latn
ben_Beng	commonvoice, GlotStoryBook, leipzigwiki, tatoeba, ud	small in HPLT 3.0
bod_Tibt	GlotStoryBook, tatoeba, the newest Wikipedia	high confusion with dzo_Tibt
cat_Latn	commonvoice, GlotStoryBook, leipzigwiki, tatoeba, ud	some confusion with oci_Latn
dan_Latn	GlotStoryBook, leipzigwiki, tatoeba, ud	high confusion with nob_Latn
dzo_Tibt	the newest Wikipedia	high confusion with bod_Tibt
frp_Latn	HFWikipedia	missed in OpenLID-v2, high confusion with lij_Latn
fuv_Latn	tatoeba	high confusion with other
gug_Latn	commonvoice, leipzigwiki, tatoeba, ud	poor in HPLT 3.0
heb_Hebr	leipzigwiki, tatoeba, ud	minor confusion with underrepresented Yiddish
kan_Knda	GlotStoryBook, leipzigwiki, tatoeba	small in HPLT 3.0
kin_Latn	afriqa, commonvoice, GlotStoryBook, tatoeba	high confusion with run_Latn
lat_Latn	tatoeba, ud, GlotStoryBook, leipzigwiki	missed in OpenLID-v2
lij_Latn	the newest Wikipedia	poor in HPLT 3.0
oci_Latn	removing OpenLID’s ‘pilar’	high confusion with lij_Latn
run_Latn	masakhanews, tatoeba, the newest Wikipedia	high confusion with kin_Latn
rus_Cyrl	commonvoice, GlotStoryBook, leipzigwiki, tatoeba, ud	high confusion with other
som_Latn	GlotStoryBook, leipzigwiki, masakhanews, tatoeba	high confusion with other
srp_Latn	ud, setimes, tatoeba	missed in OpenLID-v2, high confusion with Croatian and Bosnian
ssw_Latn	GlotStoryBook, Vukuzenzele, the newest Wikipedia	high confusion with other
sun_Latn	leipzigwiki, nusa, tatoeba	high confusion with other
tam_Tam1	commonvoice, GlotStoryBook, leipzigwiki, tatoeba, ud	small in HPLT 3.0
zxx_Zxxx	zxx_*, und_*	missed in OpenLID-v2

Table 10: OpenLID-v3 added/removed training data, compared to OpenLID-v2; added data were taken from GlotLID-corpus, if another not specified.

<i>(gold/pred)</i>	bos	hrv	srp
bos	1,922	571	-
hrv	534	1,833	-
srp	6,129	468	3,694

Table 11: GlotLID evaluation on SETimes; average document length by character count, across correctly and incorrectly labeled data.

longer documents being more likely to be correctly labeled. In addition to performance on Latin-script Serbian generally being lower, this discrepancy is likely due to common lexical overlap between Bosnian and Serbian, exacerbated by document length, and the model seemingly giving preference to the bos label, discussed further in this section.

We also acknowledge the existence of the COPA (Rupnik et al., 2023) and Heritage BCS (Romić, 2024) datasets, but, as the Serbian portion of COPA uses the standard Cyrillic script only, it was inappropriate for our evaluation; and Heritage BCS data comprise transcripts of bilingual speakers’ conversations, there is too much second-language interference in the documents for meaningful evaluation on LID task.

Montenegrin GlotLID predictions Interestingly, there are a handful of mislabeled documents

that GlotLID labels as Montenegrin (cnr) or Chakavian (ckm, a dialect of Croatian). Both the first¹⁴ and the second¹⁵ feature nonstandard syntax and repetitive *i* conjunctions and adjectives; both are, respectively, typical of Chakavian and Montenegrin poetry, which feature disproportionately in training data.

HPLT-LID reannotation details 201 samples were predicted by OpenLID-v2 to be Serbian Cyrillic, out of which 151 were pre-annotated to be false positives in the corresponding repository, but turned out to be missing annotation rather than false. 21 samples were detected to be translationese from Serbian Cyrillic and annotated with a new tag tra_Zxxx; 3 samples were found to be only valid in Montenegrin and annotated with cnr_Latn.

D Romance Languages of Italy and France

Table 12 lists the Francoprovençal and Occitan translations of the UDHR and their corresponding URLs.

¹⁴“Dijagnoza je točna, nisam i siguran da je i terapija, nisam siguran da je i terapija ispravna.”

¹⁵“Bijele noći, pa normalno da su bijele noći smetale, sve što je bijelo smetalo je, u mraku se puno bolje rade ovakvi poslovi.”

Dialect	URL
Occitan Auvergnat	http://efele.net/udhr/d/udhr_auv.txt
Occitan Languedocien	http://efele.net/udhr/d/udhr_lnc.txt
Occitan Provençal	http://efele.net/udhr/d/udhr_prv.txt
Francoprovençal Fribourg	http://efele.net/udhr/d/udhr_oci_1.txt
Francoprovençal Savoie	http://efele.net/udhr/d/udhr_oci_2.txt
Francoprovençal Vaud	http://efele.net/udhr/d/udhr_oci_3.txt
Francoprovençal Valais	http://efele.net/udhr/d/udhr_oci_4.txt

Table 12: Occitan and Francoprovençal UDHR translations and their URLs.

E Scandinavian Languages

Table 13 presents ensemble results combining GlotLID (no threshold) and OpenLID-v3 (softmax threshold 0.5). Table 14 shows performance comparison for different versions of OpenLID with GlotLID on the SLIDE, Nordic DSL, FLORES+ devtest, and UDHR test sets with and without thresholding.

F Hierarchical Models

Hierarchical LID systems decompose the classification task into a series of decision steps, typically progressing from broad language groups to the individual languages or variants (Jauhiainen et al., 2019; Agarwal et al., 2023). The motivation for these approaches is that closely related languages are naturally more challenging to discriminate for LID models (Goutte et al., 2014).

We trained hierarchical fastText-based models and evaluated their effect on the following language groups: Scandinavian languages (Danish, Faroese, Icelandic, Norwegian Bokmål, Norwegian Nynorsk, and Swedish), South Slavic languages (Bosnian, Croatian, Serbian in Latin script), along with varieties of Arabic (Egyptian, Levantine, Mesopotamian, Moroccan, Najdi, Standard Arabic, Ta’izzi-Adeni, and Tunisian), and Persian (Dari and Iranian Persian). For each group, we train a specialized language classifier on the subset of the OpenLID-v3 corpus composed of these languages. During inference, every time the base model predicts a language from one of the four groups, we replace the base model prediction with the prediction of the specialized model.

The specialized models follow the same architecture and training procedure as OpenLID-v3. For each language group, we use the corresponding language-specific subset of the OpenLID-v3 training data.

We have not found a significant difference in the performance of OpenLID-v3 and the hierarchical models in any of the related language groups when evaluating on FLORES+ or UDHR. We hypothesize that the capacity of the large LID model is not saturated yet and therefore there is nothing to be gained by using a subset of the training data. Instead, effort should be put into securing larger and higher-quality annotated data for these language groups.

G List of Supported Languages

Table 15 shows the list of language labels supported by OpenLID-v3.

Dataset	Loose Acc.	Exact Acc.	NB F1	DA F1	NN F1	SV F1	Other F1
SLIDE	<i>94.62</i>	<i>92.30</i>	92.52	<i>90.53</i>	91.52	<i>98.21</i>	91.70
Nordic DSL	93.66	93.66	<i>89.41</i>	<i>94.89</i>	97.01	95.54	92.54
FLORES+	99.97	99.97	<i>97.56</i>	<i>99.35</i>	<i>98.45</i>	99.95	99.99
UDHR	99.99	99.99	98.39	99.17	<i>99.13</i>	100.00	100.00

Table 13: Ensembling GlotLID (no threshold) and OpenLID-v3 (softmax threshold 0.5) on Scandinavian languages. Bold indicates improvement over both individual models. Italic indicates improvement over OpenLID-v3 only.

Thres.	Model	Loose Acc.	Exact Acc.	NB F1	DA F1	NN F1	SV F1	Other F1
<i>SLIDE</i>								
No	OpenLID-v3	94.46	90.55	94.33	89.27	93.61	96.30	93.13
	OpenLID-v2	94.39	90.36	94.07	89.60	94.15	95.89	93.11
	OpenLID-v1	93.61	89.77	93.91	87.56	93.03	95.51	91.42
	GlottLID	97.20	93.40	95.83	93.37	94.84	98.23	98.05
0.5	OpenLID-v3	95.55	91.87	94.25	91.60	93.52	97.62	95.16
	OpenLID-v2	95.63	91.83	94.16	91.58	94.12	97.38	95.29
	OpenLID-v1	94.81	91.14	94.00	89.26	93.36	96.84	93.76
	GlottLID	97.09	93.45	95.56	93.40	94.52	98.22	97.73
	n samples	6,950	6,950	2,098	677	1,628	1,250	1,745
<i>Nordic DSL (from 50k split)</i>								
No	OpenLID-v3	94.71	94.71	89.29	93.63	94.24	96.06	97.36
	OpenLID-v2	94.32	94.32	88.63	92.89	93.79	95.67	97.30
	OpenLID-v1	94.17	94.17	88.73	93.64	92.78	95.37	97.15
	GlottLID	96.19	96.19	94.54	95.40	96.84	95.87	97.21
0.5	OpenLID-v3	94.30	94.30	89.17	93.83	94.36	96.11	95.96
	OpenLID-v2	93.96	93.96	88.49	92.84	94.30	95.65	96.07
	OpenLID-v1	93.88	93.88	88.47	93.65	93.16	95.41	96.13
	GlottLID	96.10	96.10	94.48	95.51	96.85	95.96	96.86
	n samples	74,800	74,800	14,960	14,960	14,960	14,960	14,960
<i>FLORES+ devtest</i>								
No	OpenLID-v3	99.97	99.97	97.53	99.11	98.22	99.90	100.00
	OpenLID-v2	99.97	99.97	97.19	98.91	98.08	99.90	100.00
	OpenLID-v1	99.97	99.97	97.63	98.86	98.57	99.75	100.00
	GlottLID	99.98	99.98	98.24	99.60	98.60	100.00	100.00
0.5	OpenLID-v3	99.97	99.97	97.22	99.26	98.32	99.95	100.00
	OpenLID-v2	99.97	99.97	96.84	98.86	98.17	99.95	100.00
	OpenLID-v1	99.97	99.97	97.53	98.86	98.62	99.85	100.00
	GlottLID	99.98	99.98	98.19	99.55	98.60	100.00	100.00
	n samples	216,568	216,568	1,012	1,012	1,012	1,012	212,520
<i>UDHR</i>								
No	OpenLID-v3	99.75	99.75	93.85	97.44	82.19	78.21	99.88
	OpenLID-v2	99.74	99.74	96.12	85.71	95.80	73.49	99.87
	OpenLID-v1	99.79	99.79	83.22	95.31	91.20	88.41	99.90
	GlottLID	99.97	99.97	99.20	96.00	99.15	99.19	99.99
0.5	OpenLID-v3	99.97	99.97	98.39	98.36	97.44	98.39	99.99
	OpenLID-v2	99.97	99.97	98.41	97.56	98.28	97.60	99.99
	OpenLID-v1	99.97	99.97	98.41	99.19	95.00	99.19	99.99
	GlottLID	99.99	99.99	99.20	98.36	100.00	100.00	100.00
	n samples	27,757	27,757	62	61	58	61	27,515

Table 14: Comprehensive comparison of language identification models across all test datasets with and without softmax thresholding. Bold values indicate best performance for each metric within each dataset/threshold combination.

Lang. code	Name	Lang. code	Name	Lang. code	Name	Lang. code	Name
ace_Arab	Achinese	fon_Latn	Fon	lij_Latn	Ligurian	slk_Latn	Slovak
ace_Latn	Achinese	fra_Latn	French	lim_Latn	Limburgan (Limburger, Limburgish)	slv_Latn	Slovenian
afr_Latn	Afrikaans	frp_Latn	Arpitan (Franco-provençal)	lin_Latn	Lingala	smo_Latn	Samoan
als_Latn	Tosk Albanian	fur_Latn	Friulian	lit_Latn	Lithuanian	sna_Latn	Shona
amh_Ethi	Amharic	fuv_Latn	Nigerian Fulfulde	lmo_Latn	Lombard	snd_Arab	Sindhi
ara_Arab	Arabic	gaz_Latn	West Central Oromo	ltg_Latn	Latgalian	som_Latn	Somali
asm_Beng	Assamese	gla_Latn	Scottish Gaelic	ltz_Latn	Luxembourgish (Letzeburgesch)	sot_Latn	Southern Sotho
ast_Latn	Asturian (Asturleonese, Bable, Leonese)	gle_Latn	Irish	lua_Latn	Luba-Lulua	spa_Latn	Spanish (Castilian)
awa_Deva	Awadhi	glg_Latn	Galician	lug_Latn	Ganda	srd_Latn	Sardinian
ayr_Latn	Central Aymara	gug_Latn	Paraguayan Guaraní	luo_Latn	Luo (Kenya and Tanzania) (Dholuo)	srp_Cyrl	Serbian
azb_Arab	South Azerbaijani	guj_Gujr	Gujarati	lus_Latn	Lushai	srp_Latn	Serbian
azj_Latn	North Azerbaijani	hat_Latn	Haitian (Creole, Haitian Creole)	lvs_Latn	Standard Latvian	ssw_Latn	Swati
bak_Cyrl	Bashkir	hau_Latn	Hausa	mag_Deva	Magahi	sun_Latn	Sundanese
bam_Latn	Bambara	heb_Hebr	Hebrew	mai_Deva	Maithili	swe_Latn	Swedish
ban_Latn	Balinese	hin_Deva	Hindi	mal_Mlym	Malayalam	swh_Latn	Swahili (individual language) (Kiswahili)
bel_Cyrl	Belarusian	hne_Deva	Chhattisgarhi	mar_Deva	Marathi	szl_Latn	Silesian
bem_Latn	Bemba (Zambia)	hrv_Latn	Croatian	min_Latn	Minangkabau	tam_Taml	Tamil
ben_Beng	Bengali	hun_Latn	Hungarian	mkd_Cyrl	Macedonian	taq_Latn	Tamasheq
bho_Deva	Bhojpuri	hye_Armn	Armenian	mlt_Latn	Maltese	taq_Tfng	Tamasheq
bjn_Arab	Banjar	ibo_Latn	Igbo	mni_Beng	Manipuri	tat_Cyrl	Tatar
bjn_Latn	Banjar	ilo_Latn	Iloko	mos_Latn	Mossi	tel_Telu	Telugu
bod_Tibt	Tibetan	ind_Latn	Indonesian	mri_Latn	Maori	tgk_Cyrl	Tajik
bos_Latn	Bosnian	isl_Latn	Icelandic	mya_Mymr	Burmese	tha_Thai	Thai
bug_Latn	Buginese	ita_Latn	Italian	nld_Latn	Dutch (Flemish)	tir_Ethi	Tigrinya
bul_Cyrl	Bulgarian	jav_Latn	Javanese	nno_Latn	Norwegian Nynorsk	tpi_Latn	Tok Pisin
cat_Latn	Catalan (Valencian)	jpn_Jpan	Japanese	nob_Latn	Norwegian Bokmål	tsn_Latn	Tswana
ceb_Latn	Cebuano	kab_Latn	Kabyle	npi_Deva	Nepali (individual language)	tso_Latn	Tsonga
ces_Latn	Czech	kac_Latn	Kachin (Jingpho)	nso_Latn	Pedi (Northern Sotho, Sepedi)	tuk_Latn	Turkmen
cjk_Latn	Chokwe	kam_Latn	Kamba (Kenya)	nus_Latn	Nuer	tum_Latn	Tumbuka
ckb_Arab	Central Kurdish	kan_Knda	Kannada	nya_Latn	Chichewa (Chewa, Nyanja)	tur_Latn	Turkish
cmn_Hans	Mandarin Chinese	kas_Arab	Kashmiri	oci_Latn	Occitan (post 1500)	twi_Latn	Twi
cmn_Hant	Mandarin Chinese	kas_Deva	Kashmiri	ory_Orya	Odia (Oriya (individual language))	uig_Arab	Uighur (Uyghur)
crh_Latn	Crimean Tatar (Crimean Turkish)	kat_Geor	Georgian	pag_Latn	Pangasinan	ukr_Cyrl	Ukrainian
cym_Latn	Welsh	kaz_Cyrl	Kazakh	pan_Guru	Panjabi (Punjabi)	umb_Latn	Umbundu
dan_Latn	Danish	kbp_Latn	Kabiyè	pap_Latn	Papiamentu	urd_Arab	Urdu
deu_Latn	German	kea_Latn	Kabuverdianu	pbt_Arab	Southern Pashto	uzn_Latn	Northern Uzbek
dik_Latn	Southwestern Dinka	khk_Cyrl	Halh Mongolian	plt_Latn	Plateau Malagasy	vec_Latn	Venetian
dzo_Tibt	Dzongkha	khm_Khmr	Khmer (Central Khmer)	pol_Latn	Polish	vie_Latn	Vietnamese
ekk_Latn	Standard Estonian	kik_Latn	Kikuyu (Gikuyu)	por_Latn	Portuguese	war_Latn	Waray (Philippines)
ell_Grek	Modern Greek (1453-)	kin_Latn	Kinyarwanda	quy_Latn	Ayacucho Quechua	wol_Latn	Wolof
eng_Latn	English	kir_Cyrl	Kirghiz (Kyrgyz)	ron_Latn	Romanian (Moldavian, Moldovan)	xho_Latn	Xhosa
epo_Latn	Esperanto	kmb_Latn	Kimbundu	run_Latn	Rundi	ydd_Hebr	Eastern Yiddish
eus_Latn	Basque	kmr_Latn	Northern Kurdish	rus_Cyrl	Russian	yor_Latn	Yoruba
ewe_Latn	Ewe	knc_Arab	Central Kanuri	sag_Latn	Sango	yue_Hant	Yue Chinese
fao_Latn	Faroese	knc_Latn	Central Kanuri	san_Deva	Sanskrit	zgh_Tfng	Standard Moroccan Tamazight
fas_Arab	Persian	kor_Hang	Korean	sat_0lck	Santali	zsm_Latn	Standard Malay
fij_Latn	Fijian	ktu_Latn	Kituba (Democratic Republic of Congo)	scn_Latn	Sicilian	zul_Latn	Zulu
fil_Latn	Filipino (Pilipino)	lao_Lao	Lao	shn_Mymr	Shan	zxx_Zxxx	No linguistic content (Not applicable)
fin_Latn	Finnish	lat_Latn	Latin	sin_Sinh	Sinhala (Sinhalese)		

Table 15: OpenLID-v3 supported languages (model class labels).

Improving Dialect Robustness in Large Language Models via LoRA and Mixture-of-Experts

Sanjh Maheshwari*, Aniket Singh Rajpoot*, Oana Cocarascu, Mamta

King's College London

{sanjhmaheshwari1209, aniket161200}@gmail.com

{oana.cocarascu, mamta.name}@kcl.ac.uk

Abstract

Despite the success of large language models (LLMs) in a wide range of applications, it has been shown that their performance varies across English dialects. Differences among English dialects are reflected in vocabulary, syntax, and writing style, and can adversely affect model performance. Several studies evaluate the dialect robustness of LLMs, yet research on enhancing their robustness to dialectal variation remains limited.

In this paper, we propose two parameter-efficient frameworks for improving dialectal robustness in LLMs: *DialectFusion* where we train separate LoRA layers for each dialect and apply different LoRA merging methods, and *DialectMoE* which is built on top of Mixture of Experts LoRA and introduces multiple LoRA-based experts to the feed-forward layer to internally model the dialectal dependencies. Our comprehensive analysis on five open-source LLMs for sentiment and sarcasm tasks in zero- and few-shot settings shows that our proposed approaches enhance the dialect robustness of LLMs and outperforms instruct and LoRA fine-tuning based approaches.

1 Introduction

Large Language Models (LLMs) have been successful in a variety of tasks including sentiment and sarcasm classification (Zhang et al., 2025; Vajjala and Shimangaud, 2025). However, these models often exhibit a bias towards mainstream dialects, limiting their performance in dialect-specific contexts (Srirag et al., 2025b; Lin et al., 2025). Regional English exhibits distinct grammatical structures, vocabulary, and expressions. As LLMs are trained on datasets composed of standard English, they often struggle to generalize to more diverse linguistic varieties (Joshi et al., 2025).

Several studies have explored ways to improve dialect robustness using rule-based systems (Ziems

et al., 2023) or adapting language models for dialectal tasks (Sun et al., 2023; Liu et al., 2023). More recently Srirag et al. (2025a) introduced BESSTIE, a manually annotated dataset with three varieties of English (Australian, Indian, British), however, their evaluation focused on encoder and decoder models. Although parameter-efficient methods such as LoRA (Hu et al., 2022) have been explored for dialect adaptation, existing studies (Faisal and Anatasopoulos, 2024; Liu et al., 2023) do not systematically evaluate their performance across modern LLMs nor investigate approaches to enhance robustness across dialect groups (Srirag et al., 2025a).

To mitigate these gaps, in this paper we propose two parameter-efficient approaches, *DialectFusion* and *DialectMoE*, to improve the dialectal robustness of open-source LLMs.¹ In *DialectFusion*, we train separate LoRA adapters for each dialect and then merge them using two methods: Learnable Concatenation (CAT) (Prabhakar et al., 2025) and TrIm, Elect, and Merge (TIES) (Yadav et al., 2023). In *DialectMoE*, LoRA-based experts are added to the feed-forward layer of an LLM along with a gate router (Li et al., 2024) to enhance dialect performance. We focus on three English dialects: Australian, British, and Indian.

To evaluate the effectiveness of our proposed approaches, we fine-tune five open-source LLMs: three predominantly pre-trained using English corpora (Mistral-v0.1 7B Instruct (Jiang et al., 2023), Gemma2-9B Instruct (Team, 2024), Phi-3 Medium Instruct (et al., 2024)) and two multilingual models (Llama 3.1 8B Instruct (Weerawardhena et al., 2025), Qwen 2.5 7B Instruct (Yang et al., 2025) using the BESSTIE dataset (Srirag et al., 2025a). We compare our approaches with instruction-tuned models and LoRA fine-tuning on the full dataset without dialect distinctions in zero- and few-shot

¹Our code is available at <https://github.com/Sanjh-Maheshwari/LLM-Dialect-Robustness>.

*Equal contribution.

settings. Our results show that our proposed methods consistently outperform both instruction-tuned and LoRA fine-tuning.

To summarize, our **contributions** are as follows:

- We introduce two parameter-efficient approaches based on LoRA adapters and Mixture-of-Experts to improve dialectal robustness in LLMs;
- We fine-tune and evaluate five recent LLMs on the BESSTIE dataset, covering sentiment and sarcasm classification, enabling a systematic assessment of robustness across English dialects;
- We compare our methods against strong baselines including instruction-tuned models and LoRA fine-tuning without dialect separation, and show consistent improvements in both zero-shot and few-shot settings, demonstrating improved robustness across dialects.

2 Related Works

Large Language Models (LLMs) often perform poorly on dialectal data due to syntactic, orthographic, and lexical variation (Srirag et al., 2025b). As training data is heavily skewed toward Standard American English (Joshi et al., 2025), monolingual and multilingual LLMs often fail to generalize to dialectal text without explicit adaptation.

Several efforts have been made to improve dialect robustness. Ziems et al. (2023) proposed Multi-Value, a rule-based translation system consisting of 50 English dialects and their 189 unique linguistic features. Using these transformation rules, Liu et al. (2023) created a synthetic dataset using transformation rules, trained adapters for each linguistic feature, and then fused these adapters into a single model. Some works adapted language models for tasks requiring extensive dialectal knowledge, however their focus was on smaller task-specific language models such as mT5, Flan-T5 rather than recent LLMs (Sun et al., 2023; Liu et al., 2023). In order to increase robustness to dialectal variance without impairing downstream task performance, Sun et al. (2023) proposed a dialect-robust evaluation metric and NANO, a training schema which introduces regional and language information to pretraining. Faisal and Anastasopoulos (2024) applied LoRA (Hu et al., 2022) to adapt a multilingual instruction-tuned model to improve task performance on three South Slavic dialects.

More recently, Srirag et al. (2025a) proposed a dataset for sentiment and sarcasm classification for three varieties of English: Australian, Indian, and British, collected from Google Places reviews and Reddit comments. They evaluated nine LLMs (6 encoders and 3 decoders) and showed that models are better at sentiment classification compared to sarcasm detection. They also highlighted that monolingual models perform slightly better than multilingual models suggesting that multilingual pre-training does not adequately capture intra-language variation. The results are reported only for fine-tuned models, which offers limited insights into the actual improvements on dialectal variants.

Previous work has not explored recent advances in LLM fine-tuning including LoRA merging strategies (e.g. CAT (Prabhakar et al., 2025) and TIES (Yadav et al., 2023)) and parameter efficient implementations (e.g. MixLoRA (Li et al., 2024)) despite their proven effectiveness in multi-task learning settings that closely resemble multi-dialect settings. We bridge this gap through a systematic evaluation of six methods across five LLMs in both zero-shot and few-shot settings.

3 Methodology

We propose two parameter-efficient approaches to improve dialect robustness in LLMs:

- *DialectFusion* trains separate LoRA adapters for each dialect and then merges them using two methods: CAT (Prabhakar et al., 2025) or TIES (Yadav et al., 2023).
- *DialectMoE* employs mixture-of-experts and inserts multiple LoRA-based experts within the feed-forward network blocks of the frozen pre-trained model with a top-k router to dynamically select dialect-specific experts without requiring separate adapter training (Li et al., 2024).

We focus on three English dialects, specifically Australian (AU), British (UK), and Indian (IN).

3.1 DialectFusion

DialectFusion is a parameter-efficient method designed to improve dialect robustness of LLMs. Here, we train multiple LoRA adapters, one for each target dialect, as shown in Figure 1. This allows the model to capture dialect specific linguistic features independently.

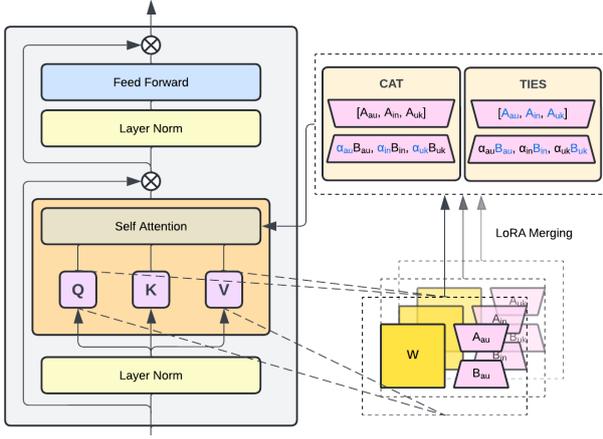


Figure 1: Overview of *DialectFusion* using LoRA merging (CAT and TIES). α_{au} , α_{in} and α_{uk} highlighted in blue in case of CAT are trainable. Similarly, A_{au} , A_{in} , A_{uk} , B_{au} , B_{in} and B_{uk} highlighted in blue in TIES are pre-trained dialect-specific LoRA adapters merged with fixed α weights.

One approach is to keep these adapters separate (LoRA Separate) where each dialect-specific adapter is loaded independently during inference based on the target dialect. While this maintains the distinct linguistic characteristics of each dialect, it requires prior knowledge of the dialect and separate inference passes for each dialect. Another approach is to merge the trained adapters using LoRA merging techniques such as CAT (Prabhakar et al., 2025) and TIES (Yadav et al., 2023) to produce a single adapter for all dialects.

CAT CAT is a LoRA merging method that first trains separate LoRA adapters for each dialect independently, then learns layer-specific (l) mixing coefficients to optimally combine these adapters, enabling the model to handle multiple dialects without retraining the base model. In each standard LoRA layer, the original weight matrix (W_0) stays frozen, while the update is calculated as $\Delta W = BA$, where $A \in \mathbb{R}^{r \times k}$ and $B \in \mathbb{R}^{d \times r}$ are low-rank decomposition matrices with $r \ll \min(d, k)$. Here, k and d are the input and output dimensions of W_0 respectively, and r is the low-rank bottleneck dimension. For an input vector x , the forward pass computes $W_0x + BAx$.

CAT extends this by merging trained LoRA adapters. As shown in Figure 1, this approach combines LoRA updates from three dialect adapters using layer-specific mixing coefficients:

$$\Delta W^l = \alpha_{au}^l B_{au} A_{au}^\top + \alpha_{in}^l B_{in} A_{in}^\top + \alpha_{uk}^l B_{uk} A_{uk}^\top$$

where $\alpha_{au}^l, \alpha_{in}^l, \alpha_{uk}^l \in [0, 1]$ are trainable layer-

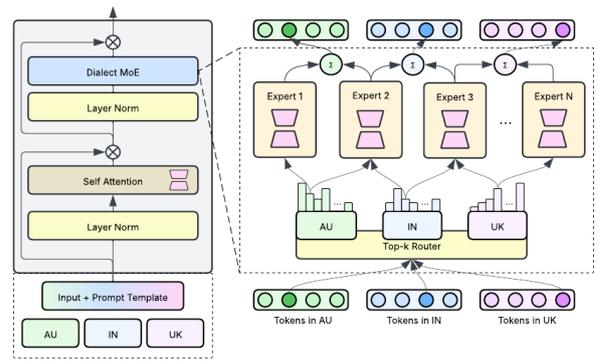


Figure 2: Overview of *DialectMoE*. LoRA adapters are applied to self-attention layers as standard. At the feed-forward layer, a sparse MoE is formed where all experts share the base Feed Forward Network weights.

specific merging coefficients for the Australian, Indian, and UK dialect adapters respectively, and A_k, B_k (for $k \in \{au, in, uk\}$) are the low-rank matrices from each pre-trained dialect-specific LoRA adapter. Unlike standard LoRA where only A and B are trained, in CAT the dialect-specific LoRAs are first trained independently, then frozen, and only the mixing coefficients α^l are learned to optimally combine them at each layer l .

TIES TIES addresses interference between merged models through a three-step process. This method applies pre-processing steps before merging. First, we prune the smallest values in (A_k, B_k) for $k \in \{au, in, uk\}$, retaining only the top $\lambda \in [0, 1]$ fraction based on magnitude. Second, we compute a majority sign mask by summing all pruned parameters and storing the sign. Finally, we apply weighted linear merging using only the parameters matching the majority sign, effectively resolving conflicts when merging adapters with overlapping parameters as shown below:

$$\Delta W^l = (\alpha_{au} B_{au} + \alpha_{in} B_{in} + \alpha_{uk} B_{uk}) (\alpha_{au} A_{au} + \alpha_{in} A_{in} + \alpha_{uk} A_{uk})^\top$$

The weights in the equation, α_{au} , α_{in} , and α_{uk} are fixed hyperparameters which are typically set uniformly, e.g., $\alpha_{au} = \alpha_{in} = \alpha_{uk} = 1/3$. (Yadav et al., 2023)

3.2 DialectMoE

DialectMoE inserts multiple LoRA-based experts within the feed-forward network blocks of the frozen pre-trained model, employing a top-k router mechanism to dynamically select the most relevant experts for each input (Li et al., 2024). In our case,

each expert is assumed to be specialized in a distinct dialect after training, allowing the model to process dialect-specific nuances without the need to train separate adapters for each dialect and overloading the GPU memory.

As shown in Figure 2, *DialectMoE* consists of two parts. The left component is the transformer architecture with frozen base model augmented with LoRA adapters at various layers (Self Attention, Layer Norm, and Feed-Forward layers). The right component shows the sparse MoE mechanism with three dialect-specific experts.

Input tokens first pass through the standard transformer layers (Self Attention and Layer Norm) in the base model. At the feed forward layers, the input tokens corresponding to each dialect are passed to the top-k router which computes routing probabilities and assigns each token from various tasks (i.e. in our case dialects: AU, IN, UK) to different expert modules. The boxes labeled "AU", "IN", and "UK" below the experts represent the weighted gating mechanism that determines how much each expert contributes to the final output.

Each expert module comprises a feed-forward network with LoRA adapters. The outputs from the selected adapters are then combined using weighted sum (\sum). The weights are determined using the routing probabilities assigned by the top-k router. The output is then passed to the next layers of transformers allowing the model to leverage dialect-specific knowledge while maintaining the efficiency of the sparse MoE architecture.

4 Experiments

Models We evaluate five open-source LLMs, three predominantly pre-trained using English corpora (Mistral-v0.1 7B Instruct (Jiang et al., 2023), Gemma2-9B Instruct (Team, 2024), Phi-3 Medium Instruct (et al., 2024)) and two multilingual models (Llama 3.1 8B Instruct (Weerawardhena et al., 2025), Qwen 2.5 7B Instruct (Yang et al., 2025)). We conduct experiments in zero-shot and few-shot settings.

All models are fine-tuned in half-precision for 5 epochs, with a batch size of 8 and Adam optimizer. The optimal learning rate is highly dependent on the LLM as well as the technique, but varies from 1e-5 to 2e-4. Learning rate selection was based on monitoring training loss convergence during preliminary runs, with particular attention to the Mistral-v0.1 7B Instruct which required

lower learning rates (1e-5) across all proposed approaches for stable training.

All experiments are performed using two NVIDIA A100 40GB GPUs and two NVIDIA A100 80GB GPUs. We implemented the proposed approaches using the MoE-PEFT library², an LLM-Ops framework designed for high-throughput fine-tuning, evaluation and inference.³

Dataset All experiments are conducted on BESSTIE, a dataset for sentiment and sarcasm classification crawled from Google Places reviews and Reddit comments (Srirag et al., 2025a). The dataset is labeled with three varieties of English: Australian (AU), British (UK), and Indian (IN).

Variety	Subset	Train	Valid	% Pos. Sent	% Pos. Sarcasm
AU	Google	946	130	73%	7%
	Reddit	1763	241	32%	42%
IN	Google	1648	225	75%	1%
	Reddit	1686	230	25%	13%
UK	Google	1817	248	75%	0%
	Reddit	1007	138	12%	22%

Table 1: Dataset statistics for the two tasks: sentiment and sarcasm classification.

Table 1 shows dataset statistics. BESSTIE comprises 11,963 samples across the three English dialects. The training set comprises 8,867 samples, the validation set 1,212 samples, and the test set 2,523 samples. The class distribution varies significantly, with the Google-sourced data indicating high positive sentiment (about 73–75%) and low sarcasm (0–7%). In contrast, the Reddit-sourced data for each type shows substantially lower positivity (varying from 12% to 32%) and greater sarcastic levels (13–42%). Overall, Google data tends to be more positive, less sarcastic, and more verbose, while Reddit data is more neutral or negative, more sarcastic, and more concise.

We utilize the provided training set for fine-tuning all models, while the validation set is used for evaluation purposes, as the test set is not included in the publicly available version.

Although both the Google and Reddit subsets are annotated for sentiment and sarcasm, the Google subset contains very few positive sarcasm instances across all dialects (0–7%) as evident from Table 1. Due to this extreme class imbalance, we do not conduct experiments on the sarcasm classification task using the Google data.

²<https://github.com/TUDB-Labs/MoE-PEFT>

³Further implementation details, including prompt templates, are in Appendix A.

Baselines As there are no baselines with a similar objective, we compare against the following:

- Zero-shot prompting: We utilise the instruction variant of LLMs using the same evaluation setup without any additional fine-tuning.
- Few-shot prompting: We give the instruction variant of LLMs two input-output examples in the prompt to demonstrate the task before evaluation, without additional fine-tuning.
- LoRA without dialect separation (LoRA Grouping): We train LoRA adapters by grouping the dialects for a subset of the data. For each model, we train a single adapter for the sarcasm classification task and two adapters for sentiment classification, one for each source (Google, Reddit).

5 Results and Analysis

The results for sarcasm and sentiment classification in zero-shot and few-shot settings are in Table 2 and Table 3, respectively. We report task performance using accuracy and weighted F_1 score. In general, sarcasm classification seems to be the more challenging task.

5.1 Sarcasm Classification

Which model performs best? In the zero-shot setting, Gemma-2 achieves the highest average performance across all dialectal variants, with an average F_1 score of 80.79% using LoRA Separate. In the few-shot setting, Llama 3.1 with *DialectMoE* achieves the highest average F_1 score of 72.47% across all dialectal variants, demonstrating substantial improvements, particularly on UK and IN dialects. Gemma-2 using LoRA Separate maintains strong performance with an average F_1 score of 76.20%.

How consistent is performance across dialects?

Performance across dialects varies significantly. Australian English is the most challenging, with an F_1 score generally in the range of 40-75%. Indian English performs best when fine-tuned (70-80% F_1), but the baseline seems to struggle (10-40% F_1). UK English shows the highest variation in F_1 scores, ranging from 20 to 80% F_1).

How do proposed techniques compare against baselines?

Baselines generally achieve low F_1 scores. For example, the results using Mistral-7B Instruct in zero-shot (33.34% on AU, 14.67% on

UK, 9.99% on IN) are significantly lower than those of *DialectMoE* (66.29% on AU, 79.52% on UK, 81.0% on IN). This holds true for few-shot where the Mistral-7B Instruct performs worse (56.07% on AU, 25.88% on UK, 17.79% on IN) than LoRA Separate (63.82%, 71.71%, 80.27% on each dialect respectively).

Which merging technique is most effective in *DialectFusion*?

LoRA Separate almost always outperforms CAT and TIES. It achieves the highest or the near-highest accuracy in almost all cases. In the zero-shot setting, CAT shows competitive performance especially for Phi-3 on UK and AU dialects, but TIES underperforms both LoRA Separate and CAT, often yielding results closer to the baseline.

How does *DialectFusion* compare to *DialectMoE*?

DialectMoE excels for some models in zero-shot settings, achieving 5-15% F_1 improvements over *DialectFusion* methods, particularly on UK and IN dialects. *DialectFusion* (especially LoRA Separate) provides more consistent and reliable performance across both few-shot and zero-shot scenarios.

5.2 Sentiment Classification

Which model performs best?

In the zero-shot setting, *DialectMoE* achieves remarkable performance on Reddit using Mistral 7B across dialects. For Google Sentiment, *DialectFusion* with CAT merging shows the strongest performance on Phi-3 and Qwen 2.5 models, with Qwen achieving 98.46% F_1 on AU and Phi-3 reaching 99.19% on UK dialect. In the few-shot setting, Gemma 2 with LoRA Grouping achieves the best performance on UK and AU on the Reddit subset, and maintains a high F_1 score on the Google subset.

How consistent is performance across dialects?

Performance across dialects varies significantly. Indian English appears to be the most challenging dialect across sentiment tasks. On the Reddit subset, Mistral 7B shows the weakest results in zero-shot setting. For Google Sentiment, Llama 3.1 struggles across all dialects in the zero-shot setting. However, using *DialectMoE* shows substantial improvements, achieving scores above 85%. In both zero- and few-shot, UK English demonstrates the strongest and most stable overall performance, with scores typically ranging from 90-99% F_1 across different models and techniques.

Model	Technique	Reddit Sarcasm				Reddit Sentiment				Google Sentiment			
		AU	UK	IN	Avg	AU	UK	IN	Avg	AU	UK	IN	Avg
Mistral-7B	Instruct	33.34	14.67	9.99	19.33	77.51	80.23	73.58	77.11	93.85	94.74	83.60	90.73
	LoRA Group	43.63	68.38	80.27	64.09	88.38	94.20	83.44	88.67	96.92	96.32	86.52	93.25
	LoRA Sep	63.86	68.38	80.27	70.84	87.38	86.85	84.76	86.33	97.68	95.44	85.88	93.00
	CAT	43.63	68.38	80.27	64.09	85.61	88.97	83.71	86.10	97.68	96.72	84.76	93.05
	TIES	48.00	18.27	25.02	30.43	80.97	83.51	77.86	80.78	95.34	94.68	83.55	91.19
	DialectMoE	66.29	79.52	81.00	75.60	90.91	96.66	90.44	92.67	96.16	97.16	88.16	93.83
Phi-3	Instruct	72.69	44.82	58.53	58.68	86.85	89.54	83.11	86.50	94.76	95.64	86.47	92.29
	LoRA Group	57.36	70.22	77.89	68.49	87.90	94.70	85.68	89.43	94.59	96.70	86.99	92.76
	LoRA Sep	66.08	68.38	80.27	71.58	88.65	92.13	86.48	89.09	95.38	97.12	87.76	93.42
	CAT	50.33	72.06	81.81	68.07	89.25	91.52	88.16	89.64	96.14	99.19	87.55	94.29
	TIES	73.41	54.37	66.71	64.83	88.82	91.12	84.98	88.31	94.73	96.43	85.53	92.23
	DialectMoE	68.67	58.60	60.83	62.70	86.35	93.84	83.93	88.04	74.79	83.91	50.51	69.74
Qwen 2.5	Instruct	61.25	30.26	46.45	45.99	85.16	91.29	84.67	87.04	96.95	98.01	87.15	94.04
	LoRA Group	74.67	73.34	82.22	76.74	89.31	93.84	85.01	89.39	95.38	97.57	86.41	93.12
	LoRA Sep	74.67	73.34	82.22	76.74	89.31	93.84	85.01	89.39	95.38	97.57	86.41	93.12
	CAT	42.69	69.63	81.06	64.46	88.65	94.03	82.8	88.49	98.46	97.55	86.52	94.18
	TIES	59.71	70.44	76.5	68.88	85.16	93.1	86.48	88.25	94.69	97.61	85.69	92.66
	DialectMoE	73.17	81.71	83.21	79.36	90.16	96.51	88.56	91.74	94.59	98.79	86.74	93.37
Llama 3.1	Instruct	57.71	23.99	38.28	40.00	78.79	91.12	81.90	83.94	78.32	75.54	76.66	76.84
	LoRA Group	51.25	75.18	79.62	68.68	89.04	94.70	84.49	89.41	97.68	97.55	87.55	94.26
	LoRA Sep	61.02	68.38	80.27	69.89	88.84	95.38	71.77	85.33	96.89	96.72	87.89	93.83
	CAT	42.69	68.38	80.27	63.78	85.55	94.86	81.66	87.36	96.89	97.15	88.06	94.03
	TIES	69.15	42.68	61.10	57.64	80.44	91.29	84.06	85.26	96.23	95.23	84.25	91.90
	DialectMoE	70.58	78.40	85.08	78.02	91.24	96.42	88.73	92.13	96.89	97.99	88.72	94.53
Gemma 2	Instruct	54.94	18.89	22.48	32.10	84.89	90.85	87.03	87.59	86.89	88.55	84.37	86.60
	LoRA Group	72.18	81.76	83.03	78.99	90.80	94.99	89.50	91.76	96.95	98.40	86.00	93.78
	LoRA Sep	78.14	78.98	87.24	81.45	89.17	96.21	87.83	91.07	94.53	97.99	87.41	93.31
	CAT	52.43	77.84	83.95	71.41	89.4	94.70	85.55	89.88	96.17	97.18	86.01	93.12
	TIES	76.08	50.56	57.10	61.25	87.59	93.23	89.50	90.11	95.50	94.91	87.28	92.56
	DialectMoE	46.52	69.24	79.92	65.23	88.74	91.29	82.41	87.48	83.21	89.42	76.56	83.06

Table 2: F_1 scores for zero-shot experiments across different models and techniques.

In the few-shot setting, several models (Mistral-7B Instruct, Gemma 2 with *DialectMoE*) yield lower F_1 compared to zero-shot. All models show moderate stability on AU with F_1 scores typically in the 88-90% range for Reddit and 94-99% for Google. Overall, sentiment classification tasks show strong dialect robustness, with Google sentiment demonstrating superior performance where most fine-tuned models achieve F_1 scores above 95% across dialects, while Reddit sentiment shows more variation with scores typically in the 85-96% range.

How do proposed techniques compare against baselines? In the zero-shot setting, for Reddit sentiment, Mistral-7B Instruct shows the weakest results for Indian English with F_1 score of 73.58%, whereas our proposed techniques achieve signif-

icantly higher scores. For instance, *DialectMoE* demonstrates remarkable improvements, with Mistral achieving F_1 scores above 90% for all dialects. *DialectFusion* using CAT merging in Phi-3 performs better than LoRA Grouping and Instruct with by 2-5% improvements, respectively, on the Indian English dialect. For Google Sentiment, Llama 3.1 Instruct struggles across all dialects. In contrast, *DialectMoE* shows significant improvements over the baseline ($\sim 20\%$ each), demonstrating the effectiveness of our proposed approach. *DialectFusion* with CAT merging shows the strongest performance on Phi-3, reaching 99.19% on UK, representing the best results for this dialect. *DialectFusion* with CAT and TIES provides robust performance, with Qwen 2.5 CAT achieving 94.20% F_1 on UK and 83.69% F_1 on IN on Reddit sentiment, outperforming baselines by approximately $\sim 2\%$. For

Model	Technique	Reddit Sarcasm				Reddit Sentiment				Google Sentiment			
		AU	UK	IN	Avg	AU	UK	IN	Avg	AU	UK	IN	Avg
Mistral-7B	Instruct	56.07	25.88	17.79	33.25	69.36	59.61	65.69	64.89	92.55	90.29	83.33	88.72
	LoRA Grouping	58.29	63.28	61.35	60.97	81.54	81.36	73.18	78.69	96.14	95.84	83.46	91.81
	LoRA Separate	63.82	71.71	80.27	71.93	79.13	68.95	76.55	74.88	94.64	94.98	84.20	91.27
	CAT	52.92	69.74	77.67	66.78	71.93	69.58	73.10	71.54	93.90	95.08	83.98	90.99
	TIES	60.02	39.72	33.13	44.29	69.86	59.61	67.43	65.63	93.27	92.53	85.73	90.51
	DialectMoE	61.98	68.41	82.29	70.89	79.70	79.60	77.45	78.92	93.27	93.64	86.07	90.99
Phi-3	Instruct	74.76	50.87	59.48	61.70	84.89	91.12	83.24	86.42	94.04	96.06	84.93	91.68
	LoRA Grouping	68.84	68.22	74.68	70.58	83.60	91.71	79.76	85.02	94.73	97.20	85.05	92.33
	LoRA Separate	67.42	68.38	80.27	72.02	87.92	92.94	83.95	88.27	94.64	96.38	88.16	93.06
	CAT	56.07	73.05	79.82	69.65	84.89	93.10	84.05	87.35	96.17	97.61	86.99	93.59
	TIES	72.82	55.36	64.14	64.11	84.08	92.32	84.41	86.94	93.27	96.45	85.73	91.82
	DialectMoE	64.30	69.32	75.96	69.86	79.32	94.51	82.25	85.36	92.97	93.90	84.90	90.59
Qwen 2.5	Instruct	69.85	53.66	64.30	62.60	84.60	87.98	81.30	84.63	91.11	94.53	86.25	90.63
	LoRA Grouping	71.12	76.69	80.97	76.26	86.03	92.32	76.80	85.05	96.17	96.38	87.40	93.32
	LoRA Separate	70.70	80.14	80.27	77.04	90.37	92.54	83.30	88.74	96.98	95.95	86.47	93.13
	CAT	42.69	68.38	80.27	63.78	88.04	94.20	83.69	88.64	96.98	97.58	85.69	93.42
	TIES	65.03	70.22	78.25	71.17	83.06	88.67	80.58	84.10	94.76	96.06	89.50	93.44
	DialectMoE	52.29	64.33	49.62	55.41	87.42	85.63	76.19	83.08	94.73	95.58	88.10	92.80
Llama 3.1	Instruct	51.48	19.85	19.48	30.27	80.98	92.63	82.96	85.52	90.39	91.12	81.39	87.63
	LoRA Grouping	54.81	42.50	62.18	53.16	89.89	92.48	86.13	89.50	92.44	95.62	87.35	91.80
	LoRA Separate	61.58	68.38	80.27	70.08	81.16	82.60	73.50	79.09	88.85	94.77	82.40	88.67
	CAT	42.69	68.38	80.27	63.78	79.48	86.00	77.47	80.98	86.73	90.32	84.69	87.25
	TIES	70.30	43.82	49.14	54.42	78.84	90.85	85.70	85.13	91.77	95.25	84.64	90.55
	DialectMoE	53.20	80.82	83.21	72.41	74.56	84.65	70.26	76.49	94.73	96.42	87.87	93.01
Gemma 2	Instruct	72.35	31.48	39.09	47.64	86.85	91.43	85.19	87.82	91.87	93.73	88.83	91.48
	LoRA Grouping	57.14	84.80	83.21	75.05	89.64	96.21	84.73	90.19	93.85	99.19	87.40	93.48
	LoRA Separate	77.49	73.09	83.21	77.93	89.93	94.35	83.32	89.20	93.85	97.14	85.60	92.20
	CAT	45.45	81.61	87.95	71.67	90.77	94.86	85.63	90.42	95.38	97.98	85.60	92.99
	TIES	74.79	64.80	69.60	69.73	88.10	93.23	87.33	89.55	94.76	95.64	88.57	92.99
	DialectMoE	48.19	69.00	79.70	65.63	84.76	90.45	81.66	85.62	79.43	79.27	79.85	79.52

Table 3: F_1 scores for few-shot experiments across different models and techniques.

Google sentiment, Gemma 2 with LoRA Grouping in the few-shot setting demonstrates strong performance with F_1 reaching 99.99% for UK. *DialectMoE* performs considerably better for Llama 3.1, outperforming both baselines. *DialectFusion* with CAT provides robust performance on Gemma 2 and Qwen 2.5.

How does *DialectFusion* compare to *DialectMoE*? *DialectMoE* proves to be effective in zero-shot setting on both Reddit and Google sentiment subsets. For Reddit Sentiment, *DialectMoE* achieves remarkable performance across Llama 3.1, Qwen 2.5, and Mistral-7B, with Mistral-7B showing the strongest performance across dialects. For Google sentiment, *DialectMoE* yields high scores across multiple models, and Llama 3.1 showing significant improvements compared to its baselines.

In the few-shot setting, *DialectFusion* with CAT and TIES is more consistent and reliable. For Reddit sentiment, *DialectFusion* with CAT demonstrates strong performance compared to *DialectMoE*, with Qwen 2.5 CAT achieving 88.04% on AU, 94.20% on UK, and 83.69% on enIN with ~8% advantage over the latter approach. Gemma 2 shows strong performance with both CAT and TIES. In contrast, *DialectMoE* exhibits mixed effectiveness in the few-shot scenarios. For Reddit sentiment, Mistral-7B achieves as low as 77.45% on IN, below results in the zero-shot setting. For Google sentiment, *DialectFusion* with CAT provides robust performance, with Gemma 2 and Qwen 2.5 achieving high scores. Overall, *DialectMoE* excels in zero-shot settings, while *DialectFusion* provides more stable performance across both settings.

ID	Dialect	Domain	Task	Example	True	Base	LoRA	CAT	TIES	MoE
1	UK	Reddit	Sentiment	"Looks like we got a genius over here"	0	1	1	0 (Phi)	1	0 (Llama)
2	IN	Google	Sentiment	"Very famous and renowned bhaaji box shop. Quality is good. Little overpriced."	1	0	0	0	0	1 (Llama)
3	UK	Google	Sentiment	"It's a very good pub. The food is also nice, but the price is too high for the size you get, so sorry to add this comment. I just couldn't believe 7 plus service on chicken strips. It's just not worth it eating out. I am not sure how that can be improved"	1	0	0	1 (Qwen)	0	0
4	AU	Reddit	Sentiment	"Urgent care clinics might be your next best option: It emergency but for non life threatening issues, anything life threatening I'd be calling an ambulance."	0	1	1	0 (Phi)	1	1
5	AU	Reddit	Sarcasm	"They present *opinion* from those sources. But of course, opinion that aligns with one's ideology is *analysis*. As you said, you shouldn't give credit to the news source for referring to *analysis* of others. Glad we agree those two measures quoted by The Spectator and The Guardian are inflationary (or at least I think we agreed, you haven't moved past the whole media watch thing yet)."	0	1	1	0 (Gemma)	1	0
6	AU	Reddit	Sarcasm	"The Decepticons are infiltrating and taking over"	1	1	1	0	1 (Gemma)	0
7	IN	Reddit	Sarcasm	"Was she playing pub G? On a serious note tho, om Shanti."	1	0	0	0	0	0
8	IN	Reddit	Sentiment	"bro woke up to revenge the out of context use of his speech"	1	0	0	0	0	0

Table 4: Behavior of different methods across different dialects, domains, and tasks. Base denotes the Instruct variant, LoRA denotes LoRA Grouping, CAT and TIES refer to *DialectFusion*, and MoE refers to *DialectMoE*. Examples demonstrate cases where: baselines fail but proposed methods succeed (1-4), CAT vs TIES divergence (5-6), and all methods fail (7-8).

5.3 Qualitative Analysis

Next, we provide a qualitative analysis of the behavior of baselines and our proposed methods, with examples shown in Table 4.

The first three examples show the case where the baselines fail, but our proposed methods, *DialectMoE* on Qwen 2.5 and Llama 3.1 as well as *DialectFusion* with CAT merging on Phi-3 classify the sentiment correctly. Example 2 contains dialectal vocabulary ("bhaaji") and mixed sentiment. While baselines fail, *DialectMoE* on Llama 3.1 correctly classifies the sentiment to be positive (1), suggesting its dialect-specific experts better capture regional linguistic patterns and sentiment expressions. Example 3 illustrates the performance of *DialectFusion* with CAT in correctly identifying the overall positive sentiment, while all other methods fail. This demonstrates CAT's effective merging of dialect-specific knowledge, which appears particularly valuable for expressions in British English.

Examples 5-6 illustrate the case where CAT succeeds but TIES fails, and vice versa. Example 5 illustrates *DialectFusion* with CAT correctly identifying sarcasm distributed across multiple sentences with embedded quotes, whereas *DialectFusion* with TIES fails. On the other hand, Example 6 shows CAT failing on brief statements with clear cultural markers, whereas TIES correctly identifies this "pop culture" reference as sarcastic.

Example 2 demonstrates the case where *DialectMoE* performs correct classification while CAT and TIES fail. The reason may be that the routing mechanism successfully directs examples to specialized experts, whereas merging approaches fail to aggregate the distributed knowledge effectively when dialectal signals ("bhaaji") are strong. However, CAT demonstrates superiority when dialectal signals are ambiguous, as evident from Example 4. While *DialectMoE* fails across all models, CAT succeeds consistently.

Error Analysis We also present error cases (Examples 7 and 8) in Table 4 to reveal limitations of proposed approaches. In Example 7, all the models misclassify the example as non-sarcastic (0); this may be due to models focusing on surface level lexical cues such as "serious" and "Om Shanti", failing to capture the pop culture reference of the game called PubG which provides the sarcastic intent. Similarly, for the informal Indian English expression (Example 8) "bro woke up to revenge", all proposed methods, including baselines, fail to capture the positive sentiment towards defensive action.

Overall, the proposed approaches demonstrate promising advantages for dialect adaptation, with effectiveness varying across zero-shot and few-shot settings as well as task. For instance, *DialectMoE* performs best for both tasks in zero-shot, *Di-*

alectFusion using CAT merging is competitive on the sentiment classification task, especially on the Google subset, and LoRA Separate is especially effective in the few-shot setting where other approaches under-perform.

Although LoRA Separate achieves robust performance in most settings since each dialect has its own dedicated adapter trained on isolated dialectal examples without parameter-sharing constraints, this causes difficulties at inference time, particularly when inputs are from mixed dialects. In contrast, fusion-based approaches produce a single model, eliminating the need for adapter selection during inference and thus reducing deployment complexities and memory overhead. These methods better handle mixed or unknown dialectal inputs. While our results indicate that merging strategies do not consistently yield positive transfer, the competitive performance of CAT demonstrates an effective strategy for translating separately trained adapters into practical systems.

DialectMoE demonstrates positive transfer, particularly in zero-shot settings, indicating that the routing mechanism enables expert specialization by directing each token to the most suitable expert. The shared feed forward network captures universal English patterns, while the LoRA experts seem to specialize in dialect-specific features. Moreover, the gating mechanism can leverage multiple experts when inputs exhibit cross-dialectal characteristics. The observed performance degradation in few-shot settings may stem from overfitting due to limited few-shot examples, causing the router to focus on in-context learning as opposed to the generalization acquired during training. *DialectMoE* does not encounter difficulties when handling mixed-dialect inputs as there is a single adapter for multiple dialects where experts represent the differentiating factor. Although *DialectMoE* incurs routing computation costs, it activates only the top-k experts per token, resulting in more efficient inference time than full multi-adapter approaches, while maintaining dynamic adaptation capabilities.

6 Conclusion

In this paper, we evaluated the performance of five open-source LLMs across Australian, British, and Indian English dialects, and proposed two parameter-efficient frameworks based on LoRA adapters and mixture of experts to enhance their robustness. Our first approach, *DialectFusion* in-

volves training separate LoRA adapters for each dialect and merging them using two techniques, CAT and TIES. Among these, CAT demonstrates stable performance improvements across most of the LLMs. Our second approach, *DialectMoE*, based on mixture-of-experts framework, shows strong overall performance, with particularly notable gains for Qwen 2.5, LLaMA 3.1, and Mistral-7B. Experimental results indicate that both approaches consistently outperform baseline methods including instruction-tuned models and standard LoRA fine-tuning, which demonstrate the effectiveness of parameter-efficient strategies for improving dialect robustness.

Limitations

Our work has several limitations. First, *DialectMoE* does not support all the models and custom implementation is needed for extending beyond the provided models which might not be a scalable approach. Additionally, our error analysis reveals that all our methods, including *DialectMoE*, struggle possibly because of their reliance on surface level lexical cues indicating limited contextual understanding. This suggests that parameter-efficient adaptation on its own might not be sufficient for capturing culture specific intent without having a complementary mechanism for external knowledge and reasoning.

Acknowledgements

Mamta and Oana Cocarascu acknowledge the support from EPSRC (grant number EP/X04162X/1).

References

- arcee-ai. 2024. The-tome: A curated dataset for instruction-following language models. <https://huggingface.co/datasets/arcee-ai/The-Tome>. Accessed: 2025-12-10.
- Fei Ding and Baiqiao Wang. 2025. Improved supervised fine-tuning for large language models to mitigate catastrophic forgetting. *arXiv e-prints*, pages arXiv-2506.
- Abdin et al. 2024. *Phi-3 technical report: A highly capable language model locally on your phone*. Technical Report MSR-TR-2024-12, Microsoft.
- Fahim Faisal and Antonios Anastasopoulos. 2024. *Data-augmentation-based dialectal adaptation for LLMs*. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 197–208, Mexico City, Mexico. Association for Computational Linguistics.

- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. *Mistral 7b*. Preprint, arXiv:2310.06825.
- Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2025. *Natural language processing for dialects of a language: A survey*. *ACM Comput. Surv.*, 57(6).
- Dengchun Li, Yingzi Ma, Naizheng Wang, Zhiyuan Cheng, Lei Duan, Jie Zuo, Cal Yang, and Mingjie Tang. 2024. *Mixlora: Enhancing large language models fine-tuning with lora based mixture of experts*. *CoRR*, abs/2404.15159.
- Fangru Lin, Shaoguang Mao, Emanuele La Malfa, Valentin Hofmann, Adrian de Wyster, Xun Wang, Si-Qing Chen, Michael J. Wooldridge, Janet B. Pierrehumbert, and Furu Wei. 2025. *Assessing dialect fairness and robustness of large language models in reasoning tasks*. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6317–6342, Vienna, Austria. Association for Computational Linguistics.
- Yanchen Liu, William Held, and Diyi Yang. 2023. *DADA: Dialect adaptation via dynamic aggregation of linguistic rules*. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13776–13793, Singapore. Association for Computational Linguistics.
- Akshara Prabhakar, Yuanzhi Li, Karthik Narasimhan, Sham Kakade, Eran Malach, and Samy Jelassi. 2025. Lora soups: Merging loras for practical skill composition tasks. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 644–655.
- Dipankar Srirag, Aditya Joshi, Jordan Painter, and Diptesh Kanojia. 2025a. *BESSTIE: A benchmark for sentiment and sarcasm classification for varieties of English*. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8413–8429, Vienna, Austria. Association for Computational Linguistics.
- Dipankar Srirag, Nihar Ranjan Sahoo, and Aditya Joshi. 2025b. *Evaluating dialect robustness of language models via conversation understanding*. In *Proceedings of the Second Workshop on Scaling Up Multilingual & Multi-Cultural Evaluation*, pages 24–38, Abu Dhabi. Association for Computational Linguistics.
- Jiao Sun, Thibault Sellam, Elizabeth Clark, Tu Vu, Timothy Dozat, Dan Garrette, Aditya Siddhant, Jacob Eisenstein, and Sebastian Gehrmann. 2023. *Dialect-robust evaluation of generated text*. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6010–6028, Toronto, Canada. Association for Computational Linguistics.
- Gemma Team. 2024. *Gemma 2: Improving open language models at a practical size*. *CoRR*, abs/2408.00118.
- Sowmya Vajjala and Shweta Shimangaud. 2025. *Text classification in the LLM era - where do we stand?* *CoRR*, abs/2502.11830.
- Gido M. van de Ven, Nicholas Soures, and Dhireesha Kudithipudi. 2025. *Continual learning and catastrophic forgetting*, page 153–168. Elsevier.
- Sajana Weerawardhena, Paul Kassianik, Blaine Nelson, Baturay Saglam, Anu Vellore, Aman Priyanshu, Supriti Vijay, Massimo Aufiero, Arthur Goldblatt, Fraser Burch, Ed Li, Jianliang He, Dhruv Kedia, Kojin Oshiba, Zhouyan Yang, Yaron Singer, and Amin Karbasi. 2025. *Llama-3.1-foundationai-securityllm-8b-instruct technical report*. *CoRR*, abs/2508.01059.
- Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. 2023. *Ties-merging: Resolving interference when merging models*. In *Advances in Neural Information Processing Systems*, volume 36, pages 7093–7115. Curran Associates, Inc.
- An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, Junyang Lin, Kai Dang, Kexin Yang, Le Yu, Mei Li, Minmin Sun, Qin Zhu, Rui Men, Tao He, and 9 others. 2025. *Qwen2.5-1m technical report*. *CoRR*, abs/2501.15383.
- Yazhou Zhang, Chunwang Zou, Zheng Lian, Prayag Tiwari, and Jing Qin. 2025. *Sarcasmbench: Towards evaluating large language models on sarcasm understanding*. *IEEE Transactions on Affective Computing*, 16(4):2560–2578.
- Caleb Ziems, William Held, Jingfeng Yang, Jwala Dhamala, Rahul Gupta, and Diyi Yang. 2023. *Multi-VALUE: A framework for cross-dialectal English NLP*. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 744–768, Toronto, Canada. Association for Computational Linguistics.

A Implementation details

We implemented the proposed approaches using the MoE-PEFT library⁴ which is an LLMops framework designed for high-throughput finetuning, evaluation and inference. We modified the

⁴<https://github.com/TUDB-Labs/MoE-PEFT>

original implementations to account for prompt templates of various instruct models so the fine-tuning remains consistent. In case of DialectMoE, to represent number of dialects (en-AU, en-UK, en-IN), we set the number of experts to 3 for every LLM except Mistral. We assign the number of experts to 6 for sarcasm classification and 8 for sentiment classification because the lower number of experts did not result in convergence. Similarly, the learning rates were lowered to $1e-5$ from $2e-4$ for convergence (DialectMoE).

Since our experiments focus on instruction-tuned LLMs, it was important to prevent catastrophic forgetting of their instruction-following capabilities while fine-tuning them on the BESSTIE dataset (van de Ven et al., 2025). For this reason, we augment the BESSTIE dataset with FineTome-100k. The FineTome-100k dataset is a filtered version of The Tome dataset, which consisted of high quality multi-turn conversation compiled from 9 public datasets and curated using a reranker, educational value scoring and composite scoring (arcee-ai, 2024). We augment all our training datasets with 2000 samples of FineTome-100k (Ding and Wang, 2025) and also transformed each sample into a user-assistant message pair with task-specific prompts as mentioned below. Finally, we also validate that all conversations maintain proper role alternation i.e. starting with user, ending with assistant and no consecutive messages from the same role. This helps in filtering out any malformed example and also perform de-duplication.

A.1 Prompt Templates

We use two task-specific prompts as outlined in the BESSTIE paper, one for each task. For sentiment classification, we then prompt the model with: *“Generate the sentiment of the given text. 1 for positive sentiment, and 0 for negative sentiment. Do not give an explanation.”*

Similarly, for sarcasm classification, we use: *“Predict if the given text is sarcastic. 1 if the text is sarcastic, and 0 if the text is not sarcastic. Do not give an explanation.”*

Given the text and a task-specific prompt, the expected behavior of the LLM is to generate either 1 (for positive) or 0 (for negative).

Evaluation Framework for Transfer Learning between Closely Related Lects: A Case Study of Lemko

Ilia Afanasev

University of Vienna

ilia.afanasev.1997@gmail.com

Abstract

The creation of a robust evaluation methodology is one of the pivotal issues for transfer learning between closely related lects. The current study proposes to resolve this issue by concisely implementing a group of evaluation methods that enable a more systematic qualitative analysis of errata (for instance, string similarity measures to assess lemmatisation more effectively). The paper introduces a robustness score, a metric that aims to assess the stability of model performance across different datasets.

The case study is a morphosyntactic tagging of a small historical (beginning of the twentieth century) corpus of Lemko (Slavic clade, Transcarpathian area). It presents a diversity of cross-dependent tasks, made rather complex by the rich Lemko morphology, highly influenced by areal convergence processes. The tagger is a pre-trained Stanza. The study uses modern standard Ukrainian as the source language, as it is the closest to the Lemko high-resource lect.

The analysis reveals that linguistically-aware metrics improve the speed and accuracy of analysis of the errata, especially those caused by the differences between source and target lects. The key data contribution is the open-source dataset of Lemko, obtained during the tagging tasks. Future research directions include a larger-scale test that applies more models to a more extensive material.

1 Introduction

The study addresses the lack of robust methodology for evaluation of transfer learning between closely related lects¹ that can foster linguistics-aware enhancement of the models (Section 2 refers

¹The word *lect* here denotes any group of individual linguistic repertoires (including a single one, idiolect) that could undergo approximation to a single system of features that form phonetic, grammatical, and lexical systems (Campbell, 2013, p. 191). The most frequently discussed types of lects are sociolects (lects of particular social groups), dialects (small territorial lects), and standard lect (nation-wide, codified, often top-down imposed lects).

to this along with the other theoretical issues, present at study of transfer learning from high-resource to low-resource lects). The case study is the morphosyntactic tagging of Lemko², a group of small East Slavic territorial lects of the Transcarpathian region (Section 3 characterises the material of the study in more detail). The main research question is to what degree fine-grained linguistics-aware evaluation assists in the errata³ analysis.

1.1 Contribution

To answer the research question, the paper devises a framework of fine-grained linguistics-aware evaluation, adapted to transfer learning between closely related lects, described in section 4. Besides using and modifying existing metrics, including string similarity measures for lemmatisation and complete predications number for dependency parsing, the study implements robustness score, a new measure that assesses the quality of performance on the original dataset, the quality of performance on the new dataset, and the quantitative difference between them to evaluate the stability of the model. Section 5 describes the application of the pipeline to the dataset, while section 6 delves into the prospects of further enhancements to the methodology.

The other pivotal contribution is a new open-access Lemko data set with gold morphosyntactic tagging. This dataset enables further inquiries into both the lect and the method. At the same time, it provides additional information about the unique culture of Lemko at the beginning of the twentieth century that could be useful to the modern Lemko

²Lemko is a denotation preferred by native speakers compared to Lemkian: <https://uwr.edu.pl/en/lemkos-who-are-they/> (date of access: February 11, 2026)

³The term *error* here denotes a disagreement between a gold (manually checked by a human expert) tag and an automatically assigned tag. A tag can be either a label (part-of-speech, morphological category) or a sequence (lemma).

community.

2 Related Work

This section consists of three parts. It begins with an overview of current methods of low-resource lects processing. The following part provides an outline of existing computational studies of Lemko. The section ends with a discussion of existing fine-grained evaluation techniques.

2.1 Processing Low-Resource Lects

The main challenge of processing low-resource lects is the lack of training material. This is when manual tagging takes too long to carry out (for example, when financing places restrictions on a research project), but models are still unable to generalise well enough from the existing data (de Graaf et al., 2022; Bloom Ström et al., 2023; Rao and Gopinath, 2023). There are two ways to approach this problem: transfer learning and rule-based processing.

Rule-based processing has been present in low-resource NLP for a long time, and it was the most efficient before the emergence of the huge corpora that allowed statistics-based approaches to gain the edge. (Mills, 1998; Chrupała, 2006; Plisson et al., 2008; Gesmundo and Samardžić, 2012; Radziszewski, 2013). However, in low-resource settings, the rule-based approach remains a strong contender against statistical methods (Sharipov and Sobirov, 2022; Lendvai et al., 2025). In lemmatisation tasks it is possible to train the model on the corpus of higher-resource lect to detect classes of word inflection (which do not change as significantly), and then to implement new rules for a lower-resource setting, saving the pre-trained model (de Graaf et al., 2022).

Transfer learning is a process of transferring the in-domain knowledge of the model to tagging the out-of-domain data (Bengoetxea et al., 2025, p.210). The most frequent way of doing transfer learning is to perform a task zero-shot (Tebbifakhr et al., 2020), not implementing any kind of fine-tuning for the target dataset (Kim et al., 2020, p. 72). Transfer learning may include normalisation as a part of the pipeline (van der Goot et al., 2017). This allows to reduce differences between analysed lect and the high-resource neighbour, simplifying the tagging. Sometimes, small territorial lect corpora adopt this ideology completely, erasing the presentation of the features for the end user

as well (von Waldenfels et al., 2014). Currently, one of the ways to perform transfer learning is to utilise LLMs even for extremely low-resource languages (Faisal and Anastasopoulos, 2024; Liang and Levow, 2025).

2.2 Lemko in Computational Linguistics

Lemko (as most of the small territorial lects) is poorly represented in terms of computational studies. Existing works, including corpora (Rabus and Šymon, 2015), generally do not treat it as a full-fledged system, but as a part of greater continua (Scherrer and Rabus, 2017; Rabus, 2018; Scherrer and Rabus, 2019). This is a continuation of an earlier trend in linguistics in general: the preferred way of studying small territorial lects for a long time was top-down. The scholars tended to consider them primarily as subdivisions of a larger group or a standard lect, studying mostly those features that differ them from an umbrella language (Kuparinen et al., 2023). This was a *differentiating approach* (Kriuchkova, 2007, p. 31).

The trend in theoretical linguistics changed in the second half of the twentieth century. Technological advances drastically increased the volume of the recorded material; and the material on small territorial lects ceased to be fragmentary. It became possible to study these lects *integrally* (Kriuchkova, 2007, pp. 31–32) as full-fledged systems of their own (Goldin, 1990; Goldin and Kryuchkova, 2011; Hromko, 2020). This enabled bottom-up studies of the dialect continua (Kalnyn', 1973, 1992). However, Lemko has not yet been incorporated into this type of computational linguistics study. This is a research gap this article aims to close.

2.3 Fine-Grained Evaluation

The evaluation of morphosyntactic tagging conventionally incorporated simple metrics, such as the accuracy score for lemmatisation (Straka and Straková, 2017; Bergmanis and Goldwater, 2018; Anastasyev, 2020; Kanerva et al., 2021; Torre Alonso, 2022). Rare exceptions include the evaluation of part-of-speech tagger performance on out-of-vocabulary items (Scherrer and Rabus, 2017, pp. 86–87), implementing string similarity measures to assess the lemmatisers' performance (Afanasev and Lyashevskaya, 2024), and scoring the correct dependencies of the predicate (Plank et al., 2015, p. 316). Overall, the evaluation focus is not on how good of a model a tool is, but on how

well the results of its use align with gold labels. Combining qualitative and quantitative evaluation is also relatively limited (Avramidis et al., 2018, pp. 245–246) to more theoretical works (Chung and Chou, 2025).

3 Data

The section starts with a short linguistic overview of Lemko, followed by a description of the most characteristic features of the dataset itself. As the lect is low-resource and relatively poorly known, and the dataset is newly compiled, both require additional introduction.

3.1 Lemko

Lemko group of small territorial lects is in the eastern part of the greater Slavic continuum, with its closest relatives being Bojko, East Carpathian, and Upper San groups (Del Gaudio, 2017, p. 78). The historical development of Slavic lects in the Carpathian region is relatively poorly understood, and their current close grouping is geography-based rather than linguistics-based (Ševel’ov, 1979, p. 37).

The main distribution area of Lemko for a long time has been the bordering region between southern Poland, northern Slovakia, and south-western Ukraine. However, during the Soviet era, many Lemkos became victims of deportation (alongside other oppressive practices) (Magocsi, 2015, pp. 336–338). The current spread of this lect does not cover its historical territory of distribution. During the twentieth century, when researchers collected the material, the Lemkos attested that their lects and their closest relatives were spoken in the villages and towns from the westernmost border of Ukraine (river Tisa) and to the Poland-Slovakia border (river Poprad in the region of Spiš) (Nakonečna and Rudnyc’kyj, 1940, p. 31). Figure 1 shows the map, reconstructed from the data provided in the interview. This reconstruction is far from being full: the interviewee clearly lived at the north of Slovakia, thus he had only a vague understanding of Lemko dispersal to the southwest. Most probably, the north-eastern part with the settlements mentioned by the interview was the territory of Lemkos, while the land to the southwest belonged to the speakers of closely related lects (Zilyns’kyj, 1933).

Linguistically, several key features separate Lemko from neighbouring small territorial lects.



Figure 1: Map of the Transcarpathian (yellow) distribution in the beginning of the twentieth century. The blue line at the top shows the modern borders of Poland (to the north) with Slovakia (at the centre) and Ukraine (to the west), the blue line at the middle shows the modern Slovakia-Ukraine border, the blue line below shows the border that separates Hungary (to the south) from Slovakia and Ukraine. The orange dots point to the Lemko settlements, mentioned in the interviews of speakers.

These features are present at the levels of phonetics, morphology, lexis, and syntax.

As this study explores the effects of linguistic differences on transfer learning, among the features in focus are those that diverge the most from modern standard Ukrainian, because the models trained on this lect are going to be the main subject of evaluation. Thus, it is necessary to predict the possible errata to improve the interpretation of the evaluation. Nakonečna and Rudnyc’kyj (1940, pp. 23–30), Ševel’ov (1979, p. 37) and Del Gaudio (2017, pp. 76–85) provide a more detailed and general overview.

3.1.1 Phonetics

Due to the form of representation (see below in paragraph 3.2.2) the phonetic differences are relatively non-essential, and most features that are going to influence the model are in the domain of morphosyntax and lexis. However, traces of phonetic changes that separated Lemko from the other Slavic lects might still impact the behaviour of the taggers.

The most significant feature in the dataset is a rather specific pronunciation of the verb ‘to speak’. The form here is *zvápumu* instead of expected *zovorить* (Kopp et al., 2023)⁴. This is an areal, and, seemingly, almost exclusively East Slavic Transcarpathian⁵ trait: Slovak has *hovorit’* (Ze-

⁴From here and onwards, for getting Ukrainian data the study relies on *UD_Ukrainian-ParlaMint corpus* (Kopp et al., 2023).

⁵Compare, for instance, *other corpora* (Rabus and Šymon,

man, 2017)⁶, and Polish has a completely different lexical unit, *mówić* (Wróblewska, 2018)⁷. While seemingly not crucial, this peculiarity may significantly influence morphological tagging (if there are no character/subtoken-based embeddings, it is going to be an OOV item and the model is going to tag it unpredictably) and lemmatisation (if morphological tagging is incorrect, the model of lemma generation is very likely to be incorrect as well).

Another issue is the differences in infinitive endings. Nakonečna and Rudnyc'kyj (1940) provide a rather inconsistent account of *-mu/-mi* alternations. However, the speaker uses exclusively *-mi*, which is different from the standard Ukrainian *-mu*. This is going to affect the accuracy score, but should not crucially damage the quality of predictions.

3.1.2 Morphology

Among the many morphological peculiarities of Lemko, the dataset in question mostly contains unique (as compared to the other Slavic lects) nominal inflection features. They span across nouns, adjectives, and pronouns alike. Among these, the following are the most frequent in the dataset.

- The feminine instrumental singular ending is *-ом* (for instance, *кóтром* 'which-FEM.INS.SG'⁸), and not *-ою*, present in modern standard Ukrainian (see *незалежною* 'independent-FEM.INS.SG').
- The instrumental plural ending is *-ма*. The example here is *німа* 'they-INS.PL'. This form is also present in declension systems of other Slavic languages, cf. Belarusian *двума* 'two-INS.PL' (Shishkina and Lyashevskaya, 2021)⁹. However, only in Lemko it is consistent and not alternating with other endings.
- Determinative pronouns have long partially reduplicated forms, for instance, *мóмо* 'DET-N.NOM.SG'. Modern standard Ukrainian lacks this feature.

2015)

⁶From here and onwards, for getting the Slovak data the study relies on *UD_Slovak-SNK corpus*.

⁷From here and onwards, for the Polish data the study relies on *UD_Polish-PDB corpus*.

⁸Glosses given according to Comrie et al. (2008)

⁹From here and onwards, for getting the Belarusian data the study relies on *UD_Belarusian-HSE corpus*.

3.1.3 Lexis

Most lexical peculiarities within the dataset that differ it from modern standard Ukrainian are borrowings. They come mainly from neighbouring Slavic languages and mostly include functional words: *вѣльо* 'many' < Slovak *veľa* and *бáрз* < Polish *bardzo* (Nakonečna and Rudnyc'kyj, 1940, p. 30). However, there are also borrowings from other languages, for instance, Hungarian: *кi-ральфія* 'prince-GEN.SG' (< Hungarian *királyfi* 'id.' (Nakonečna and Rudnyc'kyj, 1940, p. 30)).

3.1.4 Syntax

The key feature of the Lemko syntax is the copular structures (Fontański and Chomiak, 2000, p. 141). Here, the example would be *але то ест єдна бесіда*. 'but this.SHORT-N.NOM.SG AUX one-FEM.NOM.SG language-NOM.SG'. This structure is present in the neighbouring Polish language: *To jest mój debiut* 'this.-N.NOM.SG AUX my-MASC.NOM.SG debut-NOM.SG'. Modern standard Ukrainian lacks this kind of copular structures; therefore, at all stages of morphosyntactic tagging such sentences may present issues to the pre-trained model.

3.2 Dataset

This subsection describes the fragment (Baglioni and Rigobianco, 2024, pp. 1–9) of Lemko that constitutes the dataset. It discusses its source, some basic quantitative characteristics, along with the form of representation.

3.2.1 Source

The original dataset is LA1407 (Nakonečna and Rudnyc'kyj, 1940, pp. 31–35), a set of three texts, recorded in the late 1930s from a single Lemko speaker who grew in the village of Kamienka (Lemko *Камюнка*; now a part of the Prešov Region in the north of Slovakia) and transcribed as a part of more general Transcarpathian lects survey. The data presents three layers: detailed phonetic transcription, standard-like transcription, and German translation. The unit of data is a predication (some sentences are rather long, which led to them being cut into chunks). Overall, the dataset estimates to 609 tokens. Of these 609 tokens, 5.58% are adverbs, 7.88% are adjectives, 11.99% are verbs, 15.93% are punctuation marks, 22.66% are proper and common names, and 35.96% are function words.

3.2.2 Digital Representation

The dataset is represented digitally as a .conllu-file. It contains texts in all three forms of their representation: phonetic transcription, German text, and standard-like transcription. It also brings three key modifications.

Phonetic transcription uses the IPA rendering of the original system instead of copying the latter. The metadata now provide the English translation alongside the German. For better dependency parsing, the study merges predication units into clauses, where possible. These changes facilitate a better understanding of the material by both scholars and tools.

4 Method

This section starts with an experiment workflow outline. The subsequent part describes the model used for dataset tagging. The section ends with a discussion of the fine-grained evaluation metrics.

4.1 Experiment Workflow

The study represents three consecutive stages of tagging the dataset (Afanasev, 2026): part-of-speech and morphological tagging, lemmatisation, and dependency parsing. Between each stage, the research uses a *human-in-the-loop* (Jiang et al., 2024; Umphrey et al., 2024; Verma et al., 2025) approach. After conducting each stage of tagging and before each stage of evaluation, a human scholar checks the result of the tagging stage to create a gold version of the data. This approach facilitates the evaluation and, crucially, a more efficient tagging during the next stage (Anastasyev, 2020; Milintsevich and Sirts, 2021).

The part-of-speech and morphological tagging study is based on two experiments. The first involves tagging a pre-tokenized LA1407 corpus with Stanza, the model that undergoes the evaluation through all of the tasks (see subsection 4.2), with the primary aim of obtaining labels for the dataset. The second experiment compares different models on different materials and is designed to evaluate the newly introduced metrics (see subsection 4.3). For lemmatisation and dependency parsing, the study conducts two further experiments: the first uses manually corrected tagging from the preceding stages, and the second starts from raw text.

4.2 Model

The key utilised model is Stanza (Qi et al., 2020) pre-trained on the Ukrainian language (in terms of standard lects, probably the closest relative of Lemko (Ševel'ov, 1979, p. 37)). It is computationally effective, running even on relatively slow CPUs, and achieving over 90% accuracy on its test dataset. As the study focuses on evaluation and analysis, it is the only model implemented for all tasks. While Generative AI models could have provided better results, their zero-shot use is not always as effective; and research risks becoming irreproducible after a very short time (de Wynter, 2025).

As the study introduces a new metric, it is crucial to implement data and method cross-evaluation techniques (Afanasev, 2024). The model used for comparison is *stanzatagger*¹⁰, a modification of Stanza that employs character-based embeddings and is better adapted to lower-resource settings (Scherrer, 2021). The training data are, as with Stanza, *UD_Ukrainian-IU*¹¹. The training configuration follows the guideline from the project repository. The material used to assess the efficiency of the trained model includes the test subset of *UD_Ukrainian-IU*, the test subset of *UD_Ukrainian-ParlaMint* (Kopp et al., 2023), and LA1407. This setup ensures diversity of material and enables a more robust evaluation of both the model in comparison with Stanza and the newly introduced metric.

4.3 Metrics

The study implements three types of metrics. The first group facilitates a more effective comparison between the model performance on the source lect (standard Ukrainian) and the target lect (Lemko). The purpose of the second type is to provide a more linguistically-aware picture that enables the assessment of transfer learning. The third type implements fine-grained measurements, used for specific subcategories within the data.

4.3.1 Standard Metrics

To evaluate part-of-speech and morphological tagging, the study uses macro-F1 score, a traditional metric in the field (Scherrer and Rabus, 2019; Qi et al., 2020). It also implements exact match

¹⁰The code is available at the [GitHub repository](#).

¹¹The corpus is available at the [Universal Dependencies repository](#).



Figure 2: Geometric value of robustness between an ideal (top) and a result (bottom) vectors.

(overall accuracy score of part-of-speech and morphological tagging). For assessing the lemmatisation quality, the traditional metric is the accuracy score. To estimate the efficiency of dependency parsing, the study uses Unlabelled Attachment Score (UAS; how many elements have a correct head, i. e., the element from which they depend syntactically) and Labelled Attachment Score (LAS; how many elements have a correct head and a correct dependency type).

4.3.2 Quality of Transfer

The key metric to interpret the quality of transfer learning is robustness. The robustness in its core is a cosine similarity measure between the vectors of ideal transfer and result. For both vectors, the starting point is (0; in-domain accuracy score). In case of Ukrainian-trained Stanza in the task of part-of-speech tagging it is (0; 97.52). The end point of an ideal vector is (END; in-domain accuracy score). END here is a researcher-chosen number that is higher than 0. This study uses 25¹². The end point of the result vector is (END; out-of-domain accuracy score). If the out-of-domain accuracy score is higher or equal to the in-domain accuracy score, the value of robustness is positive, otherwise it is negative. The study uses robustness instead of simple comparison, because it gives better scores to the models that support stable and high level of performance in both in-domain and out-of-domain datasets. Figure 2 represents the robustness graphically.

When performing transfer learning, it is crucial to recognise not only how accurate the model is, but also how well it grasps the concept of what it has to do, and how dramatic its fall in quality on out-of-domain data is. The metrics for this task are string similarity measures: Levenshtein distance (Levenshtein, 1966) and Jaro-Winkler distance (Winkler, 1990). These metrics are very helpful in analysing lemmatisation of non-standard lects (Afanasev and Lyashevskaya, 2024), providing a clearer picture in case of non-standard inflection in the dataset.

¹²Experiments showed that if END is significantly smaller than 25, the metric loses sensitivity, if the fall in accuracy is too big, while if the END is significantly higher than 25, the metric loses sensitivity, if the fall in accuracy is rather small

For evaluating dependency parsing, the study adds three metrics: tree edit distance, Unlabelled and Labelled Complete Predications (TED, UCP and LCP). TED is a number of edits required to get a gold tree from a predicted one. UCP is a share of correctly detected dependencies of the verbal predicate. LCP is its stricter version that scores a share of correctly detected relations of the verbal predicate. This study slightly modifies the metric by Plank et al. (2015, p. 316). Unlike the original one, it fines the model for the generated dependencies that are not part of the gold dataset.

4.3.3 Fine-Grained Metrics

To get a more detailed view of the results, it is possible to evaluate the performance of the model by groups of labels. In case of part-of-speech and morphological tagging, these are parts of speech. Dependency parsing may use relationship labels. As Stanza uses a sequence-to-sequence model for lemmatisation, the word inflection classes are unavailable; the study opts for using parts of speech.

5 Results and Discussion

The section begins with a cross-evaluation of the robustness score. It provides an initial overview of the results and addresses the issues that arise during part-of-speech and morphological tagging, drawing on fine-grained metrics. The following subsection discusses the advantages of using string similarity measures as evaluation metrics for lemmatisation. Finally, based on the dependency parsing results, the study delves into linguistic aspects of transfer learning.

5.1 Cross-Evaluation of Robustness Score

Before analysing the performance of Stanza, it is crucial to place the robustness score in context. Table 1 presents the values of this metric for different datasets and compares them with a simple difference measure, defined as the subtraction of the model performance on the new dataset from its performance on the test subset of the training dataset.

Robustness score does not produce substantially different results from the simple subtraction measure. Both metrics highlight the key contrast between the two models: while *stanzatagger* not only maintains but slightly improves its tagging quality from *UD_Ukrainian-IU* to the test subset of Kopp et al. (2023), performance on LA1407 remains comparably poor (*stanzatagger* lags slightly

Dataset (model)	PoS	Exact match	Robustness (PoS)	Difference (PoS)	Robustness (Exact match)	Difference (Exact match)
UD_Ukrainian-IU (Stanza)	97.52	92.07	–	–	–	–
UD_Ukrainian-ParlaMint (Stanza)	89.38	83.59	-0.95	-8.14	-0.95	-8.48
LA1407 (Stanza)	66.78	55.99	-0.63	-30.74	-0.56	-36.08
UD_Ukrainian-IU (stanzatagger)	95.54	76.58	–	–	–	–
UD_Ukrainian-ParlaMint (stanzatagger)	96.06	76.59	1	0.52	1	0.01
LA1407 (stanzatagger)	62.89	40.56	-0.61	-33.17	-0.57	-36.02

Table 1: The results of tagging the test subsets of *UD_Ukrainian-IU*, Kopp et al. (2023) and LA1407. All the metrics, except for robustness and difference performance of the model on standard Ukrainian and Lemko, are in per cent values. Rounding is to the second digit after zero. Best values are in **bold**.

behind in PoS, whereas Stanza marginally underperforms in exact match). The main difference is that the robustness score smooths smaller discrepancies between the models, providing a clearer comparative picture than the simple difference.

5.2 Stanza Results

The overall results of the Stanza performance are rather poor. Taggers lose significantly in their accuracy, when compared to the initial dataset. Table 2 shows the results.

Lemmatisation and UAS demonstrate the best robustness scores, while exact match demonstrates the lowest one. Lemmatisation witnesses a relatively small fall; while the performance of UAS is surprisingly stable, given the initial low score of the system. Overall, robustness shows the difference between the easier tasks in the pipeline (lemmatisation with gold tags and head marking) and the more complicated ones (producing an exact match of part-of-speech and morphological tags), highlighting the spots for further enhancements. The simple difference metric between model performance on standard Ukrainian and on Lemko captures the overall drop in accuracy scores but misses important nuances: for instance, it fails to show how well the lemmatiser and part-of-speech tagger perform on the test dataset.

5.3 Fine-Grained Evaluation of Morphological Tagging

The robustness of morphological tagging placed it among the hardest tasks to transfer. To further

discover the issue, the study performs a by-part-of-speech evaluation. Figure 3 shows the results.

Accuracy score comparison (part-of-speech and morphological tagging, Stanza)

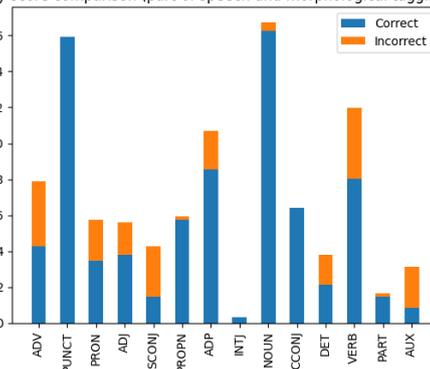


Figure 3: A by-part-of-speech evaluation of exact match results. Y-axis shows the share of part-of-speech in the dataset, the blue part of bars – the share of correctly tagged items of this part-of-speech, the orange part – the share of incorrectly tagged items of this part-of-speech. All the values are in %, rounded to the second digit.

As the figure demonstrates, the model struggles heavily with what in theoretical linguistics is sometimes called *Satellite Cluster B* (Reid, 2011, pp. 1107–1108) – verb and its adjacents. The results are especially problematic for the verb itself, adverbs, and auxiliaries. Here, the main issue is syntactical: the presence of copular sentences confuses the model. This leads to incorrect tagging of the auxiliaries. In addition, a lot of function words, and especially adverbs, like *багз* ‘very’ are loanwords from other Slavic languages. They get an X tag from the model, which may be true for modern

Metric	Standard Ukrainian	Lemko	Robustness	Difference
PoS macro-F1	97.52	66.78	-0.63	-30.74
Feats macro-F1	93.11	71.00	-0.74	-22.11
Exact match	92.07	55.99	-0.56	-36.08
Lemmatisation accuracy score (gold tagging)	96.72	75.7	-0.77	-21.02
UAS (gold tagging)	85.87	67.65	-0.77	-18.22
LAS (gold tagging)	82.77	52.71	-0.6	-30.06

Table 2: The results of tagging *LA1407* with Stanza (Qi et al., 2020). All the metrics (except for robustness and difference between performance of the model on standard Ukrainian and Lemko) are in per cent values. Rounding is to the second digit after zero. Best values are in **bold**, matched best values are in *italics*.

standard Ukrainian, but most certainly is incorrect for Lemko. Verbs are the most enigmatic case, as most errors in their tagging come from Aspect and Tense. The model often confuses the imperfective verbs with the perfective ones, as the imperfectives often took on iterative sense, cf. *полю́бят* ‘love.IPFV-PRES.3PL’.

5.4 Lemmatisation

Lemmatisation was a comparatively easier task where Stanza demonstrated robustness. To determine the principal contribution, it is crucial to perform a more in-detail view. Figure 3 shows the results.

Gold tagging seems to boost the accuracy score, but efficiency of tagging almost does not increase for string similarity measures, especially for Jaro-Winkler distance. This shows that the model generally grasps the concept of lemmatisation rather well. Still, the decrease in accuracy score is significant and requires a more thorough investigation, demonstrated in Figure 4.

Once again, the most complicated task for the model is to process verbs; the accuracy score is close to zero. On the contrary, string similarity measures show very good results (Levenshtein distance of 1.15 means the average error of 1.15 symbols, and Jaro-Winkler distance is 0.92). The inspection of the results shows that the most frequent (and almost the only) error is the wrong ending, *-mu* instead of *-mi*. This shows the value of fine-tuning or class-based lemmatisation: with a good rule system, this would not be an issue. However, the lemmatiser of Stanza is a sequence-to-sequence model, unable to precisely transfer its understanding of lemmatisation to new material.

One equally difficult but less problematic case is the auxiliaries (accuracy score of 15.79%, average Levenshtein distance of 2.88 and average Jaro-

Accuracy score comparison (lemmatisation, Stanza, gold tagging)

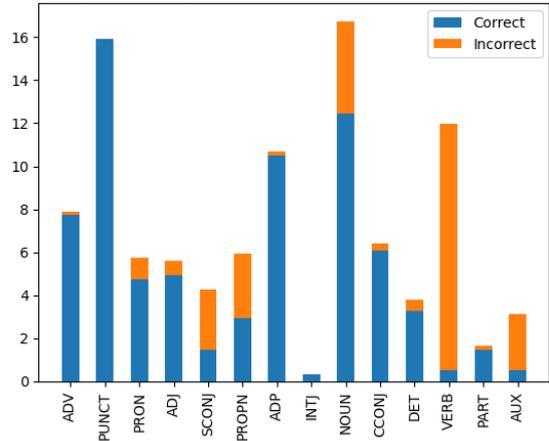


Figure 4: A by-part-of-speech evaluation of lemmatisation results. Y-axis shows the share of part-of-speech in the dataset, the blue part of bars – the share of correctly lemmatised items of this part-of-speech, the orange part – the share of incorrectly lemmatised items of this part-of-speech. All the values are in %, rounded to the second digit.

Winkler distance of 0.56). The issue at stake here is very similar to the part-of-speech and morphological tagging. The model just does not know what to do with these items, completely alien to the original dataset. They get a very random tagging that has nothing to do with their lemmatisation.

5.5 Dependency Parsing

The evaluation of dependency parsing indicates similar issues as the other types of tagging. Table 4 shows the results.

TED of 9 seems to indicate a significant issue, however, this is mostly due to the presence of large sentences that span 60 tokens. More problematic is the contrast between assignment scores (UAS and LAS) and complete predications (UCP and LCP). The model, once again, fails to correctly identify

Metric	Gold tagging	Silver tagging
Accuracy score	75.7	70.11
Levenshtein distance	1.44	1.6
Jaro-Winkler distance	0.87	0.87

Table 3: Lemmatisation evaluation results, rounded to the second digit.

Metric	Result
UAS	67.65
LAS	52.71
TED	9
UCP	47.18
LCP	35.57

Table 4: Dependency parsing evaluation results, rounded to the second digit. UAS, LAS, UCP and LCP values are percentages, TED value is an absolute value.

some Satellite Clusters B, dragging down its performance. This highlights the necessity of introducing copular structures (probably, by introducing the neighbouring languages material (Scherrer and Rabus, 2019)) into the dataset for further tagging.

5.6 Discussion

All metrics identify a single issue at the core of all the Stanza downfalls: clause structure, specifically, copular sentences, strongly characteristic of Lemko, and quite frequent in Carpathian linguistic area in general but very rarely represented in modern standard Ukrainian. The model is unable to identify the relations between tokens and, therefore, perceives the tokens themselves wrongly at the morphological level.

Lemmatisation suffers the least, because it is relatively syntax-independent, especially with lects being relatively close in terms of vocabulary. In addition, it probably is the biggest beneficiary of human-in-the-loop: correctly assigned morphology tags aid the lemmatisation the most. For instance, *ні́ма* 'PRON.3PL.INS' without gold tagging gets **Kima* as lemma, instead of intended *він* 'PRON.MASC.3SG.NOM'. Gold morphological tagging resolves the issue.

Still, the predications are clearly the weakest point (as very low values of UCP and LCP show), even with gold morphological tagging and lemmata. The key issue here is the poor identification of the lexical verb, the attractor of Satellite Cluster B. Further study needs to address the issue.

6 Conclusion

The study demonstrates that more fine-grained evaluation techniques are better at providing a concise summary of the errata that models make. This includes explaining anomalies; for instance, string similarity measures highlight the overall high quality of lemmatisation despite a low accuracy score. The metrics also help to identify key features that separate a target lect from a source lect, and thus to prepare better post-processing or fine-tuning techniques, fostering a more robust transfer learning.

The paper contributes to the study and representation of low-resource small territorial lects by making an open-access dataset of Lemko from the beginning of the twentieth century. This dataset contains more than 600 tokens with UPOS, UFeats, lemmata, and dependency parsing tags, and it is possible to use it for further experiments with other models.

Future research directions include both enriching the material and developing the metrics. There is a clear need to extend the data to other Lemko material and other Transcarpathian lects that possess their own unique features. For a more transparent demonstration of the robust evaluation advantages, it is utmost to include more models to the comparison, including region-specific ones (Scherrer and Rabus, 2019). The crucial step would be to develop a computationally effective language-aware model (Chung and Chou, 2025).

Limitations

LA1407 (Nakonečna and Rudnyc'kyj, 1940, 31–37), the main material of the research, does not represent Lemko (and Transcarpathian lects in general) of the 1920s-1930s in its entirety; a lot of material is in the process of digitisation. In addition, LA1407 represents only one speaker of Lemko, which can affect the distributions.

Ethical Considerations

The data had been published in printed form and available for research purposes for fifty to ninety

years by the time this article was written. Still, I anonymise the metatagging, where possible, masking the names of the speakers, to compensate for possible ethics violations that could have happened at the time of material collection.

The data themselves can contain slight mentions of xenophobic behaviour and religious (mostly, Christian) imagery. Discretion is advised.

Disclosure of Generative AI use

This study does not use Generative AI (in the modern colloquial meaning: the decoder models with more than a billion parameters trained on high-resource corpora; Stanza (Qi et al., 2020) is also a generative AI, but it performs on a much lesser scale, locally run and reproducible) in the research process. During the process of editing the authors utilised Generative AI (Grammarly) to polish the phrasing of the parts of the work where the authors felt that their non-native knowledge of English was not sufficient to produce grammatically and/or stylistically correct sentences. However, the written text is not a product of AI generation.

Acknowledgements

I would like to acknowledge all of the speakers, whose speech presents the recordings of the studied small territorial lects, the scholars who produced the initial transcription, as well as the research groups who produced the revised transcriptions. I also owe special thanks to Olga Fedorivna Mygolynets (ukr. Ольга Федорівна Миголінець, University of Uzhhorod), who greatly helped me with the understanding of transcription systems and phonetics of the lects analysed.

References

Iliia Afanasev. 2024. *The Cross-Evaluation Crux for Computational Phylogenetic Linguistics*, pages 75–89. Springer, Cham.

Iliia Afanasev. 2026. Evaluation framework for transfer learning between closely related lects: A case study of lemko – supplementary material.

Iliia Afanasev and Olga Lyashevskaya. 2024. *Chapter 2 String Similarity Measures for Evaluating the Lemmatisation in Old Church Slavonic*, pages 13 – 35. Brill, Leiden, The Netherlands.

Dan Anastasyev. 2020. Exploring pretrained models for joint morpho-syntactic parsing of Russian. In *Computational Linguistics and Intellectual Technologies:*

Proceedings of the International Conference “Dialogue 2020”, pages 1–12. Moscow.

Eleftherios Avramidis, Vivien Macketanz, Arle Lommel, and Hans Uszkoreit. 2018. *Fine-grained evaluation of quality estimation for machine translation based on a linguistically motivated test suite*. In *Proceedings of the AMTA 2018 Workshop on Translation Quality Estimation and Automatic Post-Editing*, pages 243–248, Boston, MA. Association for Machine Translation in the Americas.

Daniele Baglioni and Luca Rigobianco. 2024. *Chapter 1 Rethinking Fragmentariness and Reconstruction: An Introduction*, pages 1 – 25. Brill, Leiden, The Netherlands.

Jaione Bengoetxea, Mikel Zubillaga, Ekhi Azurmendi, Maite Heredia, Julen Etxaniz, Markel Ferro, and Jeremy Barnes. 2025. *HiTZ at VarDial 2025 NorSID: Overcoming data scarcity with language transfer and automatic data annotation*. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 209–219, Abu Dhabi, UAE. Association for Computational Linguistics.

Toms Bergmanis and Sharon Goldwater. 2018. *Context sensitive neural lemmatization with Lematus*. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1391–1400, New Orleans, Louisiana. Association for Computational Linguistics.

Eva-Marie Bloom Ström, Onelisa Slater, Aron Zahran, Aleksandrs Berdicevskis, and Anne Schumacher. 2023. *Preparing a corpus of spoken Xhosa*. In *Proceedings of the 2023 CLASP Conference on Learning with Small Data (LSD)*, pages 62–67, Gothenburg, Sweden. Association for Computational Linguistics.

Lyle Campbell. 2013. *Historical Linguistics: An Introduction*, ned - new edition, 3 edition. Edinburgh University Press.

Grzegorz Chrupała. 2006. Simple data-driven context-sensitive lemmatization. *Proces. del Leng. Natural*, 37:121–137.

Meng-Hsuan Chung and Chao-Ting Tim Chou. 2025. Climbing towards the nlu of the universal reading of shei ‘who’. *Concentric*, 51(2):303–348.

Bernard Comrie, Martin Haspelmath, and Balthasar Bickel. 2008. *The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses*. <https://www.eva.mpg.de/lingua/resources/glossing-rules.php>. Department of Linguistics of the Max Planck Institute for Evolutionary Anthropology and the Department of Linguistics of the University of Leipzig.

- Evelien de Graaf, Silvia Stopponi, Jasper K. Bos, Saskia Peels-Matthey, and Malvina Nissim. 2022. [AG-ILE: The first lemmatizer for Ancient Greek inscriptions](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5334–5344, Marseille, France. European Language Resources Association.
- Adrian de Wynter. 2025. [Awes, laws, and flaws from today’s LLM research](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12834–12854, Vienna, Austria. Association for Computational Linguistics.
- Salvatore Del Gaudio. 2017. *An Introduction to Ukrainian Dialectology*. Peter Lang Verlag, Berlin, Germany.
- Fahim Faisal and Antonios Anastasopoulos. 2024. [Data-augmentation-based dialectal adaptation for LLMs](#). In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 197–208, Mexico City, Mexico. Association for Computational Linguistics.
- Henryk Fontański and Mirosława Chomiak. 2000. *Gramatyka języka lemковского. Śląsk, Katowice*.
- Andrea Gesmundo and Tanja Samardžić. 2012. [Lemmatization as a tagging task](#). In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 368–372, Jeju Island, Korea. Association for Computational Linguistics.
- V. E. Goldin. 1990. K projektu tekstovogo dialektologicheskogo podfonda mashinnogo fonda russkogo jazyka [on the project of the textual dialectological sub-fund of the machine fund of the russian language]. In *Materialy III Vsesojuznoj konferencii po sozdaniju Mashinnogo fonda russkogo jazyka [Materials of the 3rd All-Union Conference on the Creation of the Machine Fund of the Russian Language]*, pages 92–103, Moscow. Izd-vo Moskovskogo universiteta.
- V. E. Goldin and O. Yu. Kryuchkova. 2011. Korpus russkoi dialektnoi rechi: kontsepsiia i parametry otsenki [Corpus of Russian Dialectal Speech: Concept and Evaluation Parameters]. In *Komp’uternaia lingvistika i intellektual’nye tekhnologii : Materialy ezhegodnoi Mezhdunarodnoi konferentsii, Bekasovo, 25–29 maia 2011 goda [Computational Linguistics and Intellectual Technologies: Proceedings of the Annual International Conference, Bekasovo, May 25–29, 2011]*, volume 10, pages 359–367, Moscow. Russian State University for the Humanities.
- Tetiana Vasylyvna Hromko. 2020. Stanovlennia monohovirkovoi deskryptsii u vitchyznianomu movoznavstvi (kinets’ xix – 40-i roky xx st.) [formation of monographic description in domestic linguistics (end of the xix – 40s of the xx century)]. In M. Pantjuk, A. Dushnyi, and I. Zymomria, editors, *Aktual’ni pytannia humanitarnykh nauk: mizhvuziv’skyj zbirnyk naukovykh prats’ molodykh vchenykh Drohobys’koho derzhavnoho pedahohichnoho universytetu imeni Ivana Franka [Current Issues of the Humanities: Interuniversity Collection of Scientific Works of Young Scientists of the Drohobych Ivan Franko State Pedagogical University]*, Vypusk 34, Tom 2, pages 118–123. Vydavnychi dim “Hel’vetyka”, Drohobych.
- Hang Jiang, Xiajie Zhang, Robert Mahari, Daniel Kessler, Eric Ma, Tal August, Irene Li, Alex Pentland, Yoon Kim, Deb Roy, and Jad Kabbara. 2024. [Leveraging large language models for learning complex legal concepts through storytelling](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7194–7219, Bangkok, Thailand. Association for Computational Linguistics.
- L. È. Kalnyn’. 1973. *Opyt modelirovanija sistemy ukrainskogo dialektnoho jazyka [An Attempt at Modeling the System of the Ukrainian Dialectal Language]*. Nauka, Moscow.
- L. È. Kalnyn’. 1992. Foneticheskij stroj odnogo gukul’skogo govora [the phonetic system of a hutul dialect]. In *Issledovanija po slavjanskoj dialektologii. [Vyp.] I: Karpato-ukrainskie dialekty [Studies in Slavic Dialectology. [Issue] 1: Carpathian-Ukrainian Dialects]*. Nauka, Moscow.
- Jenna Kanerva, Filip Ginter, and Tapio Salakoski. 2021. [Universal Lemmatizer: A sequence-to-sequence model for lemmatizing Universal Dependencies treebanks](#). *Natural Language Engineering*, 27(5):545–574.
- Hwichan Kim, Toshio Hirasawa, and Mamoru Komachi. 2020. [Zero-shot North Korean to English neural machine translation by character tokenization and phoneme decomposition](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 72–78, Online. Association for Computational Linguistics.
- Matyáš Kopp, Anna Kryvenko, and Andriana Rii. 2023. [Ukrainian parliamentary corpus ParlaMint-UA 4.0.1](#). Slovenian language resource repository CLARIN.SI.
- O. Ju. Kriuchkova. 2007. Élektronnyj korpus russkoj dialektnoj rechi i principy ego razmetki [electronic corpus of russian dialect speech and the principles of its marking]. *Izvestija Saratovskogo universiteta. Serija: Filologija. Žurnalistika [Proceedings of Saratov University. Series: Philology. Journalism]*, 7(1):51–55.
- Olli Kuperinen, Aleksandra Miletić, and Yves Scherrer. 2023. [Dialect-to-standard normalization: A large-scale multilingual evaluation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13814–13828, Singapore. Association for Computational Linguistics.

- Piroska Lendvai, Uwe Reichel, Anna Jouravel, Achim Rabus, and Elena Renje. 2025. [Retrieval of parallelizable texts across Church Slavic variants](#). In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 105–114, Abu Dhabi, UAE. Association for Computational Linguistics.
- Vladimir Iosifovich Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710. Doklady Akademii Nauk SSSR, V163 No4 845-848 1965.
- Siyu Liang and Gina-Anne Levow. 2025. [Breaking the transcription bottleneck: Fine-tuning ASR models for extremely low-resource fieldwork languages](#). In *Proceedings of the Fourth Workshop on NLP Applications to Field Linguistics*, pages 26–37, Vienna, Austria. Association for Computational Linguistics.
- Paul R Magocsi. 2015. *With their backs to the mountains : a history of Carpathian Rus' and Carpatho-Rusyns*. Central European University Press., Budapest .:
- Kirill Milintsevich and Kairit Sirts. 2021. [Enhancing sequence-to-sequence neural lemmatization with external resources](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3112–3122, Online. Association for Computational Linguistics.
- Jon Mills. 1998. [Lemmatisation of the corpus of Cornish](#). In *Workshop on Language Resources for European Minority Languages, LREC First International Conference on Language Resources and Evaluation*, pages 1–6, Granada, Spain.
- Hanna Nakonečná and Jaroslav Bohdan Rudnyc'kyj. 1940. *Ukrainische Mundarten : Südkarpatoukrainisch ; (Lemkisch, Bojkisch und Huzulisch) [Ukrainian dialects: South Carpathian Ukrainian; Lemkian, Bojkian and Huzulian]*. Arbeiten aus dem Institut für Lautforschung an der Universität Berlin ; 9. Otto Harrassowitz, Berlin.
- Barbara Plank, Héctor Martínez Alonso, Željko Agić, Danijela Merkle, and Anders Søgaard. 2015. [Do dependency parsing metrics correlate with human judgments?](#) In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 315–320, Beijing, China. Association for Computational Linguistics.
- Joël Plisson, Nada Lavrac, Dunja Mladenčić, and Tomaž Erjavec. 2008. Ripple down rule learning for automated word lemmatisation. *AI Commun.*, 21:15–26.
- 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, Online. Association for Computational Linguistics.
- A. Rabus. 2018. [Obrazovanie prošedšego vremeni v raznovidnostjach karpatorusinskogo: kvantitativnyj analiz \[the formation of the past tense in varieties of carpathian rusyn: A quantitative analysis\]](#). In Kvetoslava Koporova, editor, *20 rokov v'ysokoškolskej rusynistiky na Slovensku [20 Years of University Rusyn Studies in Slovakia]*, pages 139–151. Prešovská univerzita v Prešove, Prešov.
- A. Rabus and A. Šymon. 2015. [Na nových putjach issli-dovanja rusyns'kých dialektu. korpus rozhovornoho rusyns'koho jazýka \[on new paths of rusyn dialect research. corpus of spoken rusyn language\]](#). In Kvetoslava Koporová, editor, *Rusyn'skyj literaturnyj jazyk na Slovensku. 20 rokov kodifikaciji [The Rusyn Literary Language in Slovakia. 20 Years of Codification]*, pages 40–54. Prešov University Publishing, Prešov.
- Adam Radziszewski. 2013. [Learning to lemmatise Polish noun phrases](#). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 701–709, Sofia, Bulgaria. Association for Computational Linguistics.
- Amit Rao and Kanchi Gopinath. 2023. [A Sanskrit grammar-based approach to identify and address gaps in Google Translate's Sanskrit-English zero-shot NMT](#). In *Proceedings of the 2023 CLASP Conference on Learning with Small Data (LSD)*, pages 141–166, Gothenburg, Sweden. Association for Computational Linguistics.
- Wallis Reid. 2011. [The communicative function of English verb number](#). *Natural Language & Linguistic Theory*, 29(4):1087–1146.
- Yves Scherrer. 2021. [Adaptation of Morphosyntactic Taggers](#), page 138–166. Studies in Natural Language Processing. Cambridge University Press.
- Yves Scherrer and Achim Rabus. 2017. [Multi-source morphosyntactic tagging for spoken Rusyn](#). In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 84–92, Valencia, Spain. Association for Computational Linguistics.
- Yves Scherrer and Achim Rabus. 2019. [Neural morphosyntactic tagging for rusyn](#). *Natural Language Engineering*, 25(5):633–650.
- Maksud Sharipov and Ogabek Sobirov. 2022. [Development of a rule-based lemmatization algorithm through finite state machine for uzbek language](#). *CoRR*, abs/2210.16006.
- Yana Shishkina and Olga Lyashevskaya. 2021. [Sculpting enhanced dependencies for belarusian](#). In *Analysis of Images, Social Networks and Texts: 10th International Conference, AIST 2021, Tbilisi, Georgia, December 16–18, 2021, Revised Selected Papers*, page 137–147, Berlin, Heidelberg. Springer-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.
- Amirhossein Tebbifakhr, Matteo Negri, and Marco Turchi. 2020. [Machine-oriented NMT adaptation for zero-shot NLP tasks: Comparing the usefulness of close and distant languages](#). In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 36–46, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).
- Roberto Torre Alonso. 2022. [Automatic lemmatization of Old English class III strong verbs \(L-Y\) with ALOEV3](#). *Journal of English Studies*, 20:237–266.
- Ray Umphrey, Jesse Roberts, and Lindsey Roberts. 2024. [Investigating expert-in-the-loop LLM discourse patterns for ancient intertextual analysis](#). In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 31–40, Miami, USA. Association for Computational Linguistics.
- Rob van der Goot, Barbara Plank, and Malvina Nissim. 2017. [To normalize, or not to normalize: The impact of normalization on part-of-speech tagging](#). In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 31–39, Copenhagen, Denmark. Association for Computational Linguistics.
- Kanishk Verma, Sri Balaaji Natarajan Kalaivendan, Arefeh Kazemi, Joachim Wagner, Darragh McCashin, Isobel Walsh, Sayani Basak, Sinan Ascı, Yelena Cherkasova, Alexandrous Poullis, James O’Higgins Norman, Rebecca Umbach, Tijana Milošević, and Brian Davis. 2025. [BullyBench: Youth & experts-in-the-loop framework for intrinsic and extrinsic cyberbullying NLP benchmarking](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 2172–2208, Suzhou (China). Association for Computational Linguistics.
- Ruprecht von Waldenfels, Michael Daniel, and Nina Dobrushina. 2014. Why standard orthography? building the ustya river basin corpus, an online corpus of a russian dialect. In *Komp’juternaja lingvistika i intelektual’nye tehnologii: Po materialam ežegodnoj Meždunarodnoj konferencii «Dialog» (Bekasovo, 4 — 8 ijunja 2014 g.) [Computational Linguistics and Intellectual Technologies: Based on the materials of the Annual International Conference "Dialog" (Bekasovo, June 4-8, 2014)]*, volume 13 (20), Moscow. Izd-vo RGGU.
- William E. Winkler. 1990. String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In *Proceedings of the Section on Survey Research*, pages 354–359.
- Alina Wróblewska. 2018. Extended and enhanced polish dependency bank in universal dependencies format. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 173–182. Association for Computational Linguistics.
- Daniel Zeman. 2017. Slovak Dependency Treebank in Universal Dependencies. *Jazykovedný časopis / Journal of Linguistics*, 68(2):385–395.
- Ivan M Zilyns’kyj. 1933. *Karta ukrains’kych hovoriv : z pojasnennjamy ; mirylo 1:4.000.000*. Praci Ukraïns’koho Naukovoho Institutu 14. Ukraïns’kyj Naukovyj Instytut, Warszawa.
- Jurij Volodymyrovych Ševel’ov. 1979. *A historical phonology of the Ukrainian language*. Historical phonology of the Slavic languages ; 4. Winter, Heidelberg.

Do Large Language Models Adapt to Language Variation across Socioeconomic Status?

Elisa Bassignana^{1,2} Mike Zhang^{3,2} Dirk Hovy⁴ Amanda Cercas Curry⁵
¹IT University of Copenhagen ²Pioneer Center for AI ³University of Copenhagen
⁴Bocconi University ⁵CENTAI Institute
elba@itu.dk

Abstract

Humans adjust their linguistic style to the audience they are addressing. However, the extent to which LLMs adapt to different social contexts is largely unknown. As these models increasingly mediate human-to-human communication, their failure to adapt to diverse styles can perpetuate stereotypes and marginalize communities whose linguistic norms are less closely mirrored by the models, thereby reinforcing social stratification. We study the extent to which LLMs integrate into social media communication across different socioeconomic status (SES) communities. We collect a novel dataset from Reddit and YouTube, stratified by SES. We prompt four LLMs with incomplete text from that corpus and compare the LLM-generated completions to the originals along 94 sociolinguistic metrics, including syntactic, rhetorical, and lexical features. LLMs modulate their style with respect to SES to only a minor extent, often resulting in approximation or caricature, and tend to emulate the style of upper SES more effectively. Our findings (1) show how LLMs risk amplifying linguistic hierarchies and (2) call into question their validity for agent-based social simulation, survey experiments, and any research relying on language style as a social signal.

1 Introduction

Large-scale social media communication directly influences how people use language and how it evolves (Jebaselvi et al., 2023; Dembe, 2024). From this perspective, social media are not merely a communication tool, but a dynamic environment that actively shapes and transforms language in real-time (Akhmedova, 2024). Simultaneously, large language models (LLMs) are becoming an integral part of human communication by taking an active role in shaping how users write on social media (Yang and Menczer, 2024; Forbes, 2025), thereby directly influencing language use and communication.

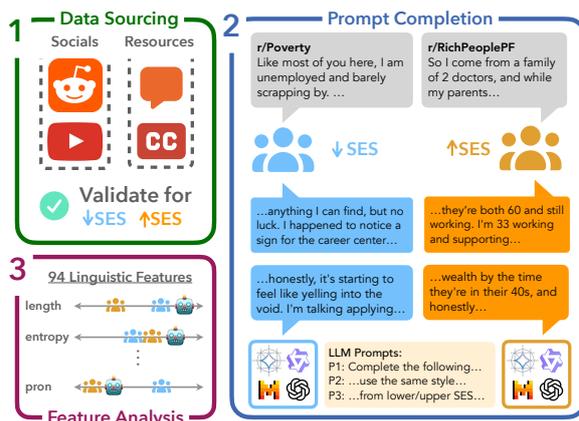


Figure 1: We compare the style of LLM-generated completions against the original text from lower and upper SES communities on Reddit and YouTube along 94 sociolinguistic dimensions.

Most work in NLP has focused on analyzing the content of LLM-generated text in terms of factuality (Min et al., 2023; Huang et al., 2025; Wang et al., 2024) and biases (Navigli et al., 2023; Fang et al., 2024; Gallegos et al., 2024; Stranisci et al., 2024). Wrong or biased LLM-generated content has a direct impact on misshaping users’ beliefs and behavior (Jakesch et al., 2023; Sharma et al., 2024). No work has investigated LLMs’ *language style* and its integration into social media communication across different communities. However, language style directly and crucially impacts communication (Meier et al., 2020; Havalдар et al., 2023; Jiang, 2024; Li and Zheng, 2024), especially large-scale social media communication (Danescu-Niculescu-Mizil et al., 2011; Muftah, 2022), and is more indicative of community than topic (Tran and Ostendorf, 2016).

We address this gap by specifically examining the language style of LLM-generated text in large-scale social media communication on Reddit and YouTube. Since style is closely tied to socioeconomic status (SES; Block, 2020), we explore the ex-

tent to which LLMs emulate the styles of different SES communities when prompted with their texts (see Figure 1). Sociolinguistics has shown substantial stylistic differences between individuals from the lower and upper ends of the socioeconomic spectrum (Flekova et al., 2016; Cercas Curry et al., 2024; Bassignana et al., 2025). In this work, we address the following research question: **To what extent do LLMs adapt their style to language variation across SES communities in social media communication?** We adopt a topic-based strategy with keyword matching to systematically collect the first open-access dataset of social media data (Reddit and YouTube), stratified by SES. We validate the stratification of language across SES against previous work (Flekova et al., 2016; Basile et al., 2019). We split each instance (i.e., Reddit posts, YouTube video captions) into two parts and use the first part to prompt four state-of-the-art LLMs with the instruction to complete the text. We then compare the LLM-generated text with the original second part. To formalize the style of LLM-generated text and compare it to in-community language, we employ 94 sociolinguistic metrics that capture syntactic, rhetorical, and lexical features.

Our results show that LLMs only partially match their output to different language styles, for example, in the level of formality marked by the ratio of pronouns and adverbs. For most other linguistic features, LLMs’ ability to replicate the specific linguistic profile of SES communities varies substantially, often resulting in approximation or caricature rather than precise emulation of the SES in-community language style. Our preliminary ablation study on the influence of the input context on model adaptation to SES language variation reveals a latent tendency of LLMs to adapt more effectively to upper SES style, when longer contexts are given. Our findings contribute to the growing research on how LLMs exacerbate socioeconomic inequalities (Bassignana et al., 2025). They also have implications for the validity of using LLMs to simulate humans in social science studies (Park et al., 2022).

Contributions. ① We provide the first openly accessible dataset of social media data (Reddit and YouTube) stratified by SES.¹ ② We quantitatively evaluate LLMs’ ability to replicate SES-specific style across two domains, four LLMs, three prompting strategies, and 94 sociolinguistic features.

¹We release our data and code at <https://github.com/elisabassignana/language-analysis-social-media>.

2 Related Work

Language Variation across SES. Since established work in sociolinguistics highlighted the differences in language use by individuals with different socioeconomic statuses (Labov, 2006), several works in NLP have attempted to quantify these differences on a large scale. In the context of social media, Preoȃuc-Pietro et al. (2015a,b) introduced a dataset of Twitter users mapped to their income through their occupational class using job-related keywords to match user profile fields. Flekova et al. (2016) employed that dataset to analyze the relation between language use and income on Twitter. However, more than for any other widely spread social media, the Twitter community is notoriously skewed towards the upper end of the socioeconomic spectrum (Sloan, 2017; Wojcik and Hughes, 2019). Basile et al. (2019) relied on the findings on Twitter by Flekova et al. (2016) to expand research on stylistic variation as a predictor of social classification in the domain of restaurant reviews.

Language Style of LLM-generated text. Muñoz-Ortiz et al. (2023) compared the language style of humans against LLM-generated news, finding measurable differences at the level of grammar, vocabulary, constituents, and dependency distances. Zamaraeva et al. (2025) expand the analysis of human-written versus LLM-generated NYT-style text through the lens of grammatical structure. Reinhart et al. (2025) investigated the ability of LLMs to match human language style in different domains (academic, news, fiction, podcasts, blogs, television, and movie scripts) and identified systematic differences across the style of distinct LLMs.

However, the extent to which LLMs integrate within large-scale communication on social media within different SES communities is largely unknown. In this work, we expand previous SES-stratified datasets to two new domains (i.e., Reddit and YouTube) and validate our data collection strategy against the findings of previous work.² Then, we use our dataset as an anchor to evaluate the language style of LLM-generated text against lower and upper SES communities along 94 features.

²We contacted Flekova et al. (2016) and Basile et al. (2019) to request their datasets, but were unsuccessful.

	YouTube		Reddit	
	Lower	Upper	Lower	Upper
# instances	472	622	1,988	1,689
# tokens	1.06M	1.13M	315.4K	227.8K

Table 1: **Statistics** of the lower and upper SES datasets. The # of instances refers to the number of video captions for YouTube and to the number of posts for Reddit.

3 Data

To explore the extent to which LLMs emulate SES in-community language on social media, we collect data from Reddit³ and from YouTube,⁴ each serving as a platform where distinct SES communities spontaneously congregate. Although SES communities naturally differ in the topics they discuss—reflecting distinct needs, lifestyles, social circles, and access to experiences—our analysis instead focuses on structural linguistic features, such as style, register, and other sociolinguistic markers, which are not affected by semantic variation. Table 1 reports the statistics of our dataset, and below we describe our data collection strategy.

3.1 Lower SES

Money matters are frequently discussed among lower SES communities (Lareau, 2018; McCaslin-Timmons and Grady, 2022; Ndou, 2024). Our data collection for the lower SES centers around topics related to *financial struggle, poverty, frugality, and benefits*. To identify relevant social media content, we manually compile a list of keywords related to these themes (e.g., ‘poverty’, ‘poor’, ‘frugal’) and use them to systematically search for subreddits (e.g., r/povertyfinance, r/Frugal) and YouTube videos.⁵ Then, to clean our data and to exclude onlookers from in-groups, we differentiate the process between Reddit and YouTube as follows:

Reddit. We perform a network analysis and clean our data by: (1) filtering only the posts by users that interacted (i.e., posting, commenting) at least 10 times within our set of subreddits and (2) interacted in at least three different subreddits within our list. We remove all usernames ending with ‘bot’, which is how bots are commonly referred to on Reddit (e.g., ‘u/sneakpeekbot’), and we manually check usernames including ‘mod’, which is often used

³We use the Pushshift API dumps and search for posts and comments from 2008 until 2024.

⁴YouTube data collected using the YouTube Data API.

⁵List of subreddits and YouTube searches in Appendix A.

to refer to automatic moderators. Finally, we omit posts including URLs and select a maximum of one post for each remaining user.

YouTube. As our goal is to collect videos spoken in the first person by individuals belonging to our target group, we query the API using the combination of ‘vlog’ and a keyword. To further clean the retrieved list of videos from non relevant content (e.g., documentaries) we filter in only the videos where the title is written in first person by checking for first-person pronouns (e.g., ‘I’, ‘we’, ‘my’, ‘us’). Then we use the YouTubeTranscriptApi to scrape the caption of our final set of videos.

To support a more robust analysis of the language style, we retain only instances (Reddit posts and YouTube captions) with at least 50 words.

3.2 Upper SES

On the other hand, online communities gathering around subreddits like r/Rich or r/millionaire mainly attract individuals seeking advice on how to achieve a higher level of wealth. In fact, upper SES individuals rarely explicitly discuss their own wealth (Hing et al., 2019). Instead, hobbies have been shown to be a highly distinctive dimension for upper SES and a symbol of social identity (Bourdieu, 1984). We center our data collection for the upper SES around lifestyle, hobbies, and leisure activities. Sawert and Bachsleitner (2021) analyzes how upper-class families use specific sports and classical music to transmit privilege across generations. Based on previous literature (Engstrom, 1974; Hwang et al., 2012; Post et al., 2018; Eime et al., 2015; Friedman and Reeves, 2020; Arnold-Forster, 2022; Schmitt et al., 2020; Sawert and Bachsleitner, 2021; Cuijuan and Hai, 2023), we collect a list of hobbies and leisure activities (e.g., golf, sailing) that are distinctive of the upper SES and we employ it for systematic data collection on Reddit and YouTube. Then, we use the same cleaning process as for the lower SES (see Section 3.1).

3.3 Dataset Validation

While collecting data from social media is mostly straightforward, inferring users’ SES is a more complex matter. As detailed in Sections 3.1 and 3.2, we identified SES-based communities from previous literature about SES topics of interest. Here, we externally validate our dataset to further ensure the soundness of our strategy. Flekova et al. (2016) showed how the readability of text correlates with

Metric	Lower SES	Upper SES
ARI*	7.18	7.36
Coleman-Liau*	6.16	6.43
Dale-Chall*	8.45	8.76
Flesch-Kincaid*	7.04	7.18
Flesch-Reading*	73.99	72.45
Gunning-Fog*	9.12	9.29
Linsear	8.81	8.88

Table 2: **Mean Readability Scores per SES Group (Reddit)**. (*) indicates a statistically significant difference in the distributions of readability scores ($p < 0.05$) between lower and upper SES, as determined by a Mann-Whitney U test (Mann and Whitney, 1947).

income. Their findings are consistent with observations that readability correlates with education (Davenport and DeLine, 2014), which has an important role in determining SES (Bourdieu, 2018). Following Basile et al. (2019) we validate our dataset against Flekova et al. (2016)’ results on readability metrics: Automated Readability Index (ARI; Smith and Senter, 1967), Coleman Liau Index (Coleman and Liau, 1975), Dale Chall Readability Score (Dale and Chall, 1948), Flesch-Kincaid Grade (Kincaid et al., 1975), Flesch-Reading Ease (Flesch, 1948), Gunning Fog (Gunning, 1952) and Linsear Write Formula (U.S. Air Force, 1975).⁶ Readability metrics are designed to estimate a text’s complexity, typically by analyzing the average number of syllables per word and words per sentence. Following Flekova et al. (2016) and Basile et al. (2019), we expect the readability scores to increase from the lower SES to upper SES subsets, except for Flesch-Reading Ease, which, by definition of the metric, leads us to expect an inverse correlation. Table 2 reports the readability metrics on our dataset, where the scores follow the expected trends.⁷ The differences are statistically significant according to a Mann-Whitney U test (Mann and Whitney, 1947). Similar to Basile et al. (2019), only the Linsear Write Formula does not show a significant difference across SES groups.

4 LLM-Generated Data

To analyze whether and to what extent LLMs emulate in-community language style based on the input prompt, we generate social media data start-

⁶We use the implementations of textstats: <https://github.com/textstat/textstat>.

⁷Readability metrics are designed for written language. Therefore, the metrics in Table 2 refer to the Reddit dataset. For completeness, we also report the readability scores of the YouTube portion of the dataset in Appendix B.

ing from the data we collected. Following Reinhart et al. (2025) we split each instance in our dataset (Reddit posts and YouTube captions) in two parts. We use the first part (i.e., 25 words) as a language cue for the input prompt and instruct the models to complete the text. We use the second part (i.e., the remaining text) as a comparison with the LLM-generated content. We analyze three variations of the prompt, which increasingly explicit the models to adapt their output with respect to the input:

1. **Implicit (IMP):** “Complete the following [Reddit post / caption for a YouTube video]. Only generate the completion and nothing else. \n{text}”;
2. **Explicit Language Style (ELS):** “Complete the following [Reddit post / caption for a YouTube video] using the same style, tone, and diction of the first part. Only generate the completion and nothing else. \n{text}”;
3. **Explicit Language Style + SES (ELS-SES):** “Complete the following [Reddit post / caption for a YouTube video] written by a user from a [lower / upper] socioeconomic status using the same style, tone, and diction of the first part. Only generate the completion and nothing else. \n{text}”.

We experiment with Gemma-3-27B-it (Team et al., 2025), Mistral-Small-3.2-24B-Instruct-2506 (Mistral AI, 2025), Qwen3-30B-A3B-Instruct-2507 (Qwen Team, 2025), and GPT-5 (OpenAI, 2025). For inference, we generate one chat completion per prompt with the default temperature. Finally, we depict the hardware, inference costs and environmental impact in Appendix E.

5 Linguistic Analysis

We rely on 94 sociolinguistic metrics to analyze the style of lower and upper SES communities, as well as the style of the LLM-generated text.

Biber features. We use Douglas Biber’s set of 67 linguistic categories collected from Biber (1991, 1995), and normalize them by the instance length. The set includes lexical, grammatical, and rhetorical features (e.g., pronouns, tense, place and time adverbials, discourse particles, adjectives before a noun). We report the full list in Appendix C.⁸

⁸We rely on the implementation of the pseudobibeR package: <https://github.com/browndw/pseudobibeR>.

Part of Speech. Texts with more nouns and articles as opposed to pronouns and adverbs are considered more formal (Pennebaker et al., 2003; Argamon et al., 2009; Rangel et al., 2014). We use Spacy’s POS-tagger and normalize each count by the length of the instance.

Length. As shorter words are considered more readable (Gunning, 1969; Pitler and Nenkova, 2008), and following Flekova et al. (2016), we calculate the number of words, syllables, lexicon, sentences, characters, letters, polysyllables, and monosyllables.

Style. We assess the level of concreteness using the list proposed by Brysbaert et al. (2014). In this collection, each word is labeled on a scale from one (abstract) to five (concrete). We compute the level of concreteness as the mean value of the words in a text. Additionally, we compute the entropy and utilize Spacy to determine the maximum depth of the syntax dependency trees as a measure of syntactic complexity. Last, following Cercas Curry et al. (2024) we compute the ratio of named entities (NE), and the ratio of hapax legomena.

6 Results

6.1 Reddit

In Figure 2, we show our results on Reddit data across three prompts of increasingly explicit language style and four state-of-the-art open and closed-source models (Section 4). We report features that are statistically significantly different according to a Mann-Whitney U test (Mann and Whitney, 1947) between lower SES (\downarrow SES) and upper SES (\uparrow SES) online communities, with a Holm-Bonferroni correction (Holm, 1979) applied to control for multiple comparisons at an $\alpha = 0.01$. We report the complete results in Appendix D. To facilitate visualization, in the forest plot in Figure 2, we use lower SES humans as an anchor and compare the frequency ratio of the features against it.

Biber features. Individuals from upper SES communities tend to have a wider and more sophisticated vocabulary (Du et al., 2022). These differences emerge in our results through higher values of ‘mean_word_length’ and token type ratio for upper SES over lower SES communities. In respect, LLMs generate sequences with similar word length for both communities, except when prompt 3 (ELS-SES) makes the SES group explicit: In this

case, all four LLMs exacerbate the differences by generating even shorter (longer) words than the corresponding lower (upper) SES communities. The token type ratio of generated text is always lower than that of human-written text, indicating a more repetitive use of language. The use of ‘that’ as a subject indicates simpler and more colloquial syntactic structures, which are more frequent among lower SES. LLMs generally generate this structure with a higher frequency with respect to the corresponding human-written text (except for Gemma). Trends across the lower and upper SES generated texts are mixed across prompt types for all four models, indicating that LLMs do not pick up on the use of ‘that’ as a subject.

Part of Speech. The higher rate of determiners (‘det’) in the language used by the upper SES is largely reflected in the output of all four LLMs, with prompt 3 exacerbating this gap. The different rates of nouns in lower and upper SES language are not substantial in the corresponding lower and upper SES-generated text, except for the output of Mistral and Qwen for prompt 3. In this case, however, both models generate an opposed trend with respect to the lower/upper human-written text. Finally, in-community language contains a higher rate of verbs, pronouns (including ‘pron’ and ‘first_person_pronouns’), and adverbs (‘time_adverbials’, ‘downtoners’) for lower SES (Jones and McMillan, 1973; Shi and Lei, 2021), indicating a more informal style (Pennebaker et al., 2003; Argamon et al., 2009; Rangel et al., 2014). The ratio of these PoS in the LLM-generated text follows the respective lower and upper SES communities trend, with exacerbated differences when including the SES in the prompt (prompt 3).

Length. LLMs have a general bias towards verbosity (Saito et al., 2023). Consistently, in our experiments, LLM-generated text is generally longer than human-written text. Especially Gemma and GPT generate text that is approximately twice as long as the corresponding human version, for all three prompt types. Mistral tends to generate shorter sequences of text and, similarly to Qwen, to differentiate more between lower and upper SES. However, both models generate longer sequences for upper SES and shorter sequences for lower SES (especially when using prompt 3), which contrasts with the trend observed in lower/upper communities. In fact, individuals with upper SES tend to express themselves with fewer words, likely due to

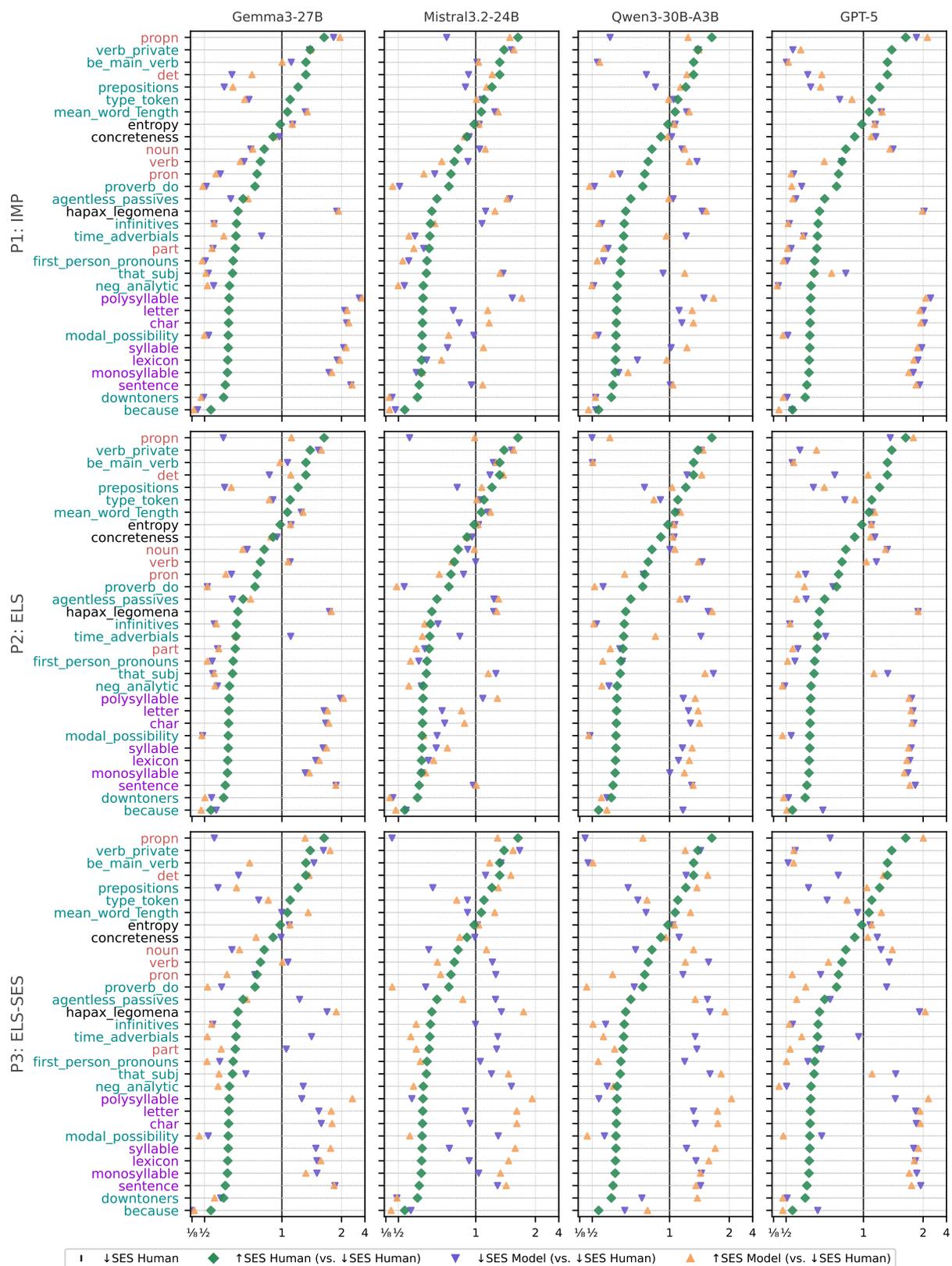


Figure 2: **Forest Plots Comparing Linguistic Features of Humans and Models on Reddit.** We *only* show the linguistic features (31) with a statistically significant difference with correction ($p < 0.01$; Mann and Whitney, 1947; Holm, 1979) in usage between lower SES (\downarrow SES; rate = 1) and upper SES (\uparrow SES) human writers. Each point indicates the frequency ratio of a feature in a model’s (or human’s) output compared to human text from the \downarrow SES group. These comparisons are presented across four models and three prompts (see Section 4). Feature types are color-coded: Biber features are cyan, length-specific features are dark violet, PoS tags are red, and style features are black.

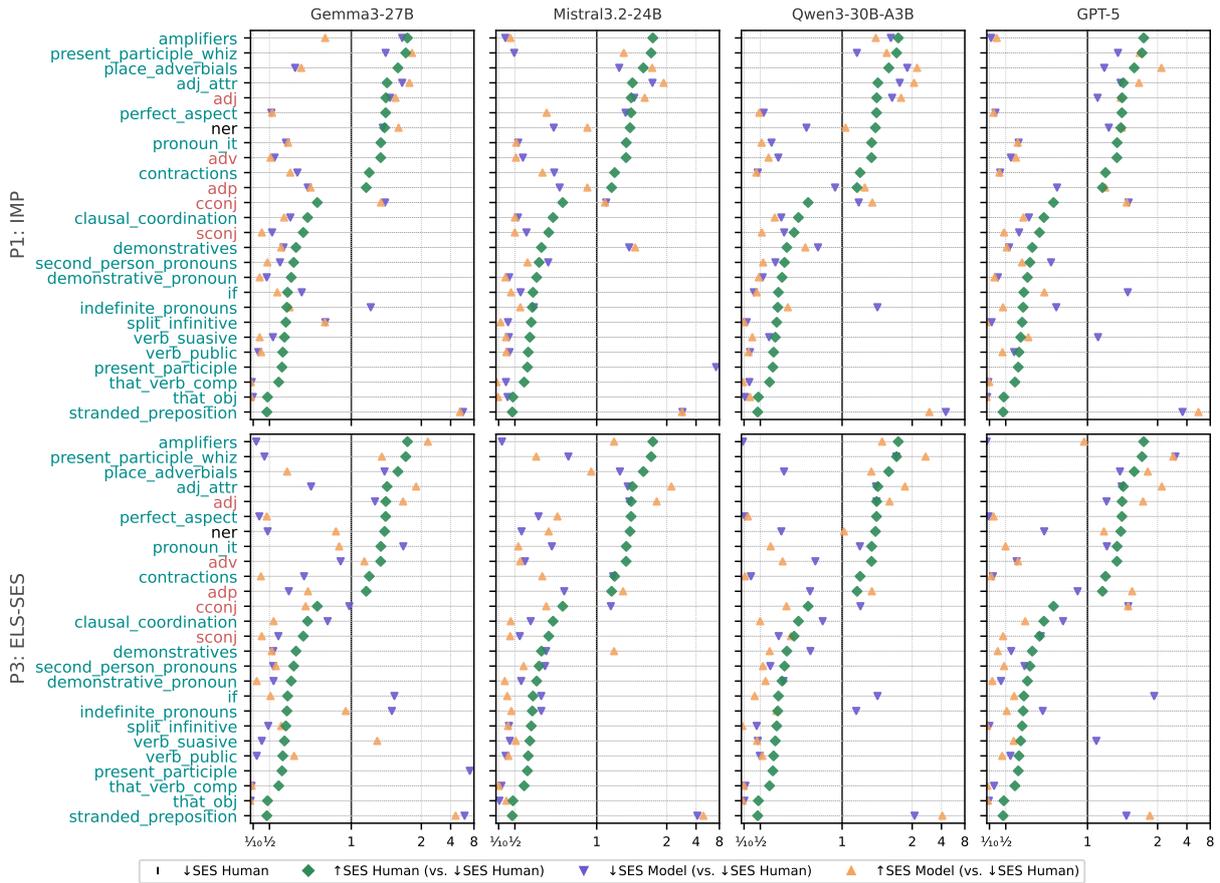


Figure 3: **Comparison of Linguistic Features on YouTube.** This plot displays *only* the linguistic features, not present in Reddit results, that show a statistically significant difference ($p < 0.01$, corrected) between lower SES (\downarrow SES) and upper SES (\uparrow SES) human writers. Each point represents the frequency ratio of a feature relative to the \downarrow SES human group. Feature types are color-coded: Biber (cyan), length (dark violet), PoS (red), and style (black).

a wider vocabulary (Bassignana et al., 2025).

Style. Bernstein (1960) posits that individuals from upper SES families are more encouraged to use language for abstract thinking in contrast to people from lower SES families, who are exposed to more concrete concepts. This difference is reflected in the language style adopted by lower and upper-SES communities on social media. LLMs pick up on the different levels of concreteness within various communities to a limited extent, which is further amplified in the results obtained with prompt 3. The hapax legomena is higher for lower SES communities (as well as to a smaller extent, the entropy) that we speculate being a consequence of lower SES language containing more non-standard expressions and slang with respect to the upper SES counterpart (Lansley and Longley, 2016). LLM-generated text does not capture these differences across SES communities (prompts 1 and 2). When making SES explicit (prompt 3), Mistral and Qwen differentiate the text; however,

the results trend in the opposite direction compared to human values (i.e., lower values of hapax legomena for the completion of the lower SES prompts).

Overall, we report little differences between the results obtained with the first and the second prompt strategies (IMP, ELS), indicating that LLMs do not easily pick up the language cues of the input text, even when explicitly prompted to replicate the style. When the input prompt contains explicit information about the social community to emulate (lower or upper SES), LLMs differentiate the generated output more distinctly. However, language differences are often exaggerated with respect to human values or do not accurately reflect the real trends across SES communities.

6.2 YouTube

In Figure 3, we show our results on YouTube data for prompts 1 and 3 (IMP and ELS-SES), focusing on the features that are distinctive of spoken language (i.e., that are not statistically significantly

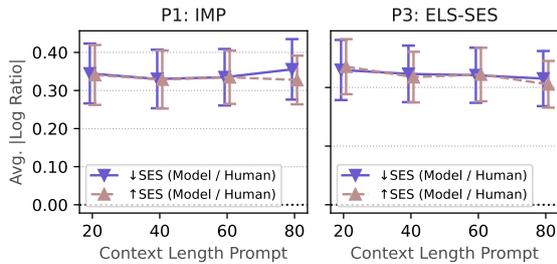


Figure 4: Average absolute logarithm of the ratio between model and human text across increasing context.

different in Reddit as well). We report the complete results in Appendix D. As for Reddit, LLMs generally struggle to match the right register in terms of formality. The results for informality markers, such as contractions and stranded prepositions, show inverse trends with respect to both lower- and upper-SES communities. The language of upper SES communities is characterized by higher syntax complexity: LLMs correctly identify the association with present participial postnominal (`present_participle_whiz`) and increase their usage when generating upper SES text (except Mistral and GPT for prompt 3), but under-produce perfect aspect. All four models tend to exhibit the correct lower/upper SES trends in terms of indefinite pronouns, with less frequent usage in upper SES communities, although with varying proportions compared to human language. Instead, they undergenerate demonstrative pronouns and fail to differentiate across SES community trends.

Similar to the Reddit results, all models demonstrate some ability to modulate their style, particularly with the explicit ELS-SES prompt. However, their ability to accurately replicate the specific linguistic profile of a language community varies substantially, often resulting in approximation or caricature rather than precise emulation of the in-community language style.

7 Ablation Study on the Input Context

We investigate the impact of context length on LLMs’ tendency to pick up on linguistic cues from the input prompt and modulate the language style of the output accordingly. We use the same methodology presented in Section 4, but we experiment with increasingly long input contexts. We filter the instances with at least 100 words and split them using, respectively, the first 20, 40, 60, 80 words for prompting and the rest for the comparison with

the LLM-generated text. For this investigation, due to computational constraints, we focus on Reddit, leave out prompt 2 (ELS), and experiment with Gemma only. Figure 4 presents the aggregated results, showing the average absolute logarithmic ratio between human and LLM text for lower and upper SES, respectively, across all features.⁹ We observe: (1) Only minor variation across increasing context length (and mainly for prompt 3), indicating that LLMs rely on linguistic cues from the input prompt to modulate their output for different language styles only to a minor extent. (2) When increasing the context to 80 words, the ratio between human and LLM is smaller for upper SES, indicating a latent tendency of LLMs (Gemma) to integrate better within upper SES language style.

8 Discussion and Conclusion

Language style is closely tied to SES (Block, 2020). We investigate the extent to which LLMs adapt their style to language variation across lower and upper SES communities. We compare the stylistic variation of online SES communities against LLM-generated text. While the language adopted by lower SES is less formal, less predictable due to the jargon adopted, and more concrete, LLMs modulate their style to SES in-community language only to a limited extent, often resulting in approximation or caricature rather than accurate emulation. A longer context does not facilitate LLMs in capturing the style of different language communities. Notably, our ablation study indicates a latent tendency for models to better emulate the style of upper SES communities compared to lower SES.

Our findings reveal severe consequences for the widespread adoption of LLMs for communication: if LLMs modulate their style more easily towards certain communities than others, they directly contribute to exacerbating social inequalities (Capraro et al., 2024). Communities that are not accurately represented may experience a degraded user experience, hindering the adoption of these technologies (Davis, 1989), and ultimately contributing to the exacerbation of the so-called ‘AI-gap’ (Bassignana et al., 2025). Additionally, our results pose a new challenge to the validity of using LLMs to emulate humans for agent-based social simulation and research relying on language style as a social signal (Argyle et al., 2022; Aher et al., 2022).

⁹We leave out of this analysis the features related to the length to focus on the style.

Limitations

While our study reveals important disparities in how LLMs model sociolinguistic variation, it is limited in terms of:

- **Dataset:** Our analysis focuses on two social media platforms. Language patterns differ based on the affordances and specific communities of each platform.
- **Operationalisation of SES:** As previous work on NLP and SES, our analysis uses proxies rather than self-reported or objective measures of SES. In addition, we do not consider a middle-SES.
- **Choice and scope of linguistic measures:** We used a set of 94 surface metrics supported from previous studies in sociolinguistics; however, it is not exhaustive, and it is possible they might miss something more nuanced.
- **Prompt design:** LLM outputs are strongly dependent on the prompt used to generate them. While we test three different prompt variations, it is possible that a different prompt or prompting strategy would result in stronger variation.
- **Models:** We tested a limited number of models, and new models are published regularly. It is possible that our findings may not hold across future model versions.

Ethical Considerations

For the analysis proposed in the paper, we utilize Reddit data from the Pushshift API dumps and YouTube data scraped using the YouTube Data API, adhering to the developer guidelines of the respective platforms. We do not collect any user-identifiable information (i.e., Reddit and YouTube usernames).

Acknowledgments

We thank Francesco Corso for his help in collecting and cleaning the Reddit data. We thank the MilaNLP group at Bocconi University for feedback on an earlier version of this draft. Elisa Bassignana is supported by a research grant (VIL59826) from VILLUM FONDEN. Mike Zhang is supported by funding from the Danish Government to Danish

Foundation Models (4378-00001B). Dirk Hovy is supported by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (grant agreement No. 949944, INTEGRATOR). He is a member the Data and Marketing Insights Unit of the Bocconi Institute for Data Science and Analysis (BIDSA).

References

- Gati Aher, RosaI. Arriaga, and A. Kalai. 2022. [Using large language models to simulate multiple humans and replicate human subject studies](#). In *International Conference on Machine Learning*.
- S Akhmedova. 2024. The influence of social media on language evolution. *Science and innovation*, 3(C10):51–58.
- Shlomo Argamon, Moshe Koppel, James W. Pennebaker, and Jonathan Schler. 2009. [Automatically profiling the author of an anonymous text](#). *Commun. ACM*, 52(2):119–123.
- Lisa P. Argyle, E. Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and D. Wingate. 2022. [Out of one, many: Using language models to simulate human samples](#). *Political Analysis*, 31:337 – 351.
- Agnes Arnold-Forster. 2022. The social and emotional world of twentieth-century anglo-american surgery: The james iv association of surgeons. *Bulletin of the History of Medicine*, 96(1):71–101.
- Angelo Basile, Albert Gatt, and Malvina Nissim. 2019. [You write like you eat: Stylistic variation as a predictor of social stratification](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2583–2593, Florence, Italy. Association for Computational Linguistics.
- Elisa Bassignana, Amanda Cercas Curry, and Dirk Hovy. 2025. [The AI gap: How socioeconomic status affects language technology interactions](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18647–18664, Vienna, Austria. Association for Computational Linguistics.
- Basil Bernstein. 1960. [Language and social class](#). *The British Journal of Sociology*, 11(3):271–276.
- Douglas Biber. 1991. *Variation across speech and writing*. Cambridge university press.
- Douglas Biber. 1995. Dimensions of register variation: A cross-linguistic comparison. *Cambridge University Press google schola*, 2:171–197.
- David Block. 2020. [Language and social class](#). *The International Encyclopedia of Linguistic Anthropology*.

- Pierre Bourdieu. 1984. *Distinction. A social critique of the judgement of taste*. Harvard, MA: Harvard University Press.
- Pierre Bourdieu. 2018. Distinction: A Social Critique of the Judgement of Taste. In *Inequality*, pages 287–318. Routledge.
- Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. [Concreteness ratings for 40 thousand generally known english word lemmas](#). *Behavior Research Methods*, 46:904–911.
- Valerio Capraro, Austin Lentsch, Daron Acemoglu, Selin Akgun, Aisel Akhmedova, Ennio Bilancini, Jean-François Bonnefon, Pablo Brañas-Garza, Luigi Butera, Karen M Douglas, et al. 2024. The impact of generative artificial intelligence on socioeconomic inequalities and policy making. *PNAS nexus*, 3(6).
- Amanda Cercas Curry, Giuseppe Attanasio, Zeerak Talat, and Dirk Hovy. 2024. [Classist tools: Social class correlates with performance in NLP](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12643–12655, Bangkok, Thailand. Association for Computational Linguistics.
- Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.
- Lin Cuijuan and Yang Hai. 2023. [Research on the influence of cultural capital on human sports practice](#). In *SHS Web of Conferences*, volume 163, page 02016. EDP Sciences.
- Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.
- Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. 2011. [Mark my words! linguistic style accommodation in social media](#). In *Proceedings of the 20th International Conference on World Wide Web, WWW '11*, page 745–754, New York, NY, USA. Association for Computing Machinery.
- James RA Davenport and Robert DeLine. 2014. The readability of tweets and their geographic correlation with education. *arXiv preprint arXiv:1401.6058*.
- Fred D Davis. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, pages 319–340.
- Tony Dembe. 2024. The impact of social media on language evolution. *European Journal of Linguistics*, 3(3):1–14.
- Xiangtao Du, Muhammad Afzaal, and Hind Al Fadda. 2022. Collocation use in efl learners’ writing across multiple language proficiencies: A corpus-driven study. *Frontiers in Psychology*, 13:752134.
- Rochelle M Eime, Melanie J Charity, Jack T Harvey, and Warren R Payne. 2015. [Participation in sport and physical activity: associations with socio-economic status and geographical remoteness](#). *BMC public health*, 15.
- Lars-Magnus Engstrom. 1974. [Physical activities during leisure time: A survey of teenage sport activities in sweden](#). *International Review of Sport Sociology*, 9(2):83–102.
- Xiao Fang, Shangkun Che, Minjia Mao, Hongzhe Zhang, Ming Zhao, and Xiaohang Zhao. 2024. [Bias of ai-generated content: an examination of news produced by large language models](#). *Scientific Reports*, 14(1):5224.
- Lucie Flekova, Daniel Preotiuc-Pietro, and Lyle Ungar. 2016. [Exploring stylistic variation with age and income on Twitter](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 313–319, Berlin, Germany. Association for Computational Linguistics.
- Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.
- Forbes. 2025. [How AI Is Changing The Way We Communicate: The Future Of Interaction](#).
- Sam Friedman and Aaron Reeves. 2020. [From aristocratic to ordinary: Shifting modes of elite distinction](#). *American Sociological Review*, 85(2):323–350.
- Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. [Bias and fairness in large language models: A survey](#). *Computational Linguistics*, 50(3):1097–1179.
- Robert Gunning. 1952. The technique of clear writing. (*No Title*).
- Robert Gunning. 1969. The fog index after twenty years. *Journal of Business Communication*, 6(2):3–13.
- Shreya Havaldar, Matthew Pressimone, Eric Wong, and Lyle Ungar. 2023. [Comparing styles across languages](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6775–6791, Singapore. Association for Computational Linguistics.
- Leanne S. Son Hing, Anne E. Wilson, Peter Gourevitch, Jaslyn English, and Parco Sin. 2019. [Failure to respond to rising income inequality: Processes that legitimize growing disparities](#). *Daedalus*, 148(3):105–135.
- Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70.

- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. [A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions](#). *ACM Trans. Inf. Syst.*, 43(2).
- Junwook Hwang, Minki Hong, Seung-Yeol Yee, and Sang-Min Lee. 2012. [Impact of sports' characteristics on the labor market](#). *International Review for the Sociology of Sport*, 47(1):60–76.
- Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. [Co-writing with opinionated language models affects users' views](#). In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.
- C Alice Evangaline Jebaselvi, K Mohanraj, A Thangamani, and M Ramesh Kumar. 2023. [The impact of social media on the evolution of language and communication trends](#). *Shanlax International Journal of English*.
- Chunlan Jiang. 2024. [Impact of changes in anchor language style on audience language habits and social communication methods: Investigation and analysis](#). In *SHS Web of Conferences*, volume 183, page 02018. EDP Sciences.
- Pauline A Jones and William B McMillan. 1973. [Speech characteristics as a function of social class and situational factors](#). *Child Development*, pages 117–121.
- J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. [Derivation of new readability formulas \(automated readability index, fog count and flesch reading ease formula\) for navy enlisted personnel](#).
- 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 ACM SIGOPS 29th Symposium on Operating Systems Principles*.
- William Labov. 2006. *The Social Stratification of English in New York City*, 2 edition. Cambridge University Press.
- Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. [Quantifying the carbon emissions of machine learning](#). *ArXiv preprint*, abs/1910.09700.
- Guy Lansley and Paul A. Longley. 2016. [The geography of twitter topics in london](#). *Computers, Environment and Urban Systems*, 58:85–96.
- Annette Lareau. 2018. [Unequal childhoods: Class, race, and family life](#). In *Inequality in the 21st Century*, pages 444–451. Routledge.
- Zhen Li and Suqin Zheng. 2024. [The influence of the language style of the anchor on consumers' purchase intention](#). *Frontiers in Psychology*, 15:1370712.
- Henry B Mann and Donald R Whitney. 1947. [On a test of whether one of two random variables is stochastically larger than the other](#). *The annals of mathematical statistics*, pages 50–60.
- Jill McCaslin-Timmons and Marilyn Grady. 2022. [Life-long effects of poverty](#).
- Tabea Meier, Ryan L. Boyd, Matthias R. Mehl, Anne Milek, James W. Pennebaker, Mike Martin, Markus Wolf, and Andrea B. Horn. 2020. [Stereotyping in the digital age: Male language is “ingenious”, female language is “beautiful” – and popular](#). *PLOS ONE*, 15(12):1–29.
- Sewon Min, Kalpesh Krishna, Xinxu Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. [FActScore: Fine-grained atomic evaluation of factual precision in long form text generation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.
- Mistral AI. 2025. [Mistral small 3.1](#). <https://mistral.ai/news/mistral-small-3-1>. Accessed: 2025-09-24.
- Muneera Muftah. 2022. [Impact of social media on learning english language during the covid-19 pandemic](#). *PSU Research Review*, 8(1):211–226.
- Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2023. [Contrasting linguistic patterns in human and llm-generated news text](#). *Artificial Intelligence Review*, 57.
- Roberto Navigli, Simone Conia, and Björn Ross. 2023. [Biases in large language models: Origins, inventory, and discussion](#). *J. Data and Information Quality*, 15(2).
- Adam Ndou. 2024. [The difference in parental financial socialisation across parental education level](#). *Research in Social Sciences and Technology*, 9(2):13–30.
- OpenAI. 2025. [Introducing gpt-5](#). <https://openai.com/index/introducing-gpt-5/>. Accessed on 2025-09-29.
- Katrin Ortman and Stefanie Dipper. 2019. [Variation between different discourse types: Literate vs. oral](#). In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 64–79, Ann Arbor, Michigan. Association for Computational Linguistics.
- Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022. [Social simulacra: Creating populated](#)

- prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, UIST '22, New York, NY, USA. Association for Computing Machinery.
- James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577.
- Emily Pitler and Ani Nenkova. 2008. [Revisiting readability: A unified framework for predicting text quality](#). In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 186–195, Honolulu, Hawaii. Association for Computational Linguistics.
- Eric G Post, Nicole E Green, Daniel A Schaefer, Stephanie M Trigsted, M Alison Brooks, Timothy A McGuine, Andrew M Watson, and David R Bell. 2018. [Socioeconomic status of parents with children participating on youth club sport teams](#). *Physical Therapy in Sport: official journal of the Association of Chartered Physiotherapists in Sports Medicine*, 32:126–132.
- Daniel Preoŕiuc-Pietro, Vasileios Lampos, and Nikolaos Aletras. 2015a. [An analysis of the user occupational class through Twitter content](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1754–1764, Beijing, China. Association for Computational Linguistics.
- Daniel Preoŕiuc-Pietro, Svitlana Volkova, Vasileios Lampos, Yoram Bachrach, and Nikolaos Aletras. 2015b. Studying user income through language, behaviour and affect in social media. *PloS one*, 10(9):e0138717.
- Qwen Team. 2025. [Qwen3 technical report](#). Preprint, arXiv:2505.09388.
- Francisco Rangel, Paolo Rosso, Irina Chugur, Martin Potthast, Martin Trenkmann, Benno Stein, Ben Verhoeven, and Walter Daelemans. 2014. Overview of the 2nd author profiling task at pan 2014. In *CLEF 2014 Evaluation Labs and Workshop Working Notes Papers*, Sheffield, UK, 2014, pages 1–30.
- Alex Reinhart, Ben Markey, Michael Laudenbach, Kachata Pantusen, Ronald Yurko, Gordon Weinberg, and David West Brown. 2025. [Do llms write like humans? variation in grammatical and rhetorical styles](#). *Proceedings of the National Academy of Sciences*, 122(8):e2422455122.
- Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. 2023. [Verbosity bias in preference labeling by large language models](#). *ArXiv*, abs/2310.10076.
- Tim Sawert and Anna Bachsleitner. 2021. [Within-track differentiation and leisure activities as strategies of distinction](#). *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 73:555 – 578.
- Anne Schmitt, Matthew Atencio, and Gaëlle Sempé. 2020. [“you’re sitting on a hot soccer field drinking gatorade. . . i’m sitting in a yacht club just enjoying the view, enjoying the drinks”](#): Parental reproduction of social class through school sport sailing. *European Physical Education Review*, 26(4):987–1005.
- Nikhil Sharma, Q. Vera Liao, and Ziang Xiao. 2024. [Generative echo chamber? effect of llm-powered search systems on diverse information seeking](#). In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.
- Yaqian Shi and Lei Lei. 2021. Lexical use and social class: A study on lexical richness, word length, and word class in spoken english. *Lingua*, 262:103155.
- Luke Sloan. 2017. [Who tweets in the united kingdom? profiling the twitter population using the british social attitudes survey 2015](#). *Social Media + Society*, 3(1):2056305117698981.
- Edgar A Smith and RJ Senter. 1967. *Automated readability index*, volume 66. Aerospace Medical Research Laboratories, Aerospace Medical Division, Air Force Systems Command.
- Marco Stranisci, Pere-Lluís Huguet Cabot, Elisa Bassignana, and Roberto Navigli. 2024. [Dissecting biases in relation extraction: A cross-dataset analysis on people’s gender and origin](#). In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 190–202, Bangkok, Thailand. Association for Computational Linguistics.
- Deborah Tannen. 1982. [Oral and literate strategies in spoken and written narratives](#). *Language*, 58(1):1–21.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Keanealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesh Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Pateron, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Pettrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch,

Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Pappas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivan, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Pöder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. [Gemma 3 technical report](#). *Preprint*, arXiv:2503.19786.

Trang Tran and Mari Ostendorf. 2016. [Characterizing the language of online communities and its relation to community reception](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1030–1035, Austin, Texas. Association for Computational Linguistics.

U.S. Air Force. 1975. Manual for readability formulas. Technical report, Aerospace Medical Research Laboratory, Wright-Patterson Air Force Base, OH.

Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi Nenkov Georgiev, Rocktim Jyoti Das, and Preslav Nakov. 2024. [Factuality of large language models: A survey](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19519–19529, Miami, Florida, USA. Association for Computational Linguistics.

Stefan Wojcik and Adam Hughes. 2019. Sizing up twitter users.

Kai-Cheng Yang and Filippo Menczer. 2024. [Anatomy of an ai-powered malicious social botnet](#). *Journal of Quantitative Description: Digital Media*, 4.

Olga Zamaraeva, Dan Flickinger, Francis Bond, and Carlos Gómez-Rodríguez. 2025. [Comparing LLM-generated and human-authored news text using formal syntactic theory](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9041–9060, Vienna, Austria. Association for Computational Linguistics.

A Data Collection

Lower SES.

- Subreddits list: povertyfinance, povertykitchen, frugalmalefashion, FrugalFemaleFashion, FrugalLiving, TrueFrugal, FrugalShopping, poor, Frugal, Cheap_Meals.
- YouTube searches: low income, living on minimum wage, surviving on food stamps, food stamps experience, living paycheck to paycheck, budgeting on minimum wage, food bank haul, food pantry haul, public housing tour, homeless, unemployed, section 8 apartment tour, feeding a family on a budget, living in poverty, eviction.

Upper SES.

- Subreddits list: RichPeoplePF, fatFIRE, Philanthropy, golf, Rowing, horseracing, Equestrian, Horses, tennis, 10s, Shooting, CompetitionShooting, Hunting, polo, sailing, opera, classicalmusic, FineArt, FineArtPhoto, literature, Fencing, Fieldhockey, yachting.
- YouTube searches: golf, rowing, equestrian, tennis, horse riding, sailing, fine arts, opera, classical music, literature, fencing, field hockey, yachting, yacht.

B Readability Metrics YouTube

Readability metrics are designed for written language, which is in principle different from spoken language (Tannen, 1982; Ortman and Dipper, 2019). However, for completeness, here we report the readability metrics computed on the YouTube portion of our dataset. As these metrics typically normalize by sentence length, and YouTube transcripts are not split into sentences, we first run the

Metric	Lower SES	Upper SES
ARI	4.03	3.78
Coleman-Liau	4.19	4.17
Dale-Chall	7.19	7.28
Flesch-Kincaid	4.61	4.21
Flesch-Reading*	85.01	85.71
Gunning-Fog	6.81	6.30
Linsear	6.73	5.87

Table 3: **Mean Readability Scores per SES Group (YouTube)**. (*) indicates a statistically significant difference in the distributions of readability scores ($p < 0.05$) between lower and upper SES, as determined by a Mann-Whitney U test (Mann and Whitney, 1947).

wtpsplit sentence tokenizer.¹⁰ Differently from Reddit, the metrics computed on YouTube captions do not differ significantly between SES communities (except for Flesch-Reading).

C Biber’s Features

Below we list the full set of Biber’s features adapted from Biber (1991, 1995) that we compute using the pseudobibeR library.¹¹

Tense and aspect markers

- past_tense: Past tense
- perfect_aspect: Perfect aspect
- present_tense: Present tense

Place and time adverbials

- place_adverbials: Place adverbials (e.g., above, beside, outdoors)
- time_adverbials: Time adverbials (e.g., early, instantly, soon)

Pronouns and pro-verbs

- first_person_pronouns: First-person pronouns
- second_person_pronouns: Second-person pronouns
- third_person_pronouns: Third-person personal pronouns (excluding *it*)
- pronoun_it: Pronoun *it*

¹⁰<https://github.com/segment-any-text/wtpsplit/>

¹¹<https://github.com/browndw/pseudobibeR>.

- demonstrative_pronoun: Demonstrative pronouns (*that, this, these, those* as pronouns)
- indefinite_pronoun: Indefinite pronouns (e.g., *anybody, nothing, someone*)
- proverb_do: Pro-verb *do*

Questions

- wh_question: Direct wh-questions

Nominal forms

- nominalization: Nominalizations (ending in *-tion, -ment, -ness, -ity*)
- gerunds: Gerunds (participial forms functioning as nouns)
- other_nouns: Total other nouns

Passives

- agentless_passives: Agentless passives
- by_passives: *by*-passives

Stative forms

- be_main_verb: *be* as main verb
- existential_there: Existential *there*

Subordination features

- that_verb_comp: *that* verb complements (e.g., *I said [that he went]*)
- that_adj_comp: *that* adjective complements (e.g., *I’m glad [that you like it]*)
- wh_clause: *wh*-clauses (e.g., *I believed [what he told me]*)
- infinitives: Infinitives
- present_participle: Present participial adverbial clauses (e.g., *[Stuffing his mouth with cookies], Joe ran out the door.*)
- past_participle: Past participial adverbial clauses (e.g., *[Built in a single week], the house would stand for fifty years.*)
- past_participle_whiz: Past participial postnominal (reduced relative) clauses (e.g., *the solution [produced by this process]*)

- **present_participle_whiz**: Present participial postnominal (reduced relative) clauses (e.g., *the event [causing this decline]*)
- **that_subj**: *that* relative clauses on subject position (e.g., *the dog [that bit me]*)
- **that_obj**: *that* relative clauses on object position (e.g., *the dog [that I saw]*)
- **wh_subj**: *wh*-relatives on subject position (e.g., *the man [who likes popcorn]*)
- **wh_obj**: *wh*-relatives on object position (e.g., *the man [who Sally likes]*)
- **pied_piping**: Pied-piping relative clauses (e.g., *the manner [in which he was told]*)
- **sentence_relatives**: Sentence relatives (e.g., *Bob likes fried mangoes, [which is disgusting].*)
- **because**: Causative adverbial subordinator (*because*)
- **though**: Concessive adverbial subordinators (*although, though*)
- **if**: Conditional adverbial subordinators (*if, unless*)
- **other_adv_sub**: Other adverbial subordinators (e.g., *since, while, whereas*)

Prepositional phrases, adjectives and adverbs

- **prepositions**: Total prepositional phrases
- **adj_attr**: Attributive adjectives (e.g., *the [big] horse*)
- **adj_pred**: Predicative adjectives (e.g., *The horse is [big].*)
- **adverbs**: Total adverbs

Lexical specificity

- **type_token**: Type–token ratio (including punctuation)
- **mean_word_length**: Average word length (excluding punctuation)

Lexical classes

- **conjuncts**: Conjuncts (e.g., *consequently, furthermore, however*)
- **downtoners**: Downtoners (e.g., *barely, nearly, slightly*)
- **hedges**: Hedges (e.g., *at about, something like, almost*)
- **amplifiers**: Amplifiers (e.g., *absolutely, extremely, perfectly*)
- **emphatics**: Emphatics (e.g., *a lot, for sure, really*)
- **discourse_particles**: Discourse particles (e.g., *well, now, anyway*)
- **demonstratives**: Demonstratives

Modals

- **modal_possibility**: Possibility modals (*can, may, might, could*)
- **modal_necessity**: Necessity modals (*ought, should, must*)
- **modal_predictive**: Predictive modals (*will, would, shall*)

Specialized verb classes

- **verb_public**: Public verbs (e.g., *assert, declare, mention*)
- **verb_private**: Private verbs (e.g., *assume, believe, know*)
- **verb_suasive**: Suasive verbs (e.g., *command, propose, insist*)
- **verb_seem**: *seem* and *appear*

Reduced forms and dispreferred structures

- **contractions**: Contractions
- **that_deletion**: Subordinator *that* deletion (e.g., *I think [he went]*)
- **stranded_preposition**: Stranded prepositions (e.g., *the candidate that I was thinking [of]*)
- **split_infinitive**: Split infinitives (e.g., *He wants [to convincingly prove] that ...*)
- **split_auxiliary**: Split auxiliaries (e.g., *They [were apparently shown] to ...*)

Co-ordination

- phrasal_coordination: Phrasal coordination (*N and N; Adj and Adj; V and V; Adv and Adv*)
- clausal_coordination: Independent clause coordination (*clause-initial and*)

Negation

- neg_synthetic: Synthetic negation (e.g., *No answer is good enough for Jones.*)
- neg_analytic: Analytic negation (e.g., *That isn't good enough.*)

D Extended Results

In Figure 5, Figure 6, and Figure 7, we show the full feature list results for Reddit. Moreover, in Figure 8, Figure 9, and Figure 10, we show the full feature list results for YouTube.

E Inference Experiments

For running GPT-5, we use the OpenAI¹² API. The costs of running inference on all data took around 112 USD. For running inference of the local models (Gemma3, Mistral, Qwen3), we make use of a large HPC cluster with hardware configurations comprising multiple nodes (depending on model size; e.g., 30B models require 4 nodes for training and 1 node for inference), each with node contains eight AMD MI250x GPU modules alongside a single 64-core AMD EPYC “Trento” CPU. The library we use for inference is vllm (Kwon et al., 2023). For all the experiments it resulted in around 16 GPU hours spent.

E.1 Environmental Impact

We acknowledge that conducting a large-scale analysis using LLMs comes with an environmental impact. Experiments were conducted using private infrastructure in [Redacted] running on green energy. A cumulative of 16 GPU hours of computation was performed on AMD MI250x GPU modules, which has a TDP of 500 Watts. The experiments were ran in September 2025. During this time, the average carbon efficiency in [Redacted] was 0.046 *kg/kWh*.¹³ This means we released about 0.368 *kg* of *CO*₂ equivalent. Estimations

were conducted using the Machine Learning Impact calculator¹⁴ presented in (Lacoste et al., 2019).

¹²<https://platform.openai.com/>

¹³According to <https://app.electricitymaps.com/map>.

¹⁴Find the tool here: <https://mlco2.github.io/impact>.

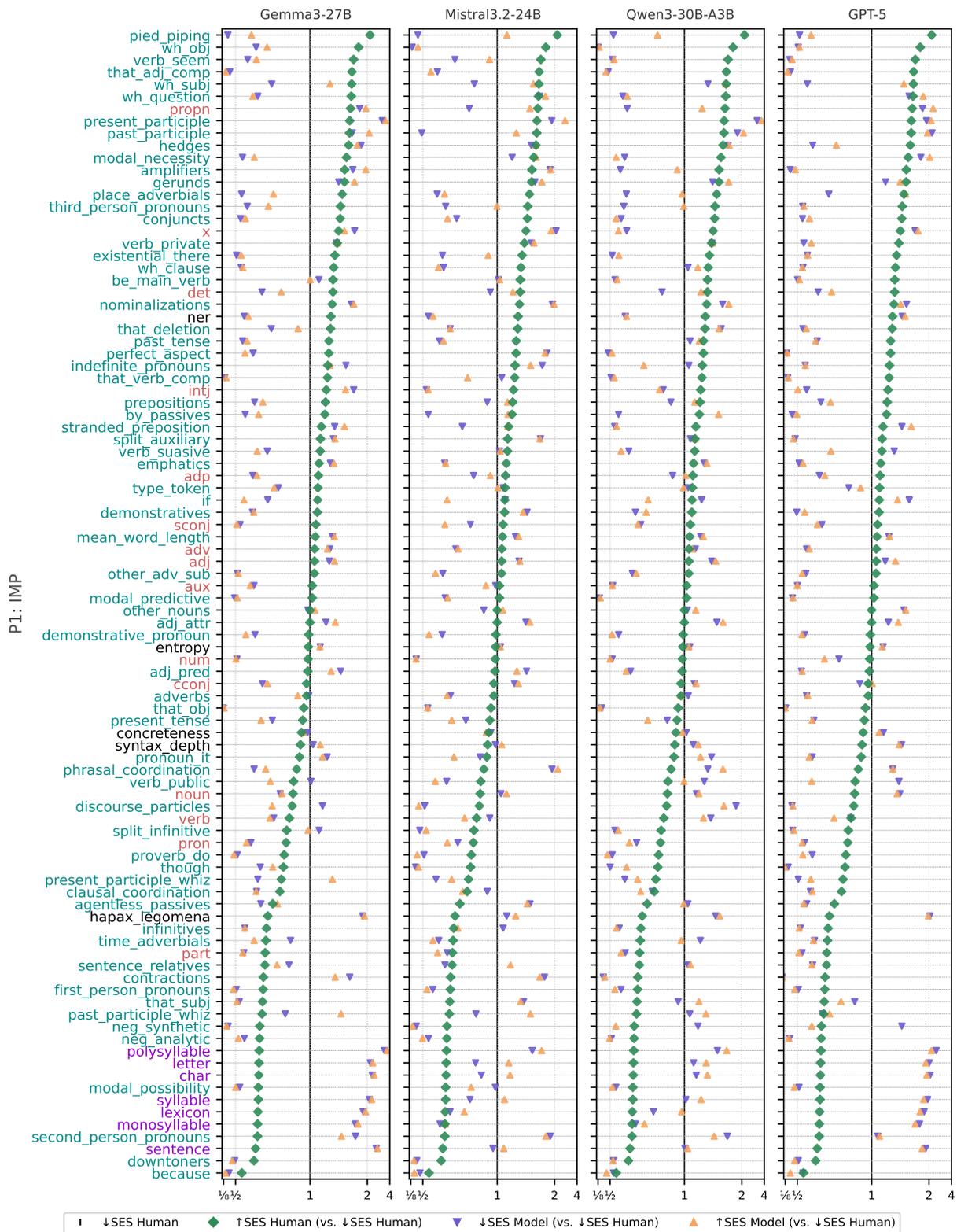


Figure 5: **Forest Plots Comparing Linguistic Features of Humans and Models on Reddit; Prompt 1, Full Features.** The plots display linguistic features between lower SES (\downarrow SES; rate = 1) and upper SES (\uparrow SES) human writers. Each point indicates the frequency ratio of a feature in a model’s (or human’s) output compared to human text from the \downarrow SES group. These comparisons are shown across four models and three prompts (see Section 4). Feature types are color-coded: Biber features are cyan, length-specific features are dark violet, PoS tags are red, and style features are black.



Figure 6: **Forest Plots Comparing Linguistic Features of Humans and Models on Reddit; Prompt 2, Full Features.** The plots display linguistic features between lower SES (\downarrow SES; rate = 1) and upper SES (\uparrow SES) human writers. Each point indicates the frequency ratio of a feature in a model’s (or human’s) output compared to human text from the \downarrow SES group. These comparisons are shown across four models and three prompts (see Section 4). Feature types are color-coded: Biber features are cyan, length-specific features are dark violet, PoS tags are red, and style features are black.

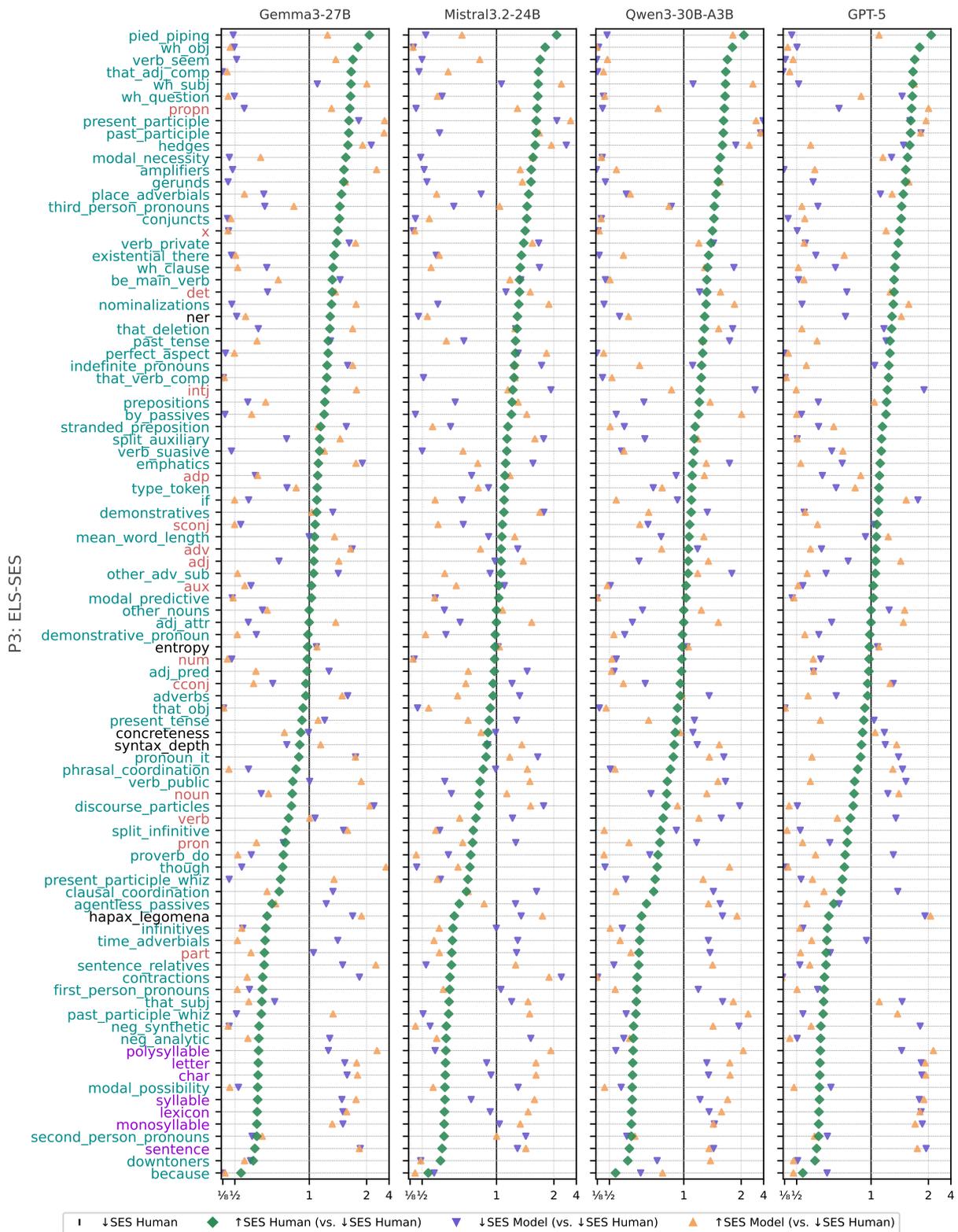


Figure 7: **Forest Plots Comparing Linguistic Features of Humans and Models on Reddit; Prompt 3, Full Features.** The plots display linguistic features between lower SES (\downarrow SES; rate = 1) and upper SES (\uparrow SES) human writers. Each point indicates the frequency ratio of a feature in a model’s (or human’s) output compared to human text from the \downarrow SES group. These comparisons are shown across four models and three prompts (see Section 4). Feature types are color-coded: Biber features are cyan, length-specific features are dark violet, PoS tags are red, and style features are black.

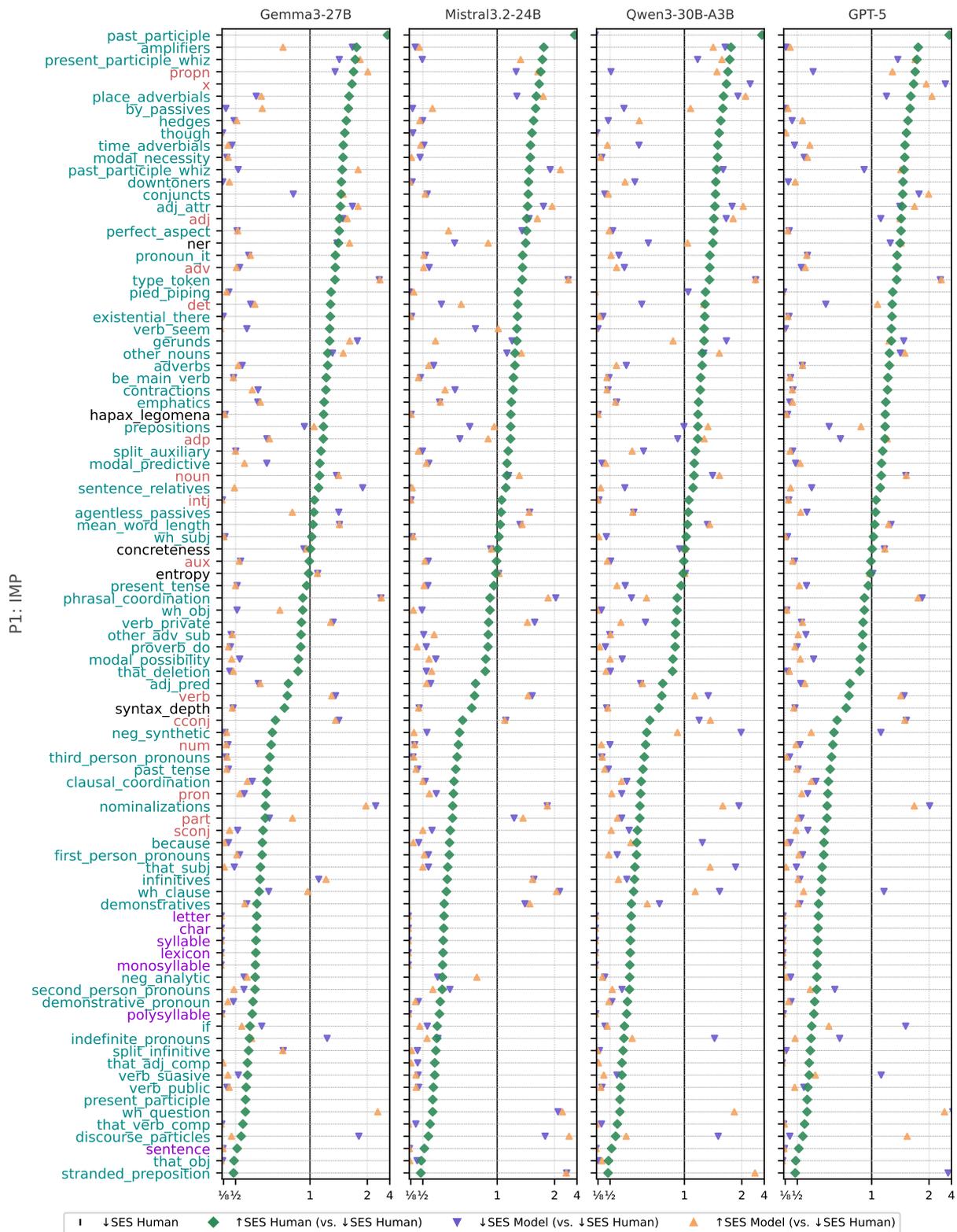


Figure 8: **Forest Plots Comparing Linguistic Features of Humans and Models on Youtube; Prompt 1, Full Features.** The plots display linguistic features between lower SES (\downarrow SES; rate = 1) and upper SES (\uparrow SES) human writers. Each point indicates the frequency ratio of a feature in a model’s (or human’s) output compared to human text from the \downarrow SES group. These comparisons are shown across four models and three prompts (see Section 4). Feature types are color-coded: Biber features are cyan, length-specific features are dark violet, PoS tags are red, and style features are black.

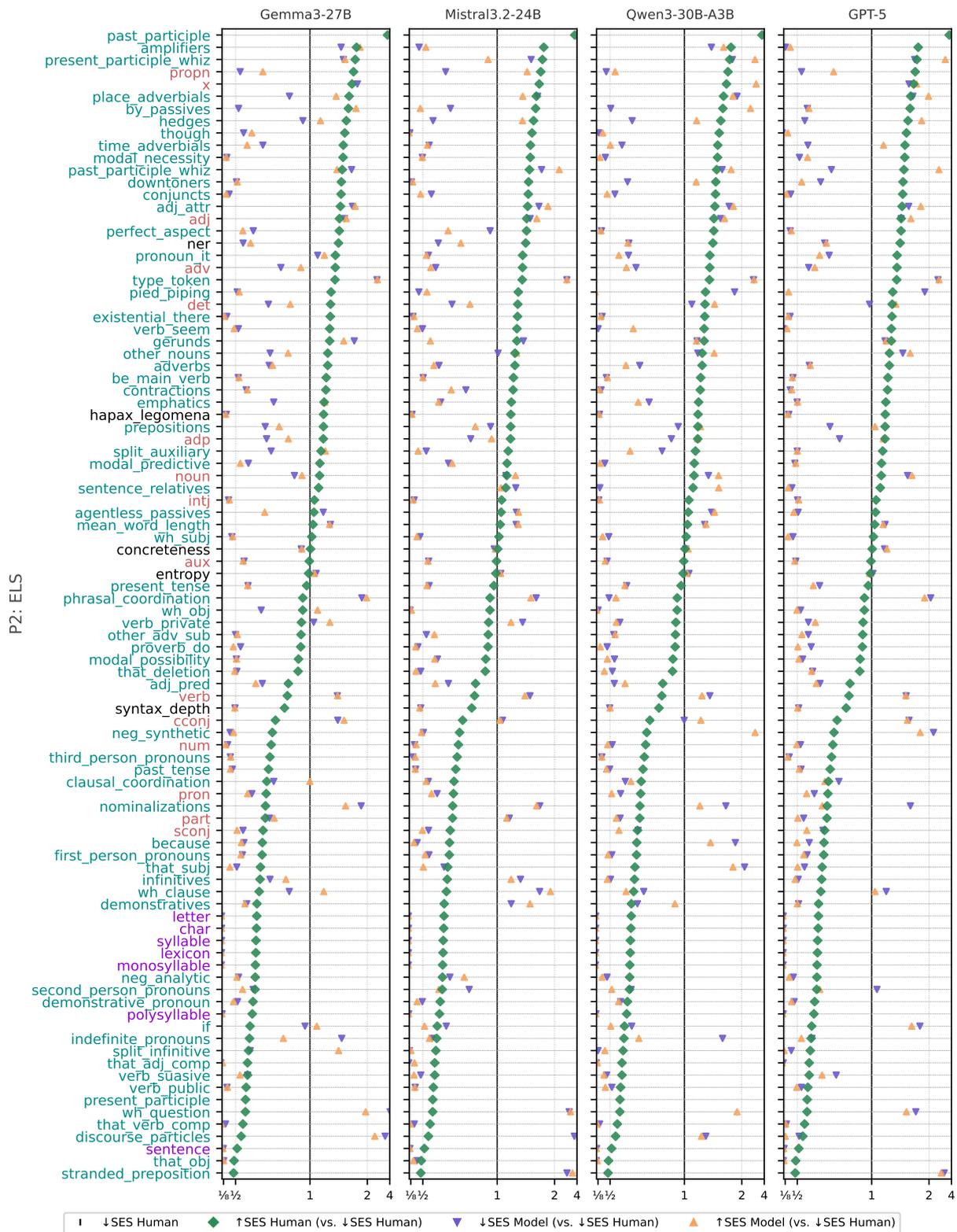


Figure 9: **Forest Plots Comparing Linguistic Features of Humans and Models on Youtube; Prompt 2, Full Features.** The plots display linguistic features between lower SES (\downarrow SES; rate = 1) and upper SES (\uparrow SES) human writers. Each point indicates the frequency ratio of a feature in a model’s (or human’s) output compared to human text from the \downarrow SES group. These comparisons are shown across four models and three prompts (see Section 4). Feature types are color-coded: Biber features are cyan, length-specific features are dark violet, PoS tags are red, and style features are black.

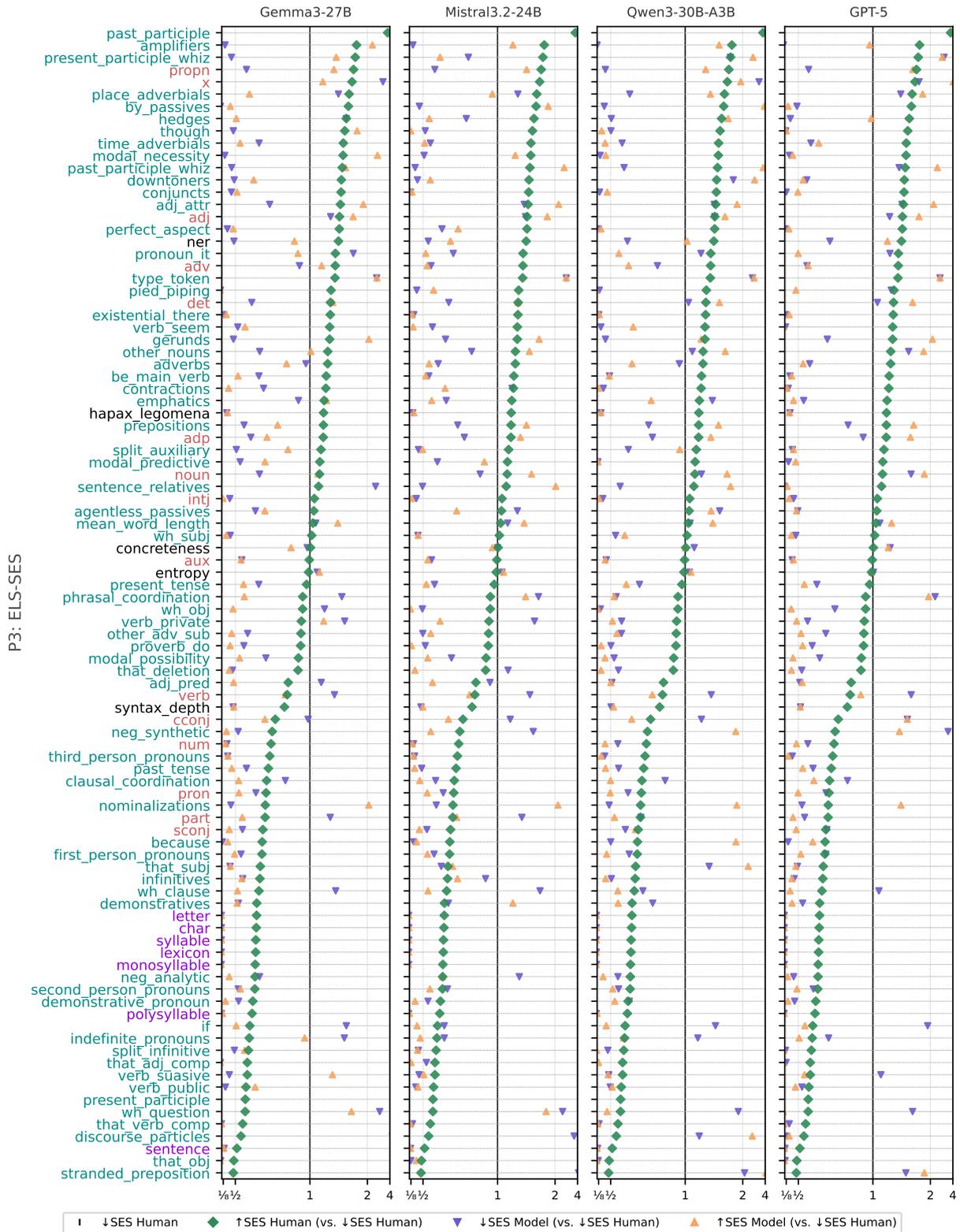


Figure 10: **Forest Plots Comparing Linguistic Features of Humans and Models on Youtube; Prompt 3, Full Features.** The plots display linguistic features between lower SES (\downarrow SES; rate = 1) and upper SES (\uparrow SES) human writers. Each point indicates the frequency ratio of a feature in a model’s (or human’s) output compared to human text from the \downarrow SES group. These comparisons are shown across four models and three prompts (see Section 4). Feature types are color-coded: Biber features are cyan, length-specific features are dark violet, PoS tags are red, and style features are black.

Aladdin-FTI @ AMIYA

Three Wishes for Arabic NLP: Fidelity, Diglossia, and Multidialectal Generation

Jonathan Mutal 🦄 Perla Al Almaoui 🦄 Simon Hengchen 🦄 🌿 Pierrette Bouillon 🦄

🦄 Faculté de traduction et d'interprétation, Université de Genève

🌿 iguanodon.ai

Correspondence: first.last@unige.ch

Abstract

Arabic dialects have long been under-represented in Natural Language Processing (NLP) research due to their non-standardization and high variability, which pose challenges for computational modeling. Recent advances in the field, such as Large Language Models (LLMs), offer promising avenues to address this gap by enabling Arabic to be modeled as a pluricentric language rather than a monolithic system. This paper presents Aladdin-FTI, our submission to the AMIYA shared task. The proposed system is designed to both generate and translate dialectal Arabic (DA). Specifically, the model supports text generation in Moroccan, Egyptian, Palestinian, Syrian, and Saudi dialects, as well as bidirectional translation between these dialects, Modern Standard Arabic (MSA), and English. The code and trained model are publicly available.¹

1 Introduction

The Thirteenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2026)² introduces the “Arabic Modeling In Your Accent” (AMIYA) shared task (Robinson et al., 2026), a benchmark designed to advance computational modeling of DA. The AMIYA shared task focuses on developing language models that capture the linguistic characteristics of spoken Arabic varieties. Such varieties remain under-represented in existing NLP research and resources (Harrat et al., 2019), although there is a growing interest in studying dialectal varieties and more resources are being created (see e.g. Al-Haff et al. (2022); Momayiz et al. (2024); Al Almaoui et al. (2025)). In this evaluation campaign, systems are assessed on their ability to model DA with respect to dialectal fidelity,

¹Code: https://github.com/drvenabili/mtfinetune_amiya, models: <https://hf.co/collections/unige-fti/aladdin-fti-amiya>.

²<https://sites.google.com/view/vardial-2026>

understanding, and generation quality using the AL-QASIDA benchmark (Robinson et al., 2025).

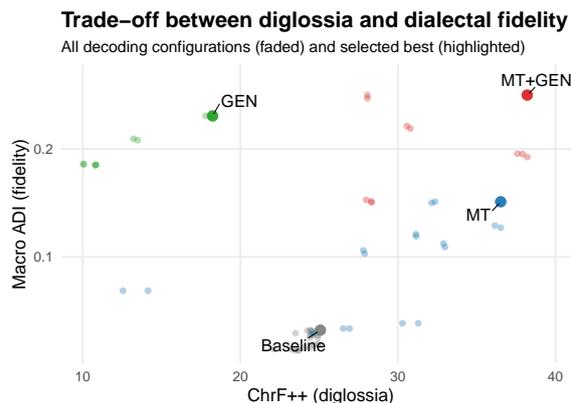


Figure 1: Trade-off between diglossia-sensitive translation accuracy (ChrF++) and dialectal fidelity (Macro ADI2). Each faded point corresponds to a decoding configuration (learning rate \times checkpoint), while highlighted points indicate the best configuration selected per model. Instruction-based generation (GEN) favours dialectal fidelity at the expense of diglossia, whereas MT exhibits the opposite behaviour. The combined MT+GEN objective achieves the best overall, improving both fidelity and diglossia.

This paper describes the participation of the team **Aladdin-FTI** 🦄, focusing on the closed data track, where the models are only fine-tuned on the official training data provided by the shared task organizers, without the use of additional external corpora. Our approach is based on translation and generation by combining two training objectives: (i) a translation objective aimed at reinforcing diglossic distinctions between MSA and DA, while also preserving semantic adequacy; and (ii) an instruction next-token generation objective designed to produce fluent and linguistically dialectal continuations from partial prompts. By jointly training with these objectives, we seek to have a balance between semantic adequacy and dialectal expressiveness (Robinson et al., 2025). We investigate the complementary roles of

translation and generation in dialectal Arabic modeling along the following questions:

- Q1** What is the impact of fine-tuning for translation on diglossia and dialectal fidelity across Arabic dialects?
- Q2** What is the impact of instruction fine-tuning for next word generation on diglossia and dialectal fidelity across Arabic dialects?
- Q3** What is the impact of both machine translation (MT) and instruction fine-tuning for next word generation on diglossia and dialectal fidelity across Arabic dialects?

We fine-tune a single large language model under different training settings and evaluate their impact on both diglossia and dialectal fidelity. Our results highlight their distinct yet complementary roles in DA generation.

Our contributions are the following:

- We propose a joint training objective that combines machine translation and instruction-conditioned next-token generation for dialectal Arabic.
- This training enables smaller models to match or even outperform substantially larger baselines in modeling Arabic dialectal variation.

The remainder of this paper is organized as follows: first, Section 2 reviews related work; next, Section 3 presents the methodology; Section 4 describes the experimental setup; Section 5 reports the results; and finally, Section 6 concludes by discussing the limitations of the study.

2 Related Work

Arabic has long been treated as a single homogeneous language, with the majority of resources, benchmarks, and models focusing almost exclusively on MSA. However, this perspective overlooks the deeply diglossic nature of Arabic-speaking communities, in which MSA is rarely a native language and is primarily used in formal and written contexts, while everyday communications take place in regionally and socially diverse dialects (Keleg et al., 2025). These dialects differ substantially from MSA in phonology, morphology, syntax, and lexicon, and each reflects the historical, cultural, and social identities of its speakers (Yushmanov, 1961).

This MSA-centric focus poses significant challenges for NLP systems, as models trained predominantly on MSA tend to normalize or suppress dialectal features when generating or translating dialectal text (Robinson et al., 2023). This gap motivates research into methods that explicitly preserve dialectal characteristics in generated text while maintaining semantic adequacy.

Machine Translation for Dialectal Arabic Prior work has shown that MT provides a natural framework for modeling distinctions between MSA and DA, as translation objectives explicitly condition generation on a source sentence rather than on the target alone (Habash et al., 2013; Zbib et al., 2012). More recent efforts include the fine-tuning of the Kuwain-1.5B small language model for translation from 15 Arabic dialects into Modern Standard Arabic, which achieved high human-rated fluency scores in evaluation studies (Hamed et al., 2025), indicating improved generation quality for dialectal inputs.

Despite these advances, multiple studies report that dialectal MT systems often favour normalised outputs, exhibiting MSA lexical or morphosyntactic choices even when dialectal targets are explicitly specified (Habash et al., 2013; Bouamor et al., 2018; Robinson et al., 2025). While translation-based approaches tend to preserve meaning effectively, they may under-represent dialect-specific variation and linguistic naturalness, a pattern that has also been observed in recent shared-task evaluations (Atwamy et al., 2024; Robinson et al., 2023).

Motivated by these findings, we evaluate the impact of machine translation training on diglossia and dialect fidelity, following Q1.

Dialectal Text Generation with LLMs Instruction fine-tuning has demonstrated strong performance in controllable text generation (Liang et al., 2024). In dialectal settings, explicit instruction conditioning and the use of dialect tokens have been shown to improve alignment between generated outputs and target dialectal varieties (Barmandah, 2025). However, prior work indicates that such approaches may also introduce generation artifacts when control constraints are emphasized, with potential negative effects on semantic fidelity (Zhang et al., 2023). We therefore evaluate the impact of instruction fine-tuning using next-word prediction on diglossia and dialect fidelity in the context of Arabic dialects (Q2).

Combining the Two Tasks Despite advances in both tasks, combining translation-based and generation-based techniques for DA remains underexplored. Prior research on Arabic NLP has begun to explore multi-task learning paradigms, for example, joint modeling of dialect identification and translation to improve MT quality (Khered et al., 2025). There have also been efforts in other settings (e.g. unsupervised MT) to couple translation objectives with language modeling to better preserve fluency (Artetxe et al., 2018). Yet, to our knowledge, no prior work has explicitly jointly optimized a large language model for translation and dialect generation in the Arabic diglossia context. This gap motivates our approach to combine MT and next-word completion (Q3).

3 Methodology

In this section, we go through the evaluation protocol (Subsection 3.1) and the training objectives (Subsection 3.2).

3.1 Evaluation Protocol

We followed the evaluation framework proposed by Robinson et al. (2025) to assess two complementary dimensions of DA generation: fidelity and diglossia.

Fidelity This dimension is evaluated using both monolingual and cross-lingual prompts, in which the model is instructed to generate text in a specific DA variety. In such settings, no single gold reference exists, as multiple valid outputs may correspond to the same prompt. Accordingly, fidelity is measured using the Macro ADI2 dialect fidelity score, which assesses whether the generated output is both dialectal and identifiable as the target variety.

Diglossia Diglossia evaluates the model’s ability to translate between MSA and DA, reflecting its capacity to differentiate between dialects and MSA. This dimension is assessed through bidirectional translation tasks (MSA→DA and DA→MSA). For these tasks, reference translations are available and performance is measured using ChrF++ (Popović, 2015).

Together, these two dimensions provide a complementary evaluation of dialectal Arabic generation: fidelity focuses on adherence to the target dialect in open-ended settings, while diglossia assesses controlled meaning under reference-based translations.

For each evaluation dimension, we compute the mean score over the different datasets. Specifically, machine translation (corpus-level) and fidelity (sentence-level) are each evaluated on their own dedicated datasets.

3.2 Training Objectives

The two training objectives previously mentioned in section 1 differ not in their optimization procedure but in the constraints imposed by the task formulation. Translation provides a reference to enforce meaning preservation with respect to a source sentence, resulting in a constrained output space. In contrast, dialectal generation is open-ended: for a given instruction and prefix, multiple continuations may be equally valid as long as they are the target dialect (or have linguistic similarities).

Formally, let \mathcal{D}_{MT} and \mathcal{D}_{GEN} denote the instruction-formatted datasets used for translation and dialectal generation, respectively. All training examples are represented using an instruction-based format and optimized with a causal language model objective, where only assistant tokens contribute to the loss. The final training objective minimizes a weighted combination of losses over the two datasets:

$$\mathcal{L}_{\text{joint}}(\theta) = \lambda \mathbb{E}_{\mathbf{z} \sim \mathcal{D}_{\text{MT}}} [\mathcal{L}(\mathbf{z})] + (1 - \lambda) \mathbb{E}_{\mathbf{z} \sim \mathcal{D}_{\text{GEN}}} [\mathcal{L}(\mathbf{z})] \quad (1)$$

where $\lambda \in [0, 1]$ controls the relative contribution of translation and generation supervision.

We evaluate the model with $\lambda \in \{0, 0.5, 1\}$. When $\lambda = 0$, the task corresponds to pure generation, whereas when $\lambda = 1$, it corresponds to a MT task.

3.2.1 Machine Translation

For the translation objective, each training example consists of an instruction in English specifying the translation direction (e.g. MSA→DA, DA→MSA, or DA↔English), followed by an assistant response containing the target sentence. After applying the chat template, the model is trained to maximize the conditional likelihood of the assistant tokens given the full preceding context. The template for this task is shown in Table 2.

Prompt examples for translation are illustrated in Table 4.

3.2.2 Instruction Next-Token Generation

For dialectal generation, training examples are formulated as instruction-conditioned sentence completion tasks. Each example provides an explicit

Instruction type	Template	Completion
English instruction	Complete the sentence starting with these 3 words in <TARGET DIALECT>: <PREFIX>	This is the full sentence in <TARGET DIALECT>: <TARGET TEXT IN DIALECT>
Dialectal instruction	<TARGET DIALECT> باللهجة كلمات دول بالتلات وابدأ بالتلات ككَل <PREFIX>:	دي الجملة كاملة باللهجة <TARGET DIALECT>:

Table 1: Templates used for dialectal sentence completion in MADAR training data. Both templates require the model to generate a complete sentence in the target dialect starting from a fixed prefix; only the language of the instruction differs.

Instruction + Source	Target (Reference or Output)	Language Pair
Translate from <SRC> into <TGT>. <SOURCE SENTENCE> Translation:	<TARGET SENTENCE>	<SRC> → <TGT>

Table 2: Template for instructing the translation across Arabic varieties and English.

instruction specifying the target dialect, followed by the first three words of a sentence, which serve as a fixed prefix. The model is trained to generate a complete sentence in the target dialect, starting from this prefix.

Two instruction variants are used. In the first variant, the instruction is formulated in English and specifies both the completion task and the target dialect. In the second variant, the instruction is formulated directly in the target dialect. In both cases, the assistant response contains a full sentence that repeats the provided prefix and continues it in a linguistically coherent manner. Although reference continuations are provided during training, they are not assumed to be unique, as the task may admit multiple valid dialectal realizations for the same prefix. Thus, the training signal encourages the model to learn distributional properties of the target dialect rather than to reproduce a single fixed continuation. The generation templates are provided in Table 1 and examples are shown in Table 5.

4 Experimental Set-Up

4.1 Models

After a hyper-parameter search, we selected SmolLM3-3B to carry out our experiments (Bakouch et al., 2025)³ as it offers a good balance between model size, performance across tasks, and has been trained with Arabic data.

We instruction fine-tuned SmolLM3-3B to support multiple training objectives and evaluation regimes, selecting the best model according to both MT and next-token generation performance. Periodic evaluation was performed every 1,000 steps

³<https://huggingface.co/HuggingFaceTB/SmolLM3-3B>

using the validation loss and a character-based metric (ChrF++). This design enables a direct comparison of models trained on (a) MT only, (b) MT combined with next-token generation, and (c) next-token generation only, while keeping the evaluation procedure consistent across experimental conditions. We let the reader refer to the model training in Appendix A.1.

To address the research questions, we compare these training configurations against a baseline. The baseline corresponds to SmolLM3-3B with a hyperparameter search over decoding settings ($top-p \in \{0.1, 0.3, 0.6, 0.9, 1\}$ and $temperature \in \{0.1, 0.3, 0.6, 0.9, 1\}$). We shorten the original template by one third to train our instruction models. We found that both templates provided similar results in preliminary evaluations.

To assess the effect of scaling to a larger model, we replicate the same experimental setting using Llama-3.1-8B-Instruct⁴ as the base model, fine-tuned with LoRA (see Appendix A.1). SmolLM-3B builds upon the Llama architecture, with modifications optimized for efficiency and long-context performance, which makes it a suitable point of comparison with Llama-3.1-8B-Instruct. We additionally compare our approach with a larger model to assess its effectiveness relative to models of different sizes, using the best configuration identified during hyperparameter search (refer to Appendix A.4).

4.2 Evaluation Data

We adopted the same evaluation data and protocol described in Al Qasida (Robinson et al., 2025). Our evaluation set comprises both monolingual and crosslingual prompts across multiple DA varieties, designed to support text generation and MT tasks. The crosslingual prompts were drawn from three different collections of LLMs user inputs: (i) a subset of Okapi prompts (Lai et al., 2023) used

⁴<https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>

with the Alpaca LLM (Taori et al., 2023), (ii) a collection of ChatGPT prompts obtained via the ShareGPT API, and (iii) a set of human-curated prompts by Marchisio et al. (2024). In addition, our evaluation corpus incorporates both monolingual and bitext sentences from the corpus-6-test-corpus-26-dev split of the MADAR26 (Bouamor et al., 2018), a large-scale parallel resource covering dialects from seven Arab countries and consisting of BTEC-style everyday utterances originally introduced by Takezawa et al. (2007). We further included data from the FLORES200 dev, a multilingual benchmark based on manually translated Wikipedia text, selecting dialectal Arabic subsets representing five major regional varieties (NLLB Team, 2022). Finally, we integrated dialectal Arabic song lyrics from the HABIBI corpus, which spans eight Arab country dialects (El-Haj, 2020).

4.3 Training Data

Following the closed-data track,⁵ we used two sources of training data: bilingual data for task (i and ii, refer to Section 3.2.1 and 3.2.2 respectively) and monolingual data for task (ii).

Bilingual Data Our bilingual training data were used to support translation tasks between English, MSA, and DA. For Saudi Arabic, the SauDial corpus (Alanazi et al., 2025) provided parallel data for EN↔DA and DA↔MSA translation. Palestinian Arabic–English parallel data were sourced from the Casablanca corpus (Talafta et al., 2024). For Jordanian Arabic, the JODA corpus (Abandah et al., 2025) was used, offering parallel data between dialectal text and its MSA-corrected version. Syrian Arabic bilingual resources included the UFAL parallel corpus (Krubinski et al., 2023), covering MSA↔DA and DA↔EN translation directions. Moroccan Arabic bilingual data combined several sources: the DODA corpus (Outchakoucht and Es-Samaali, 2024) for EN↔DA translation, and the Atlas training sets (Bounhar and Majjodi, 2025). These data were used to create the machine translation training data (see Section 3.2.1).

Monolingual Data The monolingual training data were compiled from multiple resources covering a wide range of DA varieties. Saudi Arabic monolingual data were obtained from the SDC corpus (Tarmom et al., 2014) as well as from the Saudi

Tweets Corpus (Alruily, 2018). The Shami corpus (Abu Kwaik et al., 2018) was used to provide monolingual data for Palestinian, Syrian, and Jordanian varieties, while the MASC corpus (Al-Fetyani et al., 2021) contributed data for Egyptian and Jordanian Arabic. Moroccan Arabic monolingual data were sourced from the Goud training set (Aftiss et al., 2025). Egyptian Arabic monolingual data were enriched using the EDGAD corpus (ElSayed and Farouk, 2020; Hussein et al., 2019), the EDC corpus (Tarmom et al., 2014), and the ASR-EgAr corpus (asr, 2023).

We also incorporated multidialectal monolingual data from the MADAR training set, which includes additional Arabic dialects. These data were used to create the instruction next-token generation data (see Section 3.2.2).

5 Results

Effect of the Task Figure 2 shows the scores with the different training objectives for the diglossia and fidelity tasks. Training with a translation objective improves diglossic scores, as reflected by higher ChrF++ across configurations (Q1). While MT also increases dialectal fidelity compared to the baseline, the gains remain limited and highly variable, suggesting that the LLM does not consistently generate the target dialect according to the Macro ADI2 score.

Instruction next-token generation improves dialectal fidelity, yielding the highest Macro ADI2 scores with low variance across configurations. However, this comes at the cost of diglossia, as generation-only models perform poorly on translation tasks, indicating a semantic drift (answering our Q2).

Jointly optimizing translation and instruction-based generation objectives yields a balance between diglossic scores and dialectal fidelity. Compared to MT training, the joint model substantially improves Macro ADI2 without degrading ChrF++ (avoiding the semantic shift observed in generation-only models). This indicates that translation and generation objectives provide complementary supervision signals for modeling Arabic dialects.

reprint

Trade-off Between the Tasks Figure 1 illustrates the trade-off between diglossia (measured by ChrF++) and dialectal fidelity, measured by Macro ADI2, across all decoding configurations. Faded points represent individual decoding configurations,

⁵<https://sites.google.com/view/wardial-2026/shared-tasks>

Model	Diglossia \uparrow	Fidelity \uparrow
SmolLM3-3B	22.23	0.003
+ Machine Translation + Instruction (Constrastive)	33.65	0.067
Llama-3.1-8B-Instruct	32.99	0.065
+ Machine Translation + Instruction (Primary)	35.09	0.233
Command R Arabic	46.60	0.053
GPT-OSS-120B	47.82	0.237

Table 3: Comparison of baseline LLMs in terms of diglossia (ChrF++) and dialectal fidelity (Macro ADI2) across all language. For each model, scores correspond to the best decoding configuration selected across temperature and top- p sampling settings.

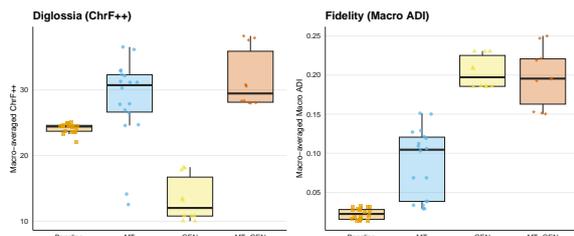


Figure 2: Performance for diglossia (ChrF++) and fidelity (Macro ADI2) across training paradigms. Each boxplot corresponds to a training paradigm (Baseline, MT, GEN, MT+GEN) using SmolLM3-3B, and each point represents a distinct decoding configuration (top- p , temperature, learning rate), with scores macro-averaged over language varieties and test sets.

highlighting the variability induced by learning rate and checkpoint selection (also decoding hyperparameters for the baseline), while the highlighted points correspond to the best-performing configuration selected per model.

The baseline model clusters in the lower-left region of the plot, exhibiting both limited diglossia and weak dialectal fidelity. Models trained exclusively with a MT objective achieve higher ChrF++ scores, indicating stronger diglossia, but remain constrained in dialectal fidelity. In contrast, instruction-based next-token generation (GEN) prioritizes dialectal generation, yielding higher Macro ADI2 scores at the cost of reduced diglossia scores.

The combined MT+GEN model is in the upper-right region of the plot, demonstrating that jointly optimizing translation and instruction-based generation objectives leads to a more favorable balance between semantic adequacy and dialectal expressiveness (**Q3**).

Comparison with Other LLMs To better understand the impact of joint training, we also carried out the same experiments using Llama-3.1-8B-

Instruct⁶ as the base model, fine-tuned with LoRA.

In addition, we included a 120B-parameter model⁷ as a reference point to assess how performance scales with substantially larger model parameters (without any optimization method). We also considered Command R Arabic, a model specifically optimized for translation across multiple Arabic varieties⁸, in order to compare our approach against a system designed for Arabic multilingual translation. Taken together, these baselines provide an approximate upper bound on the performance and help contextualize the results of smaller jointly trained models.

Table 3 reports the performance of jointly trained models in comparison with their base model. The automatic evaluation scores indicate that, joint training across tasks leads to consistent improvements in metrics reflecting both diglossia and fidelity scores even for larger model (Llama-3.1-8B-Instruct) with LoRA fine-tuning. These results provide further evidence in support of the benefits of jointly training the model with both task (**Q3**).

Compared to the other models, the much larger GPT-OSS-120B attains the highest scores overall. Despite being substantially smaller, Llama-3.1-8B-Instruct reaches the fidelity score (Macro ADI2) of 0.23, achieving performance comparable to the 120B model. This suggests that dialectal control can be enhanced through supervision strategies rather than model scaling alone.

6 Conclusion

This work presented the submission of team **Aladdin-FTI** 🇪🇬 to the AMIYA shared task, aiming to model Arabic dialects through a uni-

⁶<https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>

⁷<https://huggingface.co/openai/gpt-oss-120b>

⁸<https://huggingface.co/CohereLabs/c4ai-command-r7b-arabic-02-2025>

fied framework that combines translation and instruction-based generation.

Our results show that these objectives provide complementary supervision: translation improves diglossic awareness and semantic adequacy (**Q1**), while instruction-conditioned generation enhances dialectal fidelity (**Q2**). By combining both objectives, we obtain a more balanced model that consistently outperforms single-objective approaches across evaluation dimensions (**Q3**).

Notably, this balance is achieved with a smaller model that competes with larger systems, underscoring the importance of training objective design in dialectal Arabic modeling. These findings support treating Arabic as a pluricentric language.

Future work will focus on refining the balance between objectives, expanding coverage to additional varieties, and conducting human and linguistic evaluations to better assess dialectal naturalness.

Limitations

This study is limited to a single model architecture. While the results are encouraging, further experiments on different model families and scales are needed to assess the generality of the proposed approach. While few-shot and in-context learning approaches may be effective (see e.g. [Gao et al. \(2021\)](#); or [Mutal et al. \(2025\)](#) for use in low-resource settings), they were not considered in this work, as our objective was to keep input prompts compact and limit the number of tokens provided to the model.

Acknowledgments

The computations were performed at the University of Geneva using the Baobab HPC service.

References

2023. ASR-EgArbCSC: An Egyptian Arabic conversational speech corpus. Open-source dataset consisting of 5.5 hours of transcribed Egyptian Arabic conversational speech across nine two-speaker conversations.

Gheith Abandah, Ashraf Suyyagh, Iyad Jafar, Mohammad Abdel-Majeed, Rabie Otoum, Shorouq AlAwawdeh, and Moath Khaleel. 2025. [JODA: A dataset of jordanian dialect and erroneous modern arabic sentences coupled with proper MSA and full diacritics](#).

Kathrein Abu Kwaik, Motaz Saad, Stergios Chatzikyriakidis, and Simon Dobnik. 2018. [Shami: A corpus of Levantine Arabic dialects](#). In *Proceedings of*

the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

- Azzedine Aftiss, Salima Lamsiyah, Christoph Schommer, and Said Ouatik El Alaoui. 2025. Empirical evaluation of pre-trained language models for summarizing moroccan darija news articles. In *Proceedings of the 4th Workshop on Arabic Corpus Linguistics (WACL-4)*, pages 77–85.
- Perla Al Almaoui, Pierrette Bouillon, and Simon Hengchen. 2025. [Arabizi vs LLMs: Can the genie understand the language of aladdin?](#) In *Proceedings of Machine Translation Summit XX: Volume 2*, pages 28–41, Geneva, Switzerland. European Association for Machine Translation.
- Mohammad Al-Fetyani, Muhammad Al-Barham, Gheith Abandah, Adham Alsharkawi, and Maha Dawas. 2021. [Masc: Massive arabic speech corpus](#).
- Karim Al-Haff, Mustafa Jarrar, Tymaa Hammouda, and Fadi Zaraket. 2022. [Curras + baladi: Towards a Levantine corpus](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 769–778, Marseille, France. European Language Resources Association.
- Naif Alanazi, Mohammed Al-Batineh, and Hussein Abu-Rayyash. 2025. [Saudial: The saudi arabic dialects game localization dataset](#). *Data in Brief*, 62:111906.
- Meshrif Alruily. 2018. [Saudi tweets dataset](#).
- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. [Unsupervised statistical machine translation](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642, Brussels, Belgium. Association for Computational Linguistics.
- Hanin Atwany, Nour Rabih, Ibrahim Mohammed, Abdul Waheed, and Bhiksha Raj. 2024. [OSACT 2024 task 2: Arabic dialect to MSA translation](#). In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation @ LREC-COLING 2024*, pages 98–103, Torino, Italia. ELRA and ICCL.
- Elie Bakouch, Loubna Ben Allal, Anton Lozhkov, Noumane Tazi, Lewis Tunstall, Carlos Miguel Patiño, Edward Beeching, Aymeric Roucher, Aksel Joonas Reedi, Quentin Gallouédec, Kashif Rasul, Nathan Habib, Clémentine Fourrier, Hynek Kydlicek, Guilherme Penedo, Hugo Larcher, Mathieu Morlon, Vaibhav Srivastav, Joshua Lochner, and 4 others. 2025. SmoLLM3: smol, multilingual, long-context reasoner. <https://huggingface.co/blog/smollm3>.
- Hassan Barmandah. 2025. [Saudi-dialect-allam: Lora fine-tuning for dialectal arabic generation](#). *Preprint*, arXiv:2508.13525.

- Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouni, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. [The MADAR Arabic dialect corpus and lexicon](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Abdelaziz Bounhar and Abdeljalil El Majjodi. 2025. [Atlaset dataset for moroccan darija: From data collection, analysis, to model trainings](#). *Hugging Face Blog*.
- Mahmoud El-Haj. 2020. Habibi: A multi-dialect arabic song lyrics corpus. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*.
- Shereen ElSayed and Mona Farouk. 2020. [Gender identification for egyptian arabic dialect in twitter using deep learning models](#). *Egyptian Informatics Journal*, 21.
- Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. [Making pre-trained language models better few-shot learners](#). 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 3816–3830, Online. Association for Computational Linguistics.
- Nizar Habash, Ryan Roth, Owen Rambow, Ramy Eskander, and Nadi Tomeh. 2013. [Morphological analysis and disambiguation for dialectal Arabic](#). In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 426–432, Atlanta, Georgia. Association for Computational Linguistics.
- Mohamed Motasim Hamed, Muhammad Hreden, Khalil Hennara, Zeina Aldallal, Sara Chrouf, and Safwan AlModhayan. 2025. [Lahjawi: Arabic cross-dialect translator](#). In *Proceedings of the 4th Workshop on Arabic Corpus Linguistics (WACL-4)*, pages 12–24, Abu Dhabi, UAE. Association for Computational Linguistics.
- Salima Harrat, Karima Meftouh, and Kamel Smaili. 2019. Machine translation for Arabic dialects (survey). *Information Processing & Management*, 56(2):262–273.
- 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](#). *Preprint*, arXiv:2106.09685.
- Shereen Hussein, Mona Farouk, and ElSayed Hemayed. 2019. [Gender identification of egyptian dialect in twitter](#). *Egyptian Informatics Journal*, 20(2):109–116.
- Amr Keleg, Sharon Goldwater, and Walid Magdy. 2025. [Revisiting common assumptions about Arabic dialects in NLP](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3327, Vienna, Austria. Association for Computational Linguistics.
- Abdullah Khered, Youcef Benkhedda, and Riza Batista-Navarro. 2025. [A multi-task learning approach to dialectal Arabic identification and translation to Modern Standard Arabic](#). In *Proceedings of the First Workshop on Advancing NLP for Low-Resource Languages*, pages 21–31, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.
- Mateusz Krubiński, Hashem Sellat, Shadi Saleh, Adam Pospíšil, Petr Zemánek, and Pavel Pecina. 2023. [Multi-parallel corpus of North Levantine Arabic](#). In *Proceedings of ArabicNLP 2023*, pages 411–417, Singapore (Hybrid). Association for Computational Linguistics.
- Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023. [Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327, Singapore. Association for Computational Linguistics.
- Xun Liang, Hanyu Wang, Yezhaohui Wang, Shichao Song, Jiawei Yang, Simin Niu, Jie Hu, Dan Liu, Shunyu Yao, Feiyu Xiong, and Zhiyu Li. 2024. [Controllable text generation for large language models: A survey](#). *Preprint*, arXiv:2408.12599.
- Kelly Marchisio, Wei-Yin Ko, Alexandre Berard, Théo Dehaze, and Sebastian Ruder. 2024. [Understanding and mitigating language confusion in LLMs](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6653–6677, Miami, Florida, USA. Association for Computational Linguistics.
- Imane Momayiz, Aissam Outchakoucht, Omar Choukrani, and Ali Nirheche. 2024. [Terjamabench: A culturally specific dataset for evaluating translation models for moroccan darija](#).
- Jonathan Mutal, Raphael Rubino, and Pierrette Bouillon. 2025. [Factors affecting translation quality in in-context learning for multilingual medical domain](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 161–179, Suzhou, China. Association for Computational Linguistics.
- NLLB Team. 2022. No language left behind: Scaling human-centered machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Aissam Outchakoucht and Hamza Es-Samaali. 2024. [The evolution of darija open dataset: Introducing version 2](#). *Preprint*, arXiv:2405.13016.

- Maja Popović. 2015. [chrF: character n-gram F-score for automatic MT evaluation](#). In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.
- Nathaniel Robinson, Shahd Abdelmoneim, Anjali Kantharuban, Otba Alsoul, Salima Lamsiyah, Kelly Marchisio, and Kenton Murray. 2026. AMIYA shared task: Arabic Modeling In Your Accent at VarDial 2026. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, Rabat, Morocco. Association for Computational Linguistics.
- Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. [ChatGPT MT: Competitive for high- \(but not low-\) resource languages](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.
- Nathaniel Romney Robinson, Shahd Abdelmoneim, Kelly Marchisio, and Sebastian Ruder. 2025. [AL-QASIDA: Analyzing LLM quality and accuracy systematically in dialectal Arabic](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22048–22065, Vienna, Austria. Association for Computational Linguistics.
- Toshiyuki Takezawa, Genichiro Kikui, Masahide Mizushima, and Eiichiro Sumita. 2007. [Multilingual spoken language corpus development for communication research](#). In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 12, Number 3, September 2007: Special Issue on Invited Papers from ISCSLP 2006*, pages 303–324.
- Bashar Talafha, Karima Kadaoui, Samar Mohamed Magdy, Mariem Habiboullah, Chafei Mohamed Chafei, Ahmed Oumar El-Shangiti, Hiba Zayed, Mohamedou Cheikh Tourad, Rahaf Alhamouri, Rwa Assi, and 1 others. 2024. Casablanca: Data and models for multidialectal arabic speech recognition. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21745–21758.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models*. <https://crfm.stanford.edu/2023/03/13/alpaca.html>, 3(6):7.
- Taghreed Tarmom, William Teahan, Eric Atwell, and Mohammad Alsalka. 2014. Compression vs traditional machine learning classifiers to detect code-switching in varieties and dialects: Arabic as a case study. *Journal of Natural Language Processing*.
- Nikolai Vladimirovich Yushmanov. 1961. The structure of the arabic language.
- Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stalard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan, and Chris Callison-Burch. 2012. [Machine translation of Arabic dialects](#). In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 49–59, Montréal, Canada. Association for Computational Linguistics.
- Zhihan Zhang, Wenhao Yu, Mengxia Yu, Zhichun Guo, and Meng Jiang. 2023. [A survey of multi-task learning in natural language processing: Regarding task relatedness and training methods](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 943–956, Dubrovnik, Croatia. Association for Computational Linguistics.

A LLM Settings and Results

A.1 Instruct Fine-Tuning

We fine-tuned two instruction-tuned models: SmolLM3-3B and Llama-3.1-8B-Instruct. Both models shared the same training configuration. Evaluation and checkpointing were performed every 1,000 steps, and the best-performing model was retained according to ChrF++ and perplexity. ChrF++ evaluation relied on deterministic text generation (temperature 0.0, top- p 1.0) with up to 512 generated tokens, and translation quality was assessed using ChrF++ with a character n-gram size of 64. The best checkpoint was selected by maximizing the macro-averaged ChrF++ score over the full development set.

All models were trained for four epochs with a per-device batch size of 16, gradient accumulation over eight steps (effective batch size 128), and a per-device evaluation batch size of eight. Optimization relied on AdamW with a cosine learning-rate scheduler, a warm-up ratio of 3%, weight decay of 0.01, and gradient clipping with a maximum norm of 1.0. Learning rate values were swept over $\{2 \times 10^{-5}, 3 \times 10^{-5}, 5 \times 10^{-5}, 6 \times 10^{-5}\}$. For reproducibility, we fixed the random seed to 42.

Due to GPU resource constraints, we adopted two different instruct fine-tuning strategies:

SmolLM3-3B We fine-tuned the HuggingFaceTB/SmolLM3-3B model using a custom template to ensure alignment with the training data. Training was conducted in bfloat16 precision with TF32 enabled and gradient checkpointing. No parameter-efficient fine-tuning or quantization was applied for this model.

Llama-8B-Instruct We fine-tuned meta-llama/Meta-Llama-3.1-8B-Instruct using parameter-efficient adaptation with LoRA (Hu

et al., 2021). LoRA was applied with rank $r = 16$, scaling factor $\alpha = 32$, and dropout 0.05, targeting the attention projection layers (q, k, v, o) and the feed-forward layers. No quantization was applied.

A.2 Machine Translation Template

Instruction + Source (English)	Reference Translation (Egyptian Arabic)
<i>Translate from English into Egyptian Arabic. Output only the translation.</i> I wonder if a table is available near the window for seven tonight.	أنا بسأل لو كان فيه ترابيزة جنب الشباك تكون فاضية على الساعة سبعة بليل.
<i>Translate from English into Egyptian Arabic. Output only the translation.</i> I feel chilly and my stomach hurts badly.	حاسس إني بردان ومعدتي واجعاني جامد.
<i>Translate from English into Egyptian Arabic. Output only the translation.</i> Can I invite you out for dinner some time?	ممكن أعزملك على العشا في وقت ما؟
<i>Translate from English into Egyptian Arabic. Output only the translation.</i> The very thing I was looking for.	الحاجة اللي كنت بدور عليها.
<i>Translate from English into Egyptian Arabic. Output only the translation.</i> Flight number two zero eight, to Tokyo.	رحلة رقم ميتين وثمانية، لطوكيو.
<i>Translate from English into Egyptian Arabic. Output only the translation.</i> I'd like a typewriter ribbon.	عاز شريط آلة كاتبة.
<i>Translate from English into Egyptian Arabic. Output only the translation.</i> What's the minimum age to do this?	إيه أقل سن ممكن يعمل ده؟
<i>Translate from English into Egyptian Arabic. Output only the translation.</i> Could you call me if you find my bag?	ممكن تكلمني لو لقيت شنطتي؟

Table 4: Examples of instruction-conditioned machine translation from English to Egyptian Arabic (MADAR-26).

A.3 Instruction Next-Token Generation

Instruction language	Instruction + Prefix	Reference continuation (assistant)
English	<i>Complete the sentence starting with these 3 words in Egyptian Arabic:</i> ده قدامك هناك،	دي الجملة كاملة باللهجة المصرية: ده قدامك هناك، يادوبك قدام مكتب استعلامات السياحة.
Dialectal (Egyptian Arabic)	كَمِّلْ الجملة وابدأ بالتلات كلمات دول باللهجة المصرية: ده قدامك هناك،	دي الجملة كاملة باللهجة المصرية: ده قدامك هناك، يادوبك قدام مكتب استعلامات السياحة.
English	<i>Complete the sentence starting with these 3 words in Egyptian Arabic:</i> عمري ما سمعت	دي الجملة كاملة باللهجة المصرية: عمري ما سمعت عن العنوان ده هنا.
Dialectal (Egyptian Arabic)	كَمِّلْ الجملة وابدأ بالتلات كلمات دول باللهجة المصرية: عمري ما سمعت	دي الجملة كاملة باللهجة المصرية: عمري ما سمعت عن العنوان ده هنا.

Table 5: Examples of instruction-conditioned dialectal generation from MADAR training data. The task consists of completing a sentence in a target dialect from a fixed three word prefix. Instructions are provided either in English or directly in the target dialect, while the generation objective remains identical.

A.4 Hyperparameter Search Results

In this section, we show the performance of each model under different decoding hyperparameter settings.

Hyperparameters		SmolLM3-3B		Llama-3.1-8B-Instruct		Command R Arabic		GPT-OSS-120B	
top- p	T	Diglossia	Fidelity	Diglossia	Fidelity	Diglossia	Fidelity	Diglossia	Fidelity
0.1	0.1	21.96	0.003	32.96	0.044	45.34	0.018	47.14	0.200
0.1	0.3	22.00	0.003	32.96	0.044	45.35	0.019	47.39	0.206
0.1	0.6	22.23	0.003	32.97	0.045	45.97	0.019	47.30	0.205
0.1	0.9	22.00	0.003	32.95	0.043	46.42	0.020	47.17	0.227
0.1	1	21.85	0.003	32.83	0.044	46.19	0.021	47.60	0.214
0.3	0.1	21.96	0.003	32.97	0.044	46.42	0.020	47.50	0.207
0.3	0.3	21.93	0.003	32.97	0.044	46.46	0.020	47.55	0.204
0.3	0.6	21.83	0.003	32.86	0.043	46.47	0.021	47.49	0.224
0.3	0.9	21.40	0.003	32.49	0.044	46.52	0.020	47.37	0.216
0.3	1	20.96	0.003	32.20	0.047	46.60	0.053	47.32	0.237
0.6	0.1	21.81	0.003	32.97	0.043	46.23	0.023	47.48	0.188
0.6	0.3	21.80	0.003	32.91	0.045	46.19	0.022	47.53	0.210
0.6	0.6	21.18	0.003	32.31	0.045	46.18	0.022	47.72	0.216
0.6	0.9	18.96	0.002	30.82	0.046	46.12	0.023	47.13	0.208
0.6	1	16.86	0.002	29.69	0.043	46.05	0.021	46.78	0.203
0.9	0.1	21.85	0.002	32.97	0.045	45.19	0.021	47.45	0.209
0.9	0.3	21.68	0.002	32.67	0.044	45.08	0.020	47.48	0.223
0.9	0.6	18.91	0.003	30.81	0.049	45.13	0.021	47.39	0.215
0.9	0.9	13.76	0.002	27.34	0.055	45.23	0.021	47.19	0.197
0.9	1	11.40	0.002	25.21	0.058	45.33	0.020	46.83	0.211
1	0.1	21.78	0.003	32.99	0.045	44.83	0.020	47.50	0.217
1	0.3	21.27	0.002	32.54	0.045	44.75	0.020	47.82	0.213
1	0.6	17.44	0.002	30.06	0.047	44.88	0.021	47.31	0.202
1	0.9	12.09	0.002	25.70	0.048	44.72	0.020	46.97	0.219
1	1	9.78	0.002	23.66	0.065	44.78	0.020	45.68	0.218

Table 6: Decoding performance for different top- p and temperature (T) settings, evaluated with Diglossia (ChrF++) and Fidelity (Macro ADI2).

The table 6 reports the impact of varying decoding hyperparameters (p and temperature) on both diglossia (ChrF++) and dialectal fidelity (Macro ADI2) across several models. Overall, performance remains relatively stable for moderate decoding settings, where diglossia scores vary only slightly within each model. However, increasing temperature leads to more diverse but less controlled generation, which often results in degraded diglossia scores, particularly for smaller models such as SmolLM3-3B, whose ChrF++ drops sharply at higher temperature values ($T = 1, top - p = 1, diglossia = 9.78$ diglossia score). OpenGPT-OSS-120B (the largest model) remain more robust, maintaining high diglossia and fidelity scores across most configurations.

Maastricht University at AMIYA: Adapting LLMs for Dialectal Arabic using Fine-tuning and MBR Decoding

Abdulhai Alali Abderrahmane Issam
Department of Advanced Computing Sciences
Maastricht University

{abdulhai.alali@student., abderrahmane.issam@maastrichtuniversity.nl}

Abstract

Large Language Models (LLMs) are becoming increasingly multilingual, supporting hundreds of languages, especially high resource ones. Unfortunately, Dialect variations are still underrepresented due to limited data and linguistic variation. In this work, we adapt a pre-trained LLM to improve dialectal performance. Specifically, we use Low Rank Adaptation (LoRA) fine-tuning on monolingual and English–Dialect parallel data, adapter merging and dialect-aware MBR decoding to improve dialectal fidelity generation and translation. Experiments on Syrian, Moroccan, and Saudi Arabic show that merging and MBR improve dialectal fidelity while preserving semantic accuracy. This combination provides a compact and effective framework for robust dialectal Arabic generation.

1 Introduction

Arabic dialects exhibit substantial variation in vocabulary, morphology, and syntax, making automated generation and translation challenging. Unlike Modern Standard Arabic (MSA), Dialectal Arabic (DA) is underrepresented in NLP resources, leading to difficulties in building models that produce fluent, semantically faithful, and dialectally authentic outputs (Alabdullah et al., 2025). The AMIYA Shared Task (Robinson et al., 2026) targets these challenges by evaluating LLMs on monolingual dialect generation and cross-lingual translation, emphasizing both dialect fidelity and instruction following.

To address these issues, we adapt a Large Language Model (LLM) using parameter-efficient fine-tuning on monolingual and English–Dialect parallel data. We train separate Low-Rank Adaptation (LoRA) adapters (Houlsby et al., 2019; Bapna and Firat, 2019; Hu et al., 2021) for each task (i.e. self supervised training on monolingual data and translation on parallel data), capturing dialectal surface

forms and semantic grounding, and combine them using TIES-Merging (Yadav et al., 2023). Additionally, we apply Minimum Bayes Risk (MBR) (Bickel and Doksum, 2007; Kumar and Byrne, 2004; Deguchi et al., 2024) decoding with dialect-aware scoring to select outputs that maximize dialect authenticity during generation.

Our experiments show that merging monolingual and translation-based adapters improves the balance between dialectal fidelity and semantic accuracy. MBR decoding further enhances dialectal authenticity, leading to consistent gains over single-source fine-tuning and standard decoding. While our approach is effective, the following limitations persist: the training data is relatively small, dialect identification metrics may not capture subtle or informal usage, and MBR decoding increases inference time.

2 AMIYA Shared Task

2.1 Task Description

The AMIYA Shared Task focuses on improving LLMs for Dialectal Arabic (DA), which remains significantly underrepresented compared to Modern Standard Arabic (MSA) (Bergman and Diab, 2022). Participants are asked to develop or adapt LLMs that can generate fluent, semantically faithful, and dialectally authentic Arabic across multiple regional varieties.

Systems are evaluated using the AL-QASIDA benchmark (Robinson et al., 2025), which measures dialectal fidelity, generation quality, and robustness to MSA–DA diglossia. Evaluation includes both monolingual dialect generation and cross-lingual settings, such as English–Dialect and MSA–Dialect translation. Performance is assessed using automatic metrics—primarily Arabic Dialect Identification And Dialectness (ADI2) (Robinson et al., 2025) for dialect fidelity and character-level F-score (chrF++) (Popović, 2015; Popović, 2015)

for translation quality—as well as human judgments of fluency and instruction adherence.

2.2 Datasets

We participated in the closed data track of the shared task focusing on 3 out of 5 Arabic dialects provided by the task, namely: Syrian, Moroccan and Saudi Arabic. For each dialect, we combine two types of training data: Monolingual dialectal text which consists of unstructured sentences in the target dialect, and Machine Translation (MT) data with English and DA parallel text.

Dialect (Type)	Dataset	# Samples
Syrian (Mono.)	Shami Corpus	25,136
Syrian (MT)	UFAL	120,600
Moroccan (Mono.)	DoDa	10,000
Moroccan (MT)	DoDa	10,000
Saudi (Mono.)	SDC	14,891
Saudi (MT)	SauDial	1,000

Table 1: Datasets used per dialect and supervision type. Shami Corpus (Abu Kwaik et al., 2018), UFAL (Krućinski et al., 2023), DoDa (Darija Open Dataset Contributors, 2023), SDC (TaghreedT/SDC Contributors, 2020), SauDial (Alanazi et al., 2025).

Table 1 summarizes the datasets and sample sizes used for fine-tuning across each dialect and supervision type. While more extensive data is available for most categories, we sub-sample the datasets to maintain computational efficiency and accelerate the experimental process.

2.3 Evaluation Metrics

Evaluation is performed using different metrics depending on the task setting:

Monolingual Dialect Evaluation. For monolingual generation, **ADI2** metric is used. ADI2 score was proposed in Robinson et al. (2025) to measure whether LLMs generate outputs that are dialectal, and whether they are faithful to the specific requested dialect. The level of dialectness is measured using Arabic Level of Dialectness of text (ALDI) (Keleg et al., 2023), and the dialect class C is predicted using a dialect identification baseline model from Nuanced Arabic Dialect Identification (NADI) 2024 shared task (Abdul-Mageed et al., 2024). More formally, ADI2 score (Robinson et al., 2025) is defined as:

$$\text{score}_{\text{ADI2}}(y) = \text{score}_{\text{ALDI}}(y) * \text{score}_{\text{NADI}}(y)_C \quad (1)$$

Cross-Lingual Evaluation. For translation-based and cross-lingual generation tasks, **chrF++** is used for evaluation. chrF++ is well suited for morphologically rich languages such as Arabic, where it captures fine-grained character overlap and is robust to spelling variation, making it appropriate for dialectal evaluation where orthographic inconsistency is common.

3 System Description

3.1 LoRA Fine-tuning

To incorporate dialectal knowledge into the base model, we use parameter-efficient fine-tuning with LoRA. Fine-tuning is performed separately for each dialect and task (i.e. self-supervised and translation), allowing the model to learn different types of information without intervention between them. Table 2 reports our training hyperparameters. These hyperparameters were chosen to ensure stable training under memory constraints while maintaining sufficient capacity for effective dialect adaptation.

Hyperparameter	Value
Max. sequence length	512
Epochs	5
Learning rate	3e-5
Batch size (per device)	2
Effective batch size	32
Precision	BF16

Table 2: Key training hyperparameters.

3.1.1 Monolingual Dialect Fine-tuning

For monolingual adaptation, we fine-tune the model on raw dialectal text without any task-specific prompts. The data consists of standalone sentences written in each dialect, encouraging the model to naturally learn dialect-specific vocabulary, morphology, and sentence structure.

All sentences are tokenized using the JAIS (Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), 2025) tokenizer with a fixed maximum sequence length. We train the model using a standard causal language modeling objective. To make fine-tuning efficient, we apply LoRA adapters (Bapna and Firat, 2019; Houlby et al., 2019; Hu et al., 2021) to the attention layers of the model and update only these additional parameters during training. Furthermore, we rely on memory-optimization techniques such as gradient accumulation and gradient checkpointing, enabling larger effective batch sizes. The model is trained for multiple epochs using a standard optimization setup.

This approach allows the model to adapt strongly to dialectal surface forms while preserving the general knowledge of the base model.

3.1.2 Translation-Based Fine-tuning

In addition to monolingual data, we fine-tune the model on an English–Dialect parallel dataset. This data exposes the model to aligned semantic content across languages, helping it associate dialectal expressions with their meanings and improving controllability during generation.

We frame translation as an instruction-following task in both directions: English→Dialect and Dialect→English. Each training example includes a natural language instruction specifying the translation direction, followed by the target output. During training, the loss is computed only on the output tokens, while the instruction tokens are masked. This encourages the model to follow instructions without learning to reproduce them.

Tokenization is performed with a fixed maximum sequence length. The same LoRA setup is used as in monolingual fine-tuning to ensure compatibility across training stages. Training is carried out with a more conservative optimization setup than monolingual fine-tuning, focusing on stable learning and semantic alignment rather than aggressive adaptation.

3.1.3 Adapter Merging

The monolingual and translation-based fine-tuning strategies provide complementary supervision. Monolingual fine-tuning emphasizes dialectal fluency and authenticity, while translation fine-tuning reinforces semantic faithfulness and cross-lingual grounding. By training separate LoRA adapters for each dataset, we preserve these distinct signals and later combine them using TIES-Merging. This separation enables fine-grained control over how different sources of supervision contribute to the final dialect-aware model and minimizes intervention between them.

3.2 MBR Decoding with Dialect-Aware Scoring

While fine-tuning and merging improve the model’s internal dialect representations, decoding decisions still play a crucial role in output quality. We therefore apply MBR decoding using the *mbrs*¹ library to explicitly optimize for dialectness at inference time.

¹<https://github.com/naist-nlp/mbrs>

For each input prompt, the model generates a set of $N = 20$ candidate responses via stochastic sampling, then each candidate is scored independently using the ADI2 metric. Finally, the candidate with the highest score is selected as the final output.

3.3 Adapter Merging and MBR

TIES-based adapter merging integrates complementary dataset supervision at the parameter level, producing a compact yet expressive dialect-aware model. MBR decoding complements this by enforcing dialectal fidelity at generation time, explicitly selecting outputs that maximize dialectness. Together, fine-tuning, TIES-Merging, and MBR decoding form a unified framework that yields more consistent and authentic dialectal generation than any single technique in isolation.

4 Results

The experiments were conducted exclusively on Syrian and Moroccan DA and subsequently applied to the remaining dialects for the final submission, which was trained separately per dialect. This section presents a detailed evaluation of our dialect-aware generation framework. We report results across model variants, data configurations, and decoding strategies, with the goal of understanding (1) the impact of model choice, (2) the role of different supervision signals, and (3) the effectiveness of adapter merging and MBR decoding. The evaluation datasets used are the default datasets provided by AL-QASIDA ².

4.1 JAIS-2 vs. LLaMA 3.2

We begin by comparing two LLMs, JAIS-2³ (Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), 2025) and LLaMA 3.2⁴ (Grattafiori et al., 2024), to determine the most suitable backbone for Arabic dialect generation. For a fair comparison, both models are fine-tuned using the same data configuration: a merged setup that combines monolingual dialect data with English–Dialect parallel (MT) supervision. In addition, decoding is performed using MBR decoding with ADI2 score.

Models are evaluated using ADI2 for monolingual dialect generation and chrF++ for trans-

²<https://github.com/JHU-CLSP/al-qasida>

³<https://huggingface.co/inceptionai/Jais-2-8B-Chat>

⁴<https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct>

lation. Table 3 presents the results. On monolingual dialect generation, LLaMA 3.2 achieves a substantially higher ADI2 score (i.e. 0.78), indicating strong dialectal surface realization and fluency. However, its performance drops sharply in translation, with a significantly low chrF++ score (i.e. 0.14), suggesting weak semantic alignment when translating from English into dialectal Arabic. In contrast, JAIS-2 exhibits a more balanced performance. While its ADI2 score (0.33) is considerably lower than that of LLaMA 3.2 for monolingual generation, JAIS-2 achieves a much higher chrF++ score (0.43) on MT-based generation. This indicates stronger semantic fidelity and better handling of translation supervision.

Model	ADI2	chrF++
LLaMA 3.2	0.78	0.14
JAIS-2	0.33	0.43

Table 3: Comparison between JAIS-2 and LLaMA 3.2 after fine-tuning, TIES-Merging and generation with MBR decoding on **Syrian DA**

On overall, although LLaMA 3.2 excels in dialectal surface form generation, its poor cross-lingual performance limits its usefulness for translation-driven dialect generation. Given our goal of building a dialect-aware system that remains reliable across both monolingual and cross-lingual scenarios, we select JAIS-2 as the backbone for all subsequent experiments.

4.2 Effect of Fine-tuning Data and Adapter Merging

In this section, we analyze the impact of different fine-tuning strategies on JAIS-2. We report the results of JAIS-2 base model, JAIS-2 fine-tuned on either monolingual or translation data, and JAIS-2 with TIES-Merging. The results in Table 4 show that monolingual fine-tuning substantially improves ADI2 scores, indicating stronger dialectal identity and linguistic conformity. Furthermore, fine-tuning on parallel translations significantly improves chrF++ scores, reflecting improved semantic faithfulness and cross-lingual grounding. More importantly, merging monolingual and MT models using TIES-Merging consistently improves the balance between dialectal authenticity and semantic accuracy, leading to the best chrF++ and the second best ADI2 score performance.

Configuration	ADI2	chrF++
JAIS-2 (Base)	0.18	0.31
+ Monolingual FT	0.44	0.33
+ MT FT	0.26	0.42
+ TIES-Merging	0.38	0.44

Table 4: Effect of monolingual and MT task fine-tuning and TIES-Merging on **Moroccan DA**. Merging both tasks (i.e. TIES-Merging) leads to the best performance on chrF++ and the second best on ADI2 score.

4.3 MBR Decoding with Dialect-Aware Objectives

While adapter merging improves the model’s internal representations, standard decoding does not always select the most dialectally appropriate output. To address this, we apply MBR decoding with different objectives. MBR requires a target metric to score the candidate generations. We experiment with using ADI2 to improve dialectal fidelity, chrF++ to improve cross-lingual grounding, and their combination. As shown in 5, MBR decoding with ADI2 achieves the best overall balance, improving monolingual ADI2 to approximately 0.51 while also increasing MT ADI2 to 0.36. This represents a substantial improvement over standard decoding and demonstrates that dialect-aware reranking can recover dialectal authenticity without sacrificing semantic grounding. In contrast, chrF++-optimized MBR and the combined objective favor translation quality: they achieve higher chrF++ scores (0.42 and 0.41, respectively) but lead to a significant drop in monolingual ADI2 (0.24 and 0.29). These results indicate that chrF++-centric objectives bias the model toward more neutral or standardized Arabic forms, reducing dialectal distinctiveness. Based on these findings, and given our emphasis on dialect fidelity while maintaining acceptable translation performance, we select ADI2-based MBR decoding for the final submission.

Decoding Strategy	Mono ADI2	MT ADI2	chrF++
Standard decoding	0.38	0.27	0.44
MBR (ADI2)	0.51	0.36	0.40
MBR (chrF++)	0.24	0.30	0.42
MBR (ADI2 + chrF++)	0.29	0.37	0.41

Table 5: Effect of MBR decoding objectives on JAIS-2 after fine-tuning and TIES-Merging on **Moroccan DA**. Using ADI2 as an objective strikes the best balance in ADI2 and chrF++ performance.

4.4 Final Submission

Across all experiments, the best-performing configuration is: JAIS-2 independent fine-tuning on monolingual and translation data, merging with TIES technique, and decoding using MBR with ADI2 as an objective metric. This configuration achieves the strongest balance between dialectal authenticity and cross-lingual grounding, outperforming all alternatives in terms of combined monolingual and MT performance. Consequently, our *primary* submission applies this methodology individually to the Moroccan, Syrian, and Saudi dialects. While joint training across all dialects may offer further gains, we defer this exploration to future work.

5 AMIYA Shared Task Official Results

Table 6 summarizes the official automatic evaluation results of our *primary* submission. The results demonstrate strong generalization across dialects and task settings, confirming that the combination of parameter-efficient fine-tuning, TIES-Merging, and dialect-aware MBR decoding provides a robust and effective solution for DA generation.

Dialect	ADI2	DA→EN	EN→DA	DA→MSA	MSA→DA
Moroccan	0.679	49.93	30.02	39.53	33.77
Syrian	0.389	51.89	34.44	43.42	40.33
Saudi	0.464	0.03	19.82	37.21	24.23

Table 6: Automatic evaluation results using ADI2 and chrF++.

Besides automatic evaluation, our model generations were human evaluated for fluency and adherence to DA instructions. Table 7 shows that our highest human evaluation performance is on the Moroccan dialect.

Dialect	Adequacy	Fluency
Moroccan	1.97	3.37
Syrian	1.146	2.625
Saudi	1.122	2.378

Table 7: Human evaluation scores.

Across submissions from other teams, our system achieved the highest ADI2 for Syrian and Saudi, and the highest chrF++ scores on translation from English and MSA into Syrian (ENG→DA and MSA→DA). On Moroccan Arabic, our model performs best on the human evaluation of fluency.

6 Conclusion

We presented a method for improving Dialectal Arabic generation by combining fine-tuning on dialectal and translation data, LoRA adapter merging, and MBR decoding. This approach helps the model produce outputs that are both fluent in the target dialect and faithful to the input meaning. Our experiments across three dialects show that this combination works better than using any single technique on its own, providing a practical way to make LLMs more dialect-aware.

Limitations

This work has the following limitations: Our decoding strategy depends on automatic dialect identification (ADI2) scores, which may not always capture subtle or informal dialectal usage. The training data is limited in size and may not cover all linguistic variation within each dialect, especially code-switching and colloquial expressions. Finally, both ADI2 computation and minimum Bayes risk (MBR) decoding increase inference time. ADI2 requires an additional forward pass, and the cost of MBR grows with the number of candidate hypotheses generated per input, resulting in a several-fold slowdown compared to standard decoding. As a result, the approach may be less practical for real-time or low-latency applications.

References

- Muhammad Abdul-Mageed, Amr Keleg, AbdelRahim Elmadany, Chiyu Zhang, Injy Hamed, Walid Magdy, Houda Bouamor, and Nizar Habash. 2024. [NADI 2024: The fifth nuanced arabic dialect identification shared task](#). In *Proceedings of The Second Arabic Natural Language Processing Conference (Arabic-NLP 2024)*, pages 709–728, Bangkok, Thailand. Association for Computational Linguistics.
- Kathrein Abu Kwaik, Motaz Saad, Stergios Chatzikyriakidis, and Simon Dobnik. 2018. [Shami: A corpus of levantine arabic dialects](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA). Dataset includes Levantine Arabic tweets covering Palestinian, Jordanian, Syrian, and Lebanese dialects.
- Abdullah Alabdullah, Lifeng Han, and Chenghua Lin. 2025. Advancing dialectal arabic to modern standard arabic machine translation. *arXiv preprint arXiv:2507.20301*. Dialectal Arabic-MSA MT challenges and resource-efficient strategies.

- Naif Alanazi, Mohammed Al-Batineh, and Hussein Abu-Rayyash. 2025. **Saudial: Saudi arabic dialects game localization dataset**. <https://data.mendeley.com/datasets/mzdwkb2t6d/2>. Parallel Saudi dialect text dataset (English, MSA, Saudi varieties) with cultural context, age ratings, and dialect notes. CC-BY 4.0 license.
- Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. 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 2414–2425.
- A. Bergman and Mona Diab. 2022. **Towards responsible natural language annotation for the varieties of Arabic**. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 364–371, Dublin, Ireland. Association for Computational Linguistics.
- Peter Bickel and Kjell Doksum. 2007. *Mathematical Statistics: Basic Ideas and Selected Topics.*, volume 56.
- Darija Open Dataset Contributors. 2023. Doda: Darija open dataset. <https://github.com/darija-open-dataset/dataset/tree/main?tab=readme-ov-file#citation>. Dataset of 48,849 Darija sentences, with English translations. CC-BY-NC license. Downloadable via link. High quality.
- Hiroyuki Deguchi, Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2024. **mbrs: A library for minimum Bayes risk decoding**. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 351–362, Miami, Florida, USA. Association for Computational Linguistics.
- 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, and 542 others. 2024. **The llama 3 herd of models**. *Preprint*, arXiv:2407.21783.
- 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**. *Preprint*, arXiv:1902.00751.
- 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**. <https://arxiv.org/abs/2106.09685>. Presented at ICLR 2022; parameter-efficient adaptation method for large pre-trained models.
- Amr Keleg, Sharon Goldwater, and Walid Magdy. 2023. **ALDi: Quantifying the Arabic level of dialectness of text**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10597–10611, Singapore. Association for Computational Linguistics.
- Mateusz Krubiński, Hashem Sellat, Shadi Saleh, Adam Pospíšil, Petr Zemánek, and Pavel Pecina. 2023. **Multi-parallel corpus of north levantine arabic**. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, pages 411–417. Association for Computational Linguistics. Corpus includes 120,600 multiparallel sentences in English, French, German, Greek, Spanish and MSA manually translated into North Levantine Arabic.
- Shankar Kumar and William Byrne. 2004. **Minimum Bayes-risk decoding for statistical machine translation**. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA. Association for Computational Linguistics.
- Mohamed bin Zayed University of Artificial Intelligence (MBZUAI). 2025. Inception, cerebras and mbzuai release Jais 2 – the next generation arabic open-weight llm. <https://mbzuai.ac.ae>. Accessed: 2026-01-21.
- Maja Popović. 2015. **chrf: character n-gram f-score for automatic MT evaluation**. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.
- Maja Popović. 2015. **chrf: Character n-gram f-score for automatic mt evaluation**. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.
- Nathaniel Robinson, Shahd Abdelmoneim, Anjali Kantharuban, Otba Alsoul, Salima Lamsiyah, Kelly Marchisio, and Kenton Murray. 2026. **AMIYA shared task: Arabic Modeling In Your Accent at VarDial 2026**. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, Rabat, Morocco. Association for Computational Linguistics.
- Nathaniel R. Robinson, Shahd Abdelmoneim, Kelly Marchisio, and Sebastian Ruder. 2025. **Al-qasida: Analyzing llm quality and accuracy systematically in dialectal arabic**. *Preprint*, arXiv:2412.04193.
- TaghreedT/SDC Contributors. 2020. **Saudi dialect corpus (sdc)**. <https://github.com/TaghreedT/SDC/blob/main/SDC.txt>. A 210,396-word corpus of Saudi Arabic social media posts collected from platforms such as Facebook and Twitter. High quality text format, accessible via link; please cite Tarmom et al. (2020) when using this corpus.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. 2023. [TIES-merging: Resolving interference when merging models](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.

SDNLP at AMIYA 2026: Syrian Arabic Dialect Modeling with LoRA

Hasan Alkhder

Sakarya University, Türkiye

hasan.alkhder2@ogr.sakarya.edu.tr

Mohammad Abboush

TU Clausthal, Germany

mohammad.abboush@tu-clausthal.de

Abstract

Dialectal Arabic continues to represent a persistent challenge for contemporary large language models, which are predominantly trained and optimized for Modern Standard Arabic (MSA) and therefore exhibit limited capability when processing colloquial varieties. In this study, a dedicated system developed for participation in the AMIYA shared task focusing on Syrian Arabic is presented. The proposed solution is based on the integration of parameter-efficient fine-tuning through Low-Rank Adaptation (LoRA) with prompt-guided inference, aiming to enhance dialectal adequacy and linguistic naturalness. Rather than emphasizing strict factual precision, the system is deliberately designed to prioritize fluent and authentic Syrian Arabic generation, in accordance with the evaluation principles adopted by the AL-QASIDA benchmark. This design choice reflects a focus on human-perceived language quality and dialectal fidelity, which are central to effective dialect-aware language modeling.

1 Introduction

Dialectal Arabic constitutes the dominant form of daily communication throughout the Arab world; however, it remains largely underexplored within natural language processing research, which has traditionally focused on MSA. As a result, most pretrained large language models exhibit strong performance in formal registers while demonstrating limited capability in generating fluent and natural dialectal language, a shortcoming that is particularly pronounced for Syrian Arabic due to its highly conversational nature and substantial divergence from standardized written forms. This work is conducted within the context of the AMIYA shared task on Syrian Arabic dialect modeling (Robinson et al., 2026).

To address these challenges, the proposed system is guided by a set of explicit design choices

tailored to dialect modeling under shared-task constraints. Specifically, we adopt: (i) parameter-efficient fine-tuning via Low-Rank Adaptation (LoRA) to enable dialect specialization without modifying the full model parameters; (ii) prompt-level constraints that explicitly encourage Syrian Arabic realization while discouraging Modern Standard Arabic structures; and (iii) a qualitative evaluation strategy aligned with the AL-QASIDA benchmark (Robinson et al., 2025), prioritizing dialectal fluency and naturalness over strict literal fidelity. Together, these choices frame the task as one of dialect modeling rather than factual question answering, and directly inform the methodological decisions described in the following sections.

From a linguistic perspective, the challenge of dialectal Arabic modeling is further amplified by the phenomenon of diglossia, where speakers naturally alternate between MSA and colloquial varieties depending on context. This linguistic dynamic often leads pretrained models to default to MSA even when prompted for dialectal output, resulting in responses that sound unnatural to native speakers. Addressing this mismatch requires not only suitable training data but also modeling strategies that explicitly encourage colloquial realization and discourage formal constructions. These considerations motivate the design choices adopted in the present system.

2 Task Framing

The AMIYA shared task is approached in this work as a dialect modeling problem rather than a conventional question answering or fact retrieval task. The primary objective is to generate responses that resemble naturally spoken Syrian Arabic as used in everyday interaction, where fluency and dialectal authenticity are prioritized over strict literal faithfulness to the input prompt.

Importantly, deprioritizing strict semantic cor-

rectness does not imply unconstrained or nonsensical generation. The system remains grounded in the semantic intent of the input prompt, which serves as an anchor for generation. Rather than enforcing exact word-by-word or structure-preserving fidelity, the model is allowed to paraphrase, generalize, or abstract the input content when doing so results in more natural colloquial expression. In practice, this means that semantic adequacy is preserved, while rigid literal alignment is relaxed.

Nonsensical or hallucinated outputs are mitigated through a combination of instruction-style fine-tuning and constrained inference. During training, the model is exposed to realistic conversational data grounded in coherent dialogue contexts, which implicitly reinforces semantic plausibility. At inference time, prompts explicitly specify both the target dialect (Syrian Arabic) and the expected response format, anchoring generation to the input intent while allowing stylistic flexibility. This combination prevents unconstrained generation while still enabling dialectally natural reformulation.

This task formulation aligns with the evaluation principles of the AL-QASIDA benchmark, which penalizes outputs that drift toward Modern Standard Arabic or overly formal registers, even when they are semantically precise. Consequently, the methodological focus shifts toward controlling stylistic realization rather than enforcing exact semantic reproduction. This trade-off reflects real-world conversational usage, where communicative plausibility and dialectal naturalness often outweigh strict literal accuracy.

3 Data

3.1 Data Source

The training data used in this work consists of professionally produced dubbed television dialogues provided by the professional dubbing company NIS. The dataset is external to the official AMIYA shared task release and is therefore used under the Open Track regulations. No data provided by the task organizers or other publicly released shared-task datasets (e.g., UFAL or related corpora) were used in this work.

The original scripts are primarily in English and were translated into Modern Standard Arabic (MSA) and subsequently localized into Syrian Arabic by professional translators as part of the dubbing production workflow. Crucially, the Syrian Arabic content reflects natural spoken usage rather

than literal translation, as it is adapted for oral delivery in audiovisual media. This makes the data particularly suitable for dialect modeling tasks that emphasize conversational naturalness over formal written structure.

This localization process yields conversational, informal utterances that closely resemble everyday spoken Syrian Arabic. As the data is generated by professional dubbing practitioners, it exhibits consistent stylistic patterns and strong dialectal authenticity, distinguishing it from automatically generated or crowdsourced resources. In addition, the dialogue-oriented nature of the data supports parameter-efficient adaptation methods such as Low-Rank Adaptation (LoRA) (Hu et al., 2022), which benefit from stylistically coherent training signals.

3.2 Dataset Size and Structure

The dataset comprises approximately **30,000 dialogue utterances**. Each instance corresponds to a single conversational turn and is represented in a structured tabular format with the following fields:

- **character**: the speaker identifier associated with the dialogue turn,
- **gender**: the speaker’s gender (male or female),
- **english**: the original English script,
- **arabic_msa**: the corresponding Modern Standard Arabic version,
- **arabic_syrian**: the target Syrian Arabic utterance.

The average utterance length is approximately **9.6 words** in Syrian Arabic, **8.8 words** in MSA, and **9.9 words** in English, reflecting the dialogue-oriented and concise nature of the data. No additional datasets or auxiliary corpora were incorporated during training.

3.3 Suitability for Dialect Modeling

The dataset is inherently dialogue-driven and predominantly informal, reflecting everyday conversational scenarios rather than scripted narration or formal text. This characteristic aligns closely with the objectives of the AMIYA shared task, which prioritizes dialectal fidelity, fluency, and naturalness over strict semantic equivalence.

Moreover, the presence of parallel English and MSA representations implicitly exposes the model to distinctions between formal and colloquial registers. This property is particularly valuable in the context of Arabic diglossia, where pretrained language models often default to MSA even when prompted for dialectal output. By grounding adaptation in professionally localized dialectal utterances, the model is encouraged to internalize colloquial realization patterns that are difficult to capture using automatically generated resources.

3.4 Legal and Ethical Considerations

The data was used with explicit permission for research purposes under an agreement with the data provider and fully complies with the regulations of the AMIYA Open Track. No restricted, confidential, or off-limits resources were incorporated in the development of the proposed system, and no personal or sensitive user information is contained within the dataset.

4 System Overview

The proposed system follows a modular adaptation-based architecture designed specifically for dialect modeling under shared-task constraints. It is built upon a pretrained large language model, which is extended through the integration of a lightweight Low-Rank Adaptation (LoRA) module and a controlled prompt-guided inference strategy. In this configuration, all original parameters of the base model remain frozen, while a small set of additional trainable parameters is optimized to steer the model toward Syrian Arabic generation.

Dialectal specialization is achieved through two complementary mechanisms. First, LoRA-based parameter-efficient fine-tuning allows the model to internalize stylistic and lexical patterns characteristic of Syrian Arabic without altering its general linguistic competence. Second, prompt-guided inference is employed at generation time to explicitly control output structure and register, encouraging concise, single-sentence responses in colloquial Syrian Arabic. Together, these mechanisms enable effective dialect adaptation while maintaining model stability and reproducibility.

The architecture is intentionally designed to be modular. By decoupling the pretrained backbone from the dialect-specific adaptation layer, the system can be easily extended to other Arabic dialects by replacing or retraining only the LoRA compo-

nents, without requiring full model retraining. This modularity also facilitates experimentation with alternative prompting strategies or constraint formulations, making the approach well suited for iterative refinement in shared-task settings.

An additional motivation for adopting LoRA in this context is its alignment with the computational and practical constraints of shared tasks such as AMIYA. Parameter-efficient adaptation enables targeted dialect specialization with limited resources, avoiding the need for large-scale retraining while preserving the robustness and generalization capabilities of the underlying pretrained model. As a result, the system emphasizes controllable stylistic adaptation rather than architectural complexity.

This system is submitted as the *primary* submission for the Syrian Arabic track of the AMIYA shared task.

5 Model and Training

The proposed system is built upon the **Meta-Llama-3.1-8B** *base* language model (i.e., not the instruction-tuned variant) (Touvron et al., 2023), accessed via the HuggingFace platform.¹ Parameter-efficient fine-tuning is performed using Low-Rank Adaptation (LoRA) (Hu et al., 2022), while keeping all original model parameters frozen.

LoRA adapters are injected into the attention layers of the transformer architecture, specifically targeting the `q_proj`, `k_proj`, `v_proj`, and `o_proj` modules. The adaptation is configured with a rank of **16**, a scaling factor of **32**, and a dropout rate of **0.05**. This configuration enables effective dialect specialization while maintaining training stability and computational efficiency.

Fine-tuning is conducted in a causal language modeling setting using instruction-style input-output pairs derived from the dialogue data. The model is trained for **1.5 epochs** using the AdamW optimizer with a learning rate of 2×10^{-4} . A per-device batch size of **4** is used in conjunction with gradient accumulation over **8** steps, resulting in an effective batch size of **32**. To reduce memory consumption, the model is loaded in 8-bit precision and trained using mixed-precision (FP16). All training procedures strictly comply with the AMIYA Open Track regulations.

¹<https://huggingface.co/meta-llama/Meta-Llama-3.1-8B>

Submission	Adequacy	Fluency
SDNLP (Primary)	1.124	2.928

Table 1: Official AMIYA 2026 human evaluation results for the SDNLP primary submission on the Syrian Arabic track.

6 Prompting and Inference

During inference, a prompt-guided generation strategy is employed to steer the model toward producing dialectally appropriate outputs. The evaluation prompts are those officially released as part of the AMIYA shared task and distributed by the task organizers under the AL-QASIDA evaluation framework. These prompts typically include explicit instructions such as “Translation:” or “Answer in Syrian Arabic;” and are used in accordance with their original attribution and evaluation guidelines.

To enforce dialectal and structural consistency, the original prompts are augmented at the prompt level with explicit constraints that require the generation of a single sentence in Syrian Arabic while discouraging explanations, meta-commentary, or multi-sentence outputs. These constraints are implemented through prompt formulation rather than decoding-time penalties, ensuring that generation remains anchored to the semantic intent of the input while allowing stylistic flexibility in colloquial realization.

In addition to prompt-level control, a lightweight post-processing step is applied to remove spurious surface-level artifacts. This step consists of simple rule-based filtering to eliminate duplicated tokens, unintended Latin characters, and residual Modern Standard Arabic morphological endings that occasionally appear in generated outputs. No content-level rewriting or semantic modification is performed during post-processing; the procedure is limited to surface-form cleanup to preserve the model’s original generation behavior.

The restriction to single-sentence outputs is motivated by empirical observation rather than a formal ablation study. In preliminary experimentation, unconstrained prompts frequently resulted in longer, explanatory responses that diluted dialectal consistency. By contrast, explicitly constrained prompts yielded outputs that were shorter, more predictable, and more consistently aligned with colloquial Syrian Arabic. This qualitative observation is reflected in the example outputs presented in Section 3.

7 Evaluation

System predictions are generated for all test prompts officially released as part of the AMIYA shared task and provided by the task organizers under the AL-QASIDA evaluation framework. The prompts are processed in their original order without re-ranking or filtering, and system outputs are exported in a comma-separated values (CSV) format that preserves both the original sequence and the exact number of entries, in accordance with the shared task submission guidelines.

All evaluation is conducted centrally by the AMIYA organizers. The evaluation framework combines automatic metrics designed to assess dialectal characteristics with human judgments that focus on perceived fluency and adequacy. No additional test-time adaptation, prompt modification, or post-hoc tuning is applied beyond the inference constraints described in Section 6.

Table 1 presents the human evaluation scores provided by the task organizers. Adequacy measures how well the generated responses align with the intended meaning of the prompt, while Fluency reflects the perceived naturalness and conversational quality of the output in Syrian Arabic. These human judgments play a central role in the AMIYA shared task, where dialectal naturalness and spoken plausibility are prioritized.

Table 2 reports the official automatic evaluation scores. The ADI2 metric reflects overall dialectal performance, while the direction-specific chrF-based scores measure cross-lingual and cross-variety consistency between Dialectal Arabic (DA), English (ENG), and Modern Standard Arabic (MSA). These metrics are intended to capture stylistic and distributional alignment rather than strict semantic equivalence.

Taken together, the automatic and human evaluation results indicate that the proposed system achieves competitive performance within the Syrian Arabic track. While automatic metrics provide supporting evidence of dialectal alignment, the evaluation framework places particular emphasis on human judgments, which better reflect colloquial fluency and dialectal authenticity. The reported scores are consistent with the system’s design goal of prioritizing natural spoken Syrian Arabic over strict semantic optimization.

Submission	ADI2	DA→ENG	ENG→DA	DA→MSA	MSA→DA
SDNLP (Primary)	0.280	36.211	14.813	6.743	15.609

Table 2: Official AMIYA 2026 automatic evaluation results for the SDNLP primary submission on the Syrian Arabic track. Directional scores are reported using chrF/chrF++ metrics as defined in the AL-QASIDA evaluation framework.

Source (English)	Base Model Output	LoRA-Adapted Output
Where are you going now?	يلا قومي على رجلكي وين بدك تروحي لا تروحي من هون	انتي شو عم تعملي هون وين رايحة؟
I don't want to talk anymore.	انا ما بدني احكي اكثر من هيك خليكي فيقي	بدي روح من هون ما عاد بدني حكي معك
What should we do now?	شو بدنا نعمل؟ كيف ح نعيش هيك؟	شو بدنا نعمل هلاً؟
It's too late now.	انت ما بتسوي شي لاليف حالة رابعة	ما عاد في داعي نضيع وقت هيك

Table 3: Qualitative comparison of output fluency between the base Llama-3.1-8B model and the LoRA-adapted version.

8 Observations and Example Outputs

To assess the impact of dialect-specific adaptation, we conduct a qualitative comparison between the outputs of the LoRA-adapted model and those produced by the original Meta-Llama-3.1-8B base model under identical prompting conditions. This comparison is designed to isolate the effect of parameter-efficient fine-tuning on stylistic and dialectal generation behavior, independent of task formulation or prompt variation.

The evaluation prompts follow a controlled instruction-based format. Each input explicitly requests generation in Syrian Arabic using formulations such as:

```
Respond in Syrian Arabic:
<sentence>
Rewrite in Syrian Arabic:
<sentence>
```

For clarity and readability, Table 3 reports only the core semantic intent of each prompt, while the full instruction templates are described here and were applied consistently during evaluation.

Table 3 presents four representative examples selected from a broader set of prompts used for qualitative analysis. The examples correspond to common conversational scenarios, including refusal, hesitation, decision-making, and expressions

of temporal finality. They are not intended to be exhaustive, but rather to illustrate systematic stylistic differences in dialectal realization between the base and adapted models.

Across these examples, the base model frequently exhibits one or more of the following behaviors: (i) introduction of additional narrative or explanatory content not implied by the prompt, (ii) partial drift toward less context-appropriate or more verbose responses, and (iii) overextended utterances that exceed what is typical in spontaneous spoken Syrian Arabic. For instance, in the prompt “*Where are you going now?*”, the base model produces a longer utterance that introduces additional imperatives and discourse elements beyond the pragmatic intent of the input.

In contrast, the LoRA-adapted model consistently produces shorter, more direct utterances that better reflect colloquial Syrian Arabic norms. For example, in “*What should we do now?*”, the adapted model favors a concise spoken formulation that aligns closely with everyday conversational usage, while the base model expands the response with speculative or emotionally loaded content. Across the remaining examples, similar tendencies toward brevity, pragmatic compression, and reduced overgeneration can be observed.

These qualitative differences illustrate a con-

trolled form of generalization rather than semantic degradation. While certain outputs sacrifice literal completeness, they remain contextually appropriate and fluent within a spoken dialogue setting. Importantly, these patterns are consistent with the system’s observed behavior under the AL-QASIDA evaluation framework, which emphasizes dialectal fidelity, fluency, and naturalness over strict factual accuracy. The examples therefore serve to concretely illustrate how the proposed system satisfies the primary evaluation criteria of the AMIYA shared task.

9 Conclusion

This work introduces a Low-Rank Adaptation (LoRA)-based system for modeling the Syrian Arabic dialect, developed and submitted within the context of the AMIYA shared task. By conceptualizing the problem as a dialect modeling task and integrating parameter-efficient fine-tuning with prompt-guided inference, the proposed approach attains a high level of dialectal fluency in accordance with the AL-QASIDA evaluation framework. The presented results underscore the suitability and effectiveness of lightweight adaptation strategies for advancing dialectal Arabic language modeling while maintaining computational efficiency.

Limitations

The proposed system is not designed with the primary objective of maximizing factual accuracy or delivering detailed technical explanations. Consequently, when confronted with prompts involving precise dates, named entities, or complex informational content, the generated responses may exhibit a degree of abstraction or generalization. This behavior is an intentional outcome of the adopted design strategy, which prioritizes dialectal authenticity and conversational naturalness over exhaustive semantic completeness.

Concrete examples of this abstraction can be observed in Table 3, where the LoRA-adapted model produces shorter, pragmatically appropriate responses that condense or paraphrase the input prompt rather than reproducing its full literal content. While this may result in reduced semantic specificity, the outputs remain fluent, contextually appropriate, and aligned with the evaluation objectives of the AMIYA shared task.

Acknowledgments

The authors would like to express their gratitude to Mr. Nabeel Al-Dakkak from NIS for granting access to the professional dubbing dialogue data used in this study.

References

- Edward J. Hu and 1 others. 2022. Lora: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations*.
- Nathaniel Robinson, Shahd Abdelmoneim, Anjali Kantharuban, Otba Alsboul, Salima Lamsiyah, Kelly Marchisio, and Kenton Murray. 2026. AMIYA shared task: Arabic Modeling In Your Accent at VarDial 2026. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, Rabat, Morocco. Association for Computational Linguistics.
- Nathaniel R. Robinson, Shahd Abdelmoneim, Anjali Kantharuban, Otba Alsboul, Salima Lamsiyah, Kelly Marchisio, and Kenton Murray. 2025. Al-qasida: Evaluating dialectal arabic in large language models. In *Proceedings of the AL-QASIDA Shared Task*. Association for Computational Linguistics.
- Hugo Touvron and 1 others. 2023. Llama: Open and efficient foundation language models. In *Proceedings of the International Conference on Machine Learning*.

NUS-IDS at AMIYA/VarDial 2026: Improving Arabic Dialectness in LLMs with Reinforcement Learning

Sujatha Das Gollapalli,¹ Mouad Hakam,¹ Mingzhe Du,^{1,2} See-Kiong Ng¹

¹Institute of Data Science, National University of Singapore

²College of Computing and Data Science, Nanyang Technological University
{idssdg,mouad.hk,mingzhe,seekiong}@nus.edu.sg

Abstract

In this paper, we describe models developed by our team, NUS-IDS, for the *Closed Data* track at the Arabic Modeling In Your Accent (AMIYA) shared task at VarDial 2026. The core idea behind our solution involves data augmentation enabled by a dialect classifier trained on AMIYA data. We effectively combine various translation, summarization, and question answering prompts with the training data to form dialectal prompts for use with state-of-the-art Large Language Models (LLMs). Next, dialect predictions on outputs from these LLMs are used to compile preference data for Reinforcement Learning (RL). We report model performance on dialectal Arabic from Egypt, Morocco, Palestine, Saudi Arabia, and Syria using FLORES+, a benchmark dataset for multilingual machine translation. Our experiments illustrate that though our RL models show significant performance gains on dialectness scores, they underperform on translation metrics compared to base LLMs.

1 Introduction

Recent AI innovation in form of Large language models (LLMs) has shown revolutionary potential for solving various data processing and language understanding tasks. Indeed, state-of-the-art (SOTA) proprietary LLMs from OpenAI, Google, etc. demonstrate competitive performance on various linguistic tasks including translation, summarization, question answering, and sentiment analysis even in zero-shot settings using simple prompts (Vatsal and Dubey, 2024). However, ongoing multilingual research studies note that SOTA LLMs are not uniformly proficient in all languages and their error rates are often significantly higher when handling instructions in non-Latin scripts such as Arabic, Russian, and Hindi (Hasan et al., 2024; He et al., 2024). Similarly, there is room for improvement in current LLM capabilities for handling “dialects” or *variants of a language that*

are characteristic of a particular group of the language’s speakers as shown in lower NLP task performances when dialects are involved (Joshi et al., 2025).

The lower performance of SOTA LLMs for non-English languages, and dialect variants has often been attributed to the absence of large-scale resources for these language variants available for LLM pretraining (Inoue et al., 2021). Indeed, for languages such as Arabic, though numerous region-specific dialects are used in spoken and informal communications (such as tweets and blogs), most available corpora are based on formal settings (such as news and education) for which Modern Standard Arabic (MSA) is employed. Possibly as a result of this disparity between the scale of MSA and Dialectal Arabic (DA) corpora available for LLM pretraining, a recent study noted that SOTA LLMs do not naturally respond in dialectal Arabic and instead employ MSA even when the prompts used are in DA (Robinson et al., 2025). Motivated by these observations on LLMs for Dialectal Arabic and to encourage research on handling the same, the Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial) 2026 included the Arabic Modeling In Your Accent (AMIYA) Shared Task (Robinson et al., 2026).

Task Description and Datasets: We (the NUS-IDS team) participated in the AMIYA Shared Task, which requires participants to contribute LLMs trained or adapted for Dialectal Arabic. Specifically, in the *Closed Data* track, task participants are required to only use the provided training data to fine-tune open-source LLMs such as those from the Llama¹ and Qwen families.²

The training data released for this competition includes sentences in Dialectal Arabic from five countries: Egypt, Morocco, Palestine, Saudi

¹<https://huggingface.co/meta-llama>

²<https://huggingface.co/Qwen>

Arabic and Syria (*Egy/Mar/Pse/Sau/Syr*) and includes sentences from parallel corpora (same sentences written in multiple dialects) such as MADAR (Bouamor et al., 2018). In addition, the development split from the FLORES+ benchmark (Perez-Ortiz et al., 2024) was made available for model tuning with final evaluation centered around a different set of dialectal prompts from the benchmark.

The metrics proposed by task organizers for evaluation include the ADI2 dialect fidelity score (Robinson et al., 2025), ChrF++ scores (Popović, 2017) used in machine translation (MT) as well as human evaluation of fluency and adherence to DA instructions. Note that the translation score ChrF++ involves the computation of character n-gram overlap with ground-truth (available for MT data) whereas ADI2 combines dialectness (MSA is considered non-dialectal) with whether the output was in the specific regional dialect (requested in the prompt) to assign a “reference-free” dialect fidelity score. For the experiments in this paper, we directly employ the implementations provided by the task organizers in AL-QASIDA³ for computing the above metrics. We refer the reader to the Task Overview Paper for details on the datasets and metrics used for this task (Robinson et al., 2026).

2 Data Augmentation

Building on the observations from earlier studies (Robinson et al., 2025) and our own investigations with smaller open-source LLMs, we note the tendency of LLMs to respond in Modern Standard Arabic (MSA) instead of Dialectal Arabic (DA). We posit that this limitation may be addressed by incentivizing the LLM to respond in DA when the instruction/prompt calls for it. For example, when the prompt is in DA (without an explicit dialect mentioned) or prompts requesting the LLM output to be in a specific dialect. That is, a pairwise preference dataset that can be used to train a given LLM with Reinforcement Learning (RL) to prefer a DA response over an MSA one is required for suitably aligning the LLM. Indeed, this notion of fine-tuning LLMs using preference data compiled from human feedback through reinforcement learning (RLHF) is a widely-used technique for aligning LLM behavior with user intent in various scenarios (Ouyang et al., 2022; Lee et al., 2024). *How*

³<https://github.com/JHU-CLSP/al-qasida>

can we build preference pairs data for RL using the given AMIYA datasets? We employ the data augmentation pipeline illustrated in Figure 1 for this precise purpose.

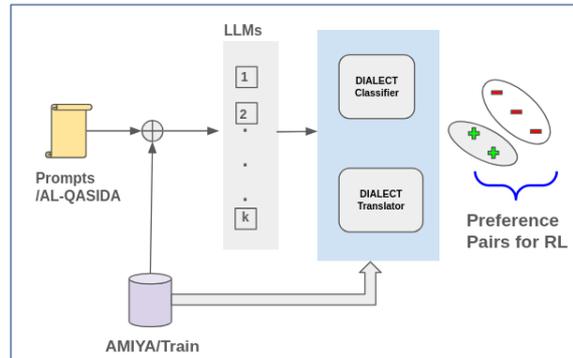


Figure 1: Data Augmentation Pipeline

First, the training data from AMIYA was used to learn dialect classification and translation models by instruction tuning open-source LLMs from Meta⁴ and Qwen.⁵ We employed these open-source LLMs and proprietary models (GPT-4o⁶ and Gemini-2.5-Flash⁷) for creating preference pairs as follows:

1. The summarization, paraphrasing, and story writing prompts from the AL-QASIDA dataset were combined with the sentences from the AMIYA training data sampled after applying thresholds on dialectness scores (Keleg et al., 2023) and number of words.⁸
2. The (prompt, input) pairs compiled in the above step were used with our selection of LLMs to obtain multiple output responses for the same pair.
3. The dialect classifier is used to label the dialects of LLM responses from the second step. If the label does not match the desired dialect, we employ the translation model to convert the response to the desired dialect and re-check

⁴<https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>

⁵<https://huggingface.co/Qwen/Qwen2.5-7B-Instruct>

⁶<https://platform.openai.com/docs/models/gpt-4o>

⁷<https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-flash>

⁸For example, short sentences (with fewer words) can be used to complete a story whereas long texts comprise inputs to summarization prompts.

using our classifier. In this fashion, we obtain “positive” and “negative” responses for the same LLM input.

4. Finally, positive and negative responses obtained above are used to form preference pairs for RL. That is, if d_i is the desired dialect, and R refers to an LLM response, $(prompt, input, R_{d_i}) \succ (prompt, input, R_{d_j})$ where $j \neq i$.

3 Experiments and Results

Datasets and Setup: Our dialect classifier was trained on $\sim 300K$ sentences obtained by randomly sampling around $\sim 50K$ sentences for each dialect from the training data in AMIYA. For training our dialect translation model, we used $\sim 697K$ sentences compiled from the parallel data provided in the training data and employing the NLLB translation model (Costa-jussa et al., 2024) when parallel translations were unavailable.⁹ In particular, we used NLLB for translating from MSA to English or vice-versa when only the MSA-DA pair or the English-DA pair was included for a given sentence in the training data from AMIYA.

For compiling preference pairs, we used about 24 prompts available for each dialect in AL-QASIDA¹⁰ and randomly sampled about 1000 sentences with dialectness scores (Keleg et al., 2023) above 0.36 in the training data, for each dialect. Additionally, we employed translation prompts for MSA to DA using random samples of sentences provided in the training data. Overall our data augmentation pipeline resulted in a set of approximately 9.6K preference pairs for use with RL. The datasets compiled by us for the AMIYA competition along with descriptions of the process are shared to enable further research on the topic.¹¹

We used the LoRA (low-rank adaptation) training scripts provided in LlamaFactory (Hu et al., 2022; Yao et al., 2024) for all our supervised fine-tuning and RL experiments. The number of epochs were set to three and default values provided in the training scripts in LlamaFactory were used for other parameters. For the rest of this paper, we refer to the base models, Llama-3.1-8B-Instruct⁴

⁹<https://huggingface.co/facebook/nllb-200-3.3B>

¹⁰https://github.com/JHU-CLSP/al-qasida/blob/main/data_processing/prompt_templates/json/dial8.json

¹¹<https://github.com/mouad157/Amiya26>

and Qwen-2.5-7B-Instruct,⁵ as **llama** and **qwen**, respectively. The base models fine-tuned with translation data are referenced as **llama-T** and **qwen-T**, respectively. We tested RL using the base models as well as our fine-tuned translation models as starting points. Models after Reinforcement Learning using the direct preference optimization algorithm (Rafailov et al., 2023) are indicated using the **+RL** suffix. All our experiments were performed on an NVIDIA A6000 server with 2 GPUs each having 48GB RAM.

Classification Performance: We show the F1 scores on a sample held-out test dataset from AMIYA that contains about 200 sentences for each dialect *Egy/Mar/Pse/Sau/Syr* as well as MSA in Figure 3. Our dialect classifier which is a Llama 3.1-8B-Instruct model fine-tuned (**FT-Llama**) on the considerably large sentence-level training data from AMIYA yields high F1 scores with the performance on the Saudi Arabian dialect being the lowest at 0.83 F1 and the Palestinian dialect being the highest at 0.94 F1. Our fine-tuned dialect classifier is able to accurately discriminate among the five dialects and MSA, and shows remarkable performance gains compared to the zero-shot setting with the base Llama model (FT-Llama versus Llama/zero in Figure 3).

We tested the NADI baseline¹² employed in previous works for dialect prediction (Abdul-Mageed et al., 2024; Robinson et al., 2025). Our **FT-Llama** predictor is instruction fine-tuned using a multiple-choice QA-style instruction “Which of the following dialects (a) Egyptian (b) Palestinian...is most applicable...”. We specifically trained our model for the five dialects and MSA targeted in the AMIYA competition using the dataset shared by them. In contrast, the NADI baseline comprises a fine-tuned MARBERT model (Abdul-Mageed et al., 2021) for discriminating up to 18 dialects of Arabic (and does not include MSA). As such, on our evaluation dataset, the F1 scores with NADI were 0.68, 0.80, 0.59, 0.79, 0.83 for Egyptian, Moroccan, Palestinian, Saudi, Syrian dialects, respectively and are significantly lower than those obtained with our **FT-Llama** model on these dialects.

Translation Performance: We evaluated our models on the FLORES+ dataset (Perez-Ortiz et al., 2024), using the dev portion included in

¹²<https://huggingface.co/AMR-KELEG/NADI2024-baseline>

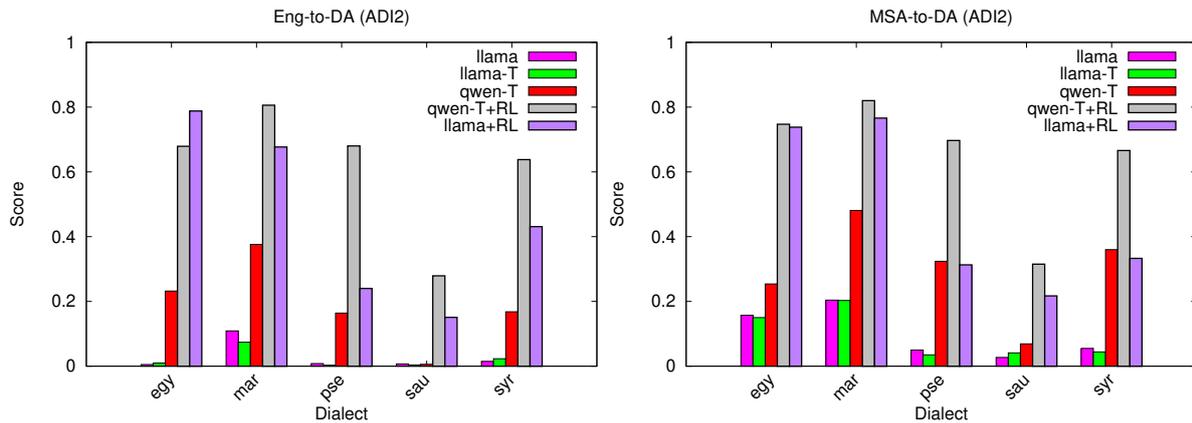


Figure 2: ADI2 Macro Scores for Eng-DA and MSA-DA in FLORES/dev

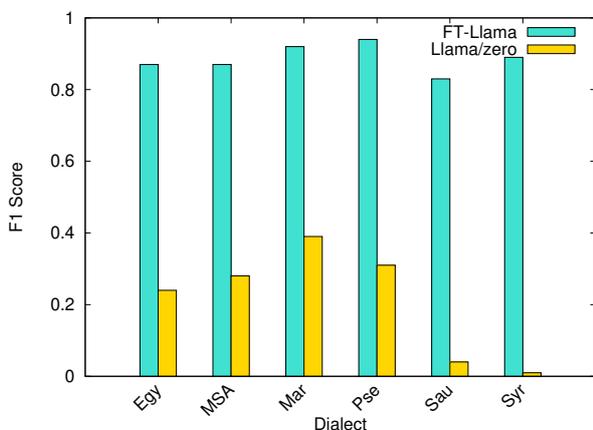


Figure 3: Dialect Classification Performance

AL-QASIDA and provided in the competition. The ADI2 macro scores for the different dialects are highlighted in Figure 2. As observed in previous studies, the dialectness scores with the base models (**llama** and **qwen**) are very low highlighting the tendency of these LLMs to “avoid” dialectal Arabic in their responses. Fine-tuning the base models with translation data yields considerable improvements for most dialects in **qwen-T** but not in **llama-T**. To test if this difference is a result of error-prone augmented data from NLLB model⁹ used for training the translation variants of the base models, we tested the models using only the “clean” translation data for which ground truth was available without employing NLLB for augmentation.

The dialectness macro scores for the different MT directions (FLORES data) are shown for the Qwen and Llama base models as well those trained with augmented (‘-Aug’) and non-augmented translation data (‘-Clean’) in Figure 7. As can be seen in this figure, for Qwen models, fine-tuning with

translation data improves performance on most dialects, with NLLB augmented data producing significant increases in dialectness scores. However, fine-tuning with translation data did not improve performance when the base model is from Llama. We hope to investigate these contradictory results in Llama versus Qwen in our on-going research beyond this paper. For the present, we speculate if this lack of improvement in Llama is possibly due to having seen the data during model pretraining and potential overlap with NLLB datasets (since NLLB is also a model released by Facebook/Meta).

In Figure 2, we note from the Eng-DA and MSA-DA plots that ADI2 macro scores significantly improve with RL. However, the overall dialectness scores are lowest for the Saudi Arabian dialect. The character n-gram overlap scores (ChrF++) (Popović, 2017) used in MT evaluation are shown for MT directions *Eng-DA*, *MSA-DA*, *DA-Eng*, *DA-MSA* in Figures 4 and 5, respectively. In these plots we note that the MT performance notably degrades for all dialects after RL. This reduction in scores reveals a core weakness in our data augmentation pipeline—focusing on using “dialect” accuracy to form preference pairs for RL is inadequate. With dialect correctness alone as an incentive, the models seem to veer off their objective to optimize the quality of the generated response. We confirmed this possibility by manually analyzing a small sample of model outputs for each MT direction.

In Figure 6, we show sample generated responses from our **qwen-T+RL** and **llama+RL** models alongside the ground-truth references. While the translations are reasonably good in some cases (Set 2), our models seem to make errors when

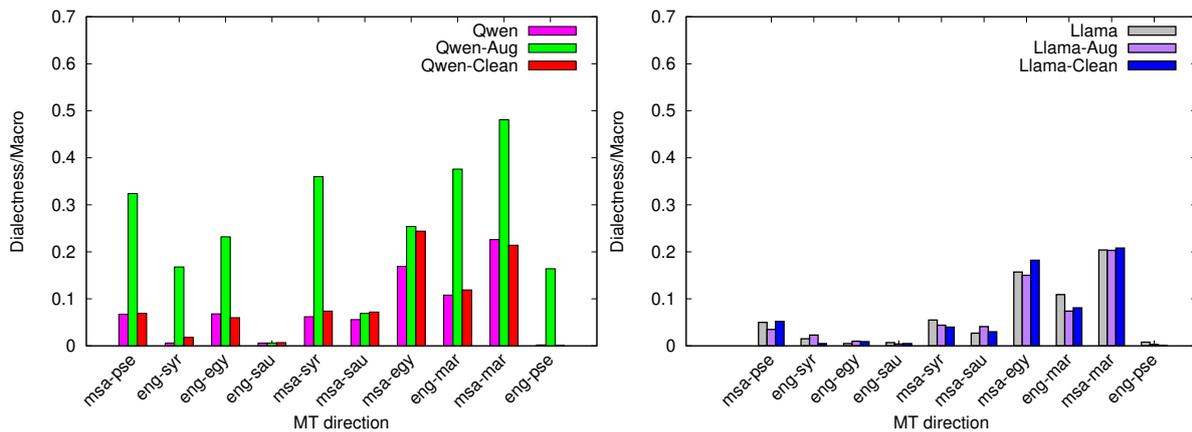


Figure 7: Dialectness Macro scores on FLORES/dev are shown for our translation models.

Model	SpBLEU	chrF++	AvgALDi
llama	15.48	32.09	0.1659
qwen	10.12	33.35	0.1429
llama-T	16.67	32.88	0.1663
qwen-T	16.71	30.22	0.2441
* <i>qwen3-T</i>	12.43	26.52	0.2194
* <i>llama+RL</i>	7.74	23.03	0.6949
qwen+RL	7.75	26.54	0.4192
llama-T+RL	8.07	24.93	0.6491
* <i>qwen-T+RL</i>	10.30	23.89	0.5780

Table 1: Validation Performance across all models

the translation inputs involve numeric information such as times and numbers and proper names. For several cases, (in models with and without RL), characters are indefinitely repeated resulting in garbage responses (‘msa-sau’ example in Set 1). RL models seem to have a propensity to choose dialectal responses. This aspect was noticed while testing DA-Eng translation data where several LLM outputs are in dialects that can be written using the Latin script (‘egy-eng’ example in Set 1) instead of in English as required in the prompt.

We note from Figures 4 and 5 that the translation scores on the Saudi dialect for base models for MT directions (MSA-DA and DA-MSA) are unusually higher than for the rest of the dialects. This peculiarity coupled with the lower classification performance in Figure 3 as well as the low dialectness scores in Figure 2 for this dialect is indicative of the proximity between the Saudi dialect and MSA compared to the other dialects considered in this evaluation (Alsudais et al., 2022) and also seems to be an artifact of the FLORES dataset which mostly contains MSA for Saudi portion of

the dataset (Perez-Ortiz et al., 2024).

Model selection for the competition: For selecting the top-3 models for consideration in the final evaluation of the shared task, we set aside about 50 examples for each dialect from our data augmentation pipeline for validation, by treating the GPT-generated response in the desired dialect (as identified by our classifier) as “gold/correct” response. All model variations were tested on this subset and the models indicated with an asterisk (*) in Table 1 were chosen for final evaluation. **llama+RL** outputs correspond to our *primary* run in AMIYA whereas ‘**qwen-T+RL**’ and ‘**qwen3-T**’ outputs correspond to the *contrastive1* and *contrastive2* runs, respectively.¹³

Our test/competition performance (Robinson et al., 2026), provided by the task organizers, is shown in Table 2. In this table, we included the best scores among our different models for each metric and the overall best scores from the competition. The primary and contrastive runs are indicated using ‘p’ and ‘c’, respectively. Mirroring our own evaluation on the FLORES dataset, our primary model obtains a significantly high score on the ADI2 metric but the translation scores are significantly lower than the best scores in the competition.

4 Conclusions

We participated in the Dialectal Arabic modeling shared task (AMIYA@VarDial 2026) in the *Closed Data* track. To handle current LLM models’ tendency to respond in Modern Standard Arabic

¹³The ‘qwen3-T’ model is similar to qwen-T but uses the base model from <https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507>.

Team	Overall ADI2	DA-ENG	ENG-DA	DA-MSA	MSA-DA
NUS-IDS	0.629 (p)	21.338 (c1)	15.764 (c1)	19.738 (c1)	17.280 (p)
MBZUAI	0.452 (c2)	53.436 (c1)	34.314 (c1)	50.639 (c1)	43.728 (c1)

Table 2: Our best performing scores are compared with the overall best scores (in **bold** font) from the competition. Columns 3-6 show ChrF++ scores for the MT directions indicated in the header

rather than dialectal variants despite prompting, we adopted a Reinforcement Learning based solution. Our main contribution to enable RL pertains to the development of a novel data augmentation pipeline that uses the training data from the competition to learn a dialect classifier and a translator and combine it with LLM outputs from various state-of-the-art proprietary and open-source LLMs. Our experiments showcase significant performance improvements in dialectness-related metrics with RL-tuned LLMs but at the cost of a degradation in MT performance. We hope to investigate this contradictory behavior in future work.

Acknowledgments

This research was supported by Google South & Southeast Asia Research Award 2022. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the funding agency.

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.
- Muhammad Abdul-Mageed, Amr Keleg, AbdelRahim Elmadany, Chiyu Zhang, Injy Hamed, Walid Magdy, Houda Bouamor, and Nizar Habash. 2024. **NADI 2024: The fifth nuanced Arabic dialect identification shared task**. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 709–728, Bangkok, Thailand. Association for Computational Linguistics.
- Abdulkareem Alsudais, Wafa Alotaibi, and Faye Alomary. 2022. **Similarities between arabic dialects: Investigating geographical proximity**. *Information Processing Management*, 59:102770.
- Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouni, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. **The MADAR Arabic dialect corpus and lexicon**. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Marta Costa-jussa, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Gonzalez, Prangthip Hansanti, John Hoffman, and Jeff Wang. 2024. **Scaling neural machine translation to 200 languages**. *Nature*, 630.
- Md. Arid Hasan, Prerona Tarannum, Krishno Dey, Imran Razzak, and Usman Naseem. 2024. **Do large language models speak all languages equally? a comparative study in low-resource settings**. *Preprint*, arXiv:2408.02237.
- Yun He, Di Jin, Chaoqi Wang, Chloe Bi, Karishma Mandyam, Hejia Zhang, Chen Zhu, Ning Li, Tengyu Xu, Hongjiang Lv, Shruti Bhosale, Chenguang Zhu, Karthik Abinav Sankararaman, Eryk Helenowski, Melanie Kambadur, Aditya Tayade, Hao Ma, Han Fang, and Sinong Wang. 2024. **Multi-if: Benchmarking llms on multi-turn and multilingual instructions following**. *Preprint*, arXiv:2410.15553.
- 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 *ICLR 2022*.
- 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*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.
- Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dipold. 2025. **Natural language processing for dialects of a language: A survey**. *ACM Comput. Surv.*, 57(6).
- Amr Keleg, Sharon Goldwater, and Walid Magdy. 2023. **ALDi: Quantifying the Arabic level of dialectness of text**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10597–10611, Singapore. Association for Computational Linguistics.

- Janghwan Lee, Seongmin Park, Sukjin Hong, Minsoo Kim, Du-Seong Chang, and Jungwook Choi. 2024. [Improving conversational abilities of quantized large language models via direct preference alignment](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11346–11364, Bangkok, Thailand. Association for Computational Linguistics.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). *Preprint*, arXiv:2203.02155.
- Juan Antonio Perez-Ortiz, Felipe Sánchez-Martínez, Víctor M. Sánchez-Cartagena, Miquel Esplà-Gomis, Aaron Galiano Jimenez, Antoni Oliver, Claudi Aventín-Boya, Alejandro Pardos, Cristina Valdés, Jusèp Loís Sans Socasau, and Juan Pablo Martínez. 2024. [Expanding the FLORES+ multilingual benchmark with translations for Aragonese, aranese, Asturian, and Valencian](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 547–555, Miami, Florida, USA. Association for Computational Linguistics.
- Maja Popović. 2017. [chrF++: words helping character n-grams](#). In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. [Direct preference optimization: your language model is secretly a reward model](#). In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23*, Red Hook, NY, USA. Curran Associates Inc.
- Nathaniel Robinson, Shahd Abdelmoneim, Anjali Kantharuban, Otba Alsboul, Salima Lamsiyah, Kelly Marchisio, and Kenton Murray. 2026. [AMIYA shared task: Arabic Modeling In Your Accent at VarDial 2026](#). In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, Rabat, Morocco. Association for Computational Linguistics.
- Nathaniel Romney Robinson, Shahd Abdelmoneim, Kelly Marchisio, and Sebastian Ruder. 2025. [AL-QASIDA: Analyzing LLM quality and accuracy systematically in dialectal Arabic](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22048–22065, Vienna, Austria. Association for Computational Linguistics.
- Shubham Vatsal and Harsh Dubey. 2024. [A survey of prompt engineering methods in large language models for different nlp tasks](#). *Preprint*, arXiv:2407.12994.
- Zhiheng Yao, Changsong Yu, and 1 others. 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 658–671.

MBZUAI at AMIYA Shared Task 2026: Adapting Open-Source LLMs for Dialectal Arabic

Rana Gaber^{1*}, Yara Allam^{1*}, Serag Amin^{1*}, Ranwa Aly^{1*}, Bashar Alhafni²

¹Alexandria University

²Mohamed bin Zayed University of Artificial Intelligence
cds.{ranaahmed30309,yaraibrahim23394}@alexu.edu.eg
cds.{seragamin23144,ranwakhaled30408}@alexu.edu.eg
bashar.alhafni@mbzuai.ac.ae

Abstract

This paper presents our contribution to the closed data track of the AMIYA Shared Task on Dialectal Arabic text generation. In this track, we train fully open-source Large Language Models (LLMs) on five Arabic dialects: Egyptian, Moroccan, Palestinian, Saudi, and Syrian, using the provided training datasets. We experiment with different base and instruct models using several pretraining and instruction tuning approaches. In total, five models were submitted, with three variants per dialect. Our best-performing models for the five dialects are ALLaM for Egyptian, LLaMa for Moroccan, and Palestinian, and Aya for Saudi and Syrian.

1 Introduction

Arabic is a morphologically rich language characterized by diglossia (Ferguson, 1959), a linguistic phenomenon in which Modern Standard Arabic (MSA), the standard and formal form of the language, co-exists with a non-standard variety, Dialectal Arabic (DA). Complicating matters, there are multiple DA varieties, each differing from one another and from MSA in phonology, morphology, syntax, and lexicon (Keleg et al., 2023), and are commonly classified regionally (e.g., Egyptian, Levantine, Gulf).

While MSA is used in formal settings such as news and education, it is not the native language of any Arabic speaker. In contrast, dialects constitute the true native varieties of Arabic, historically connected to Classical Arabic and shaped by regional linguistic contact. Although DA is primarily spoken, it is widely used in informal written communication, particularly on social media and online platforms.

Despite their widespread use, Arabic dialects are severely underrepresented in large-scale textual

resources. Compared to abundant and standardized MSA corpora, dialectal data is scarce, noisy, and highly heterogeneous, with inconsistent orthography and frequent code-switching. As a result, LLMs are predominantly pretrained on massive quantities of MSA text, with limited exposure to dialectal Arabic. This imbalance leads to substantial performance gaps, causing LLMs to struggle with understanding, generating, and reasoning over dialectal input (Bergman and Diab, 2022).

In this paper, we aim to bridge this gap by describing our system submission to the AMIYA Shared Task (Robinson et al., 2026). We explore adapting 12 open-source LLMs to five Arabic dialects: Egyptian (EGY), Moroccan (MAR), Palestinian (PSE), Saudi (SAU), and Syrian (SYR). Our models are evaluated using AL-QASIDA benchmark (Robinson et al., 2025), an evaluation suite that measures an LLM’s dialectal fidelity, understanding, generation quality, and sensitivity to MSA-DA diglossia. Our adaptation framework consists of two primary phases: Continual Pretraining (CPT), applied to base models only, and Instruction Tuning, applied to both base and instruction-tuned models. For CPT, we explore two strategies: 1) Curriculum CPT, where we train first on MSA and English before transitioning to DA-only data; 2) Mixed CPT, where we train on a mix of MSA, English, and DA data. Our results indicate that the superiority of either CPT setup relies on the model’s architecture, while fine-tuned instruct models achieve superior performance across all dialects compared to base models adapted using CPT.

2 Background

2.1 Dialectal Arabic NLP

DA NLP research has received growing attention in the Arabic NLP community, driven largely by the development of monolingual and multilingual dialectal resources and benchmarks (McNeil and

*Equal contribution.

Faiza, 2011; Zaidan and Callison-Burch, 2011; Zbib et al., 2012; Cotterell and Callison-Burch, 2014; Salama et al., 2014; Jeblee et al., 2014; Al-Badrashiny and Diab, 2016; Zaghouni and Charfi, 2018; Abdul-Mageed et al., 2018; Bouamor et al., 2019; Sajjad et al., 2020; Abdul-Mageed et al., 2020, 2021, 2022; Nagoudi et al., 2022; Abdul-Mageed et al., 2023, 2024). These efforts have enabled research across a wide range of dialectal NLP tasks, including machine translation (MT), dialect identification, and cross-dialect modeling.

Despite this progress, DA resources remain substantially more limited compared to MSA. Dialectal data is unevenly distributed across varieties, with certain dialects receiving considerably more coverage than others. This imbalance stems in part from the fact that DA is primarily a spoken language; when written, it is most commonly produced in informal contexts such as social media and online communication. Consequently, dialectal corpora are often noisy and heterogeneous, exhibiting a lot of spelling variation. These challenges are further exacerbated by the absence of standardized orthographies for Arabic dialects. Although prior work has proposed conventional orthography for DA (Habash et al., 2012; Jarrar et al., 2016; Khalifa et al., 2018; Habash et al., 2018; Eryani et al., 2020; Alhafni et al., 2024), adoption remains limited, and most real-world data continues to reflect unconstrained user-generated text.

With the emergence of LLMs, recent research has increasingly focused on evaluating and adapting Arabic-centric models such as AceGPT (Liang et al., 2024), ALLaM (Bari et al., 2024), Fanar (Fanar Team et al., 2025), Jais (Sengupta et al., 2023; Anwar et al., 2025), NileChat (Shang et al., 2025a), and Atlas-Chat (Shang et al., 2025b) alongside the development of large-scale benchmarks targeting Arabic and its dialects (Koto et al., 2024; Faisal et al., 2024; Hijazi et al., 2024; Ashraf et al., 2025; Alghamdi et al., 2025; Mousi et al., 2025; Sadallah et al., 2025; Alwajih et al., 2025; Almatham et al., 2025; Robinson et al., 2025). Our work builds on this line of research by systematically studying the adaptation of open-source LLMs to multiple Arabic dialects under controlled training conditions, and by evaluating their dialectal fidelity, generation quality, and sensitivity to MSA-DA diglossia.

2.2 AMIYA Shared Task

The Arabic Modeling in Your Accent (AMIYA) Shared Task focuses on building and adapting

LLMs for DA. All models are evaluated using the AL-QASIDA benchmark (Robinson et al., 2025). AL-QASIDA consists of multiple evaluation tasks, including monolingual generation (DA→DA) and cross-lingual generation (EN→DA) for measuring DA *fidelity*; DA→EN MT for assessing *understanding*; EN→DA MT for evaluating generation *quality*; and MSA↔DA MT for probing *diglossia*.

The shared task has three tracks: 1) Closed Data Track; 2) Closed Models Track; and 3) Open Track. For all tracks, the shared task accepts submissions for five Arabic dialects: Egyptian (EGY), Moroccan (MAR), Palestinian (PSE), Saudi (SAU), and Syrian (SYR). We participate in the **Closed Data Track** across **all five** dialects.

3 System Overview

We investigate the adaptation of 12 open-source LLMs for DA, spanning two categories: base models and instruction-tuned models (Table 3). Our adaptation framework consists of two phases: **Continual Pretraining (CPT)** and **Instruction Tuning (IT)**. CPT is applied exclusively to base models, whereas IT is applied to both CPT-adapted base models and instruction-tuned models. For CPT, we examine two training strategies: 1) **Curriculum CPT**, in which models are first trained on MSA and English data before training on DA-only data; and 2) **Mixed CPT**, which trains models on a shuffled mixture of MSA, English, and DA data.

4 Data

We use the training datasets provided the closed data track of the shared task, namely: SauDial (Alanazi et al., 2025), ASR-EGARBCSC, MASC, DoDa (Outchakoucht and Es-Samaali, 2021), Shami Corpus (Abu Kwaik et al., 2018), Atlaset (Bounhar and Majjodi, 2025), SDC and EDC (Tarmom et al., 2019), Saudi Tweets (Alruily, 2020), SADSSlyC (Alahmari, 2025), EDGAD (ElSayed and Farouk, 2020), Casablanca (Talafha et al., 2024), JODA (Abandah et al., 2025) and UFAL (Sellat et al., 2023), with MADAR (Bouamor et al., 2018) (excluding the off-limits portion) being used for IT only. These datasets include one or multiple Arabic dialects with a subset containing bitexts in English and/or MSA.

The distribution of data used for CPT across both Curriculum CPT and Mixed CPT is in Table 1.

Dialect	DA	DA-MSA	DA-EN
EGY	223.9K	–	–
MAR	1.3M	–	233.2K
PSE	84.2K	58.5K	–
SAU	235.1K	2170	404
SYR	124.1K	127.8K	24.6K
Total	2M	188.4K	258.25K

Table 1: CPT data statistics in terms of sentences for DA monotexts and DA \leftrightarrow MSA/EN bitexts.

4.1 Data Preprocessing

To standardize the input across training data, we apply a unified preprocessing pipeline using CAMEL tools (Obeid et al., 2020). The pipeline includes diacritic removal; normalization of Alif, Ya, and Ta Marbuta; character de-duplication; and the removal of special characters, emojis, and URLs found in social media text.

Moreover, the training data we use exhibits substantial imbalance across dialects, as shown in Tables 1 and 2 for CPT and IT data, respectively. Moroccan dominates the corpus, while Egyptian, Syrian, and Saudi data are smaller, and Palestinian is particularly limited with only 12K samples after preprocessing. To address this, we augment Palestinian data with linguistically similar Jordanian samples and downsample Moroccan data, retaining approximately 40% of the available Moroccan samples to reduce majority-dialect bias.

4.2 Instruction Tuning Data

For IT, we train the model on three tasks: MT, Monolingual Text generation (MonoGen), and cross-lingual Text generation (XGen).

For MT, we utilize all available bitexts in our dataset, excluding MADAR’s off-limits portion, covering bidirectional translation for EN \leftrightarrow DA and MSA \leftrightarrow DA. We use the following instruction prompt:

Translate from {source_language} into {target_language}. Output only the translation. Do NOT output anything else before or after it.

To address the scarcity of monolingual and cross-lingual text generation data, we generate synthetic instruction-response pairs using fully open-source models: Command-A-111B (Cohere et al., 2025) for Moroccan, Palestinian, and Saudi, and Command-R-35B for Egyptian and Syrian. The

Dialect	MonoGen	XGen	MT
EGY	695	613	35.3K
MAR	633	1857	65.4k
PSE	570	525	60.4k
SAU	644	602	11.9k
SYR	752	534	248.2k
Total	3.3K	4.1K	421.2K

Table 2: IT Dataset distribution by dialect and tasks. Counts represent sentences.

selection of these models was based on a preliminary inspection to observe which model is better for each dialect. For the MonoGen task (DA \rightarrow DA), we randomly sample a subset of DA training data and use the DA instruction prompts defined by the AL-QASIDA benchmark to produce synthetic DA outputs. For the XGen task (EN \rightarrow DA), we sample a different subset of DA data and prompt LLMs to generate corresponding English instruction prompts paired with the original DA text. Prompts used to generate the synthetic data are in Appendix B.

Table 2 reports the statistics of the IT datasets. While the synthetically generated monolingual and cross-lingual data is relatively small compared to the MT data, this design choice avoids introducing large volumes of synthetic content that may negatively impact data quality.

5 Experimental Setup

All models used in our experiments are listed in Table 3. For evaluation, we adopt the metrics defined by AL-QASIDA benchmark. Specifically, we use the Arabic Dialect Identification and Dialectness (ADI2) score for both monolingual and cross-lingual generation tasks, and chrF++ (Popović, 2015, 2017) for MT. The hyper-parameters we used for our CPT and IT setups are detailed in Appendix A. To streamline the evaluation process, we set the generation limit for new tokens to 64. In addition, to ensure comparability between models, we directly provide inputs to the model without any prompts or instruction templates.

6 Results

Baselines We evaluate all models in a zero-shot setting using the instruction prompts described in Section §4.2. Baseline results for all five dialects across three tasks are reported in Table 8 in the

Model	Type
Qwen3-8B-Base (Team, 2025)	Base
Qwen3-8B (Team, 2025)	Instruct
Llama-3.1-8B (Grattafiori et al., 2024)	Base
Llama-3.1-8B-Instruct (Grattafiori et al., 2024)	Instruct
Aya-Expansive-8B (Dang et al., 2024)	Instruct
ALLaM-7B-Instruct (Bari et al., 2025)	Instruct
Gemma-2-9B (Team et al., 2024)	Base
Bloom-7b1 (Workshop, 2024)	Base
Command-R7B-Arabic (Alnumay et al., 2025)	Instruct
Fanar-1-9B (Fanar Team et al., 2025)	Base
Fanar-1-9B-Instruct (Fanar Team et al., 2025)	Instruct
Jais-2-8B-Chat (Anwar et al., 2025)	Instruct

Table 3: Models used in our experiments and their types, Base or Instruct.

Dialect	Model	MonoGen	XGen	MT
EGY	Fanar-B	0.72	0.03	29
MAR	Fanar-B	0.48	0.02	27
PSE	Jais-2	0.02	0.05	49
SAU	Jais-2	0.04	0.03	61
SYR	Fanar-B	0.32	0.02	28

Table 4: Top performing baselines for each dialect in terms of ADI2 for MonoGen and XGen, and chrF++ for MT. Best results are in **bold**.

Appendix. Table 4 summarizes the top-performing baseline model for each dialect, averaged across tasks. For Egyptian, Moroccan, and Syrian, the base version of Fanar (Fanar-B) performs best, achieving very high monolingual scores despite lower cross-lingual and MT performance. For Palestinian and Saudi, Jais-2 ranks highest, showing strong MT results.

Candidate Model Selection We conduct lightweight adaptation runs to identify promising candidate models by evaluating our three training strategies on reduced subsets of the data. For Curriculum CPT, models are continually pretrained on all available English and MSA data, followed by 30% of the DA data. For Mixed CPT, models are pretrained on a 30% mixture of English, MSA, and DA data. For IT, models are trained on 30% of the IT dataset. All results for these settings are reported in Table 9 in the Appendix. Based on these experiments, we select a subset of models for full-scale pretraining and/or continued instruction tuning, which we detail here.

We select six models for full training: three base models (**LLaMA 3.1**, **Gemma**, and **Fanar Base**) and three instruction-tuned models (**Aya**, **ALLaM**,

Dialect	Model	MonoGen	XGen	MT
EGY	ALLaM	0.67	0.27	46
MAR	LLaMa	0.72	0.51	38
PSE	ALLaM	0.21	0.05	44
SAU	Aya	0.2	0.03	61
SYR	Aya	0.25	0.1	46

Table 5: Top performing instruction tuned models for each dialect in terms of ADI2 for MonoGen and XGen, and chrF++ for MT. Best results are in **bold**.

and **Command R**). During training, we save checkpoints at multiple stages (30%, 50%, 70%, and 100% of total steps), enabling performance analysis throughout training and selection of the most effective checkpoint for submission. Table 10 in the Appendix reports the evaluation results of all shortlisted models across checkpoints.

6.1 Final Models

Following the shared task regulations, we submitted three runs per dialect, totaling 15 submissions. Runs were selected from checkpoints obtained at different training stages based on validation performance. Our submissions include early IT checkpoints of **ALLaM**, **Aya**, and **LLaMA**, a fully trained instruction-tuned **Aya** model, and a **Fanar Base** model trained using Mixed CPT followed by full IT. Table 11 in the Appendix details the submitted models per dialect.

Table 5 reports the best-performing model per dialect used for our submissions. These correspond to IT checkpoints selected after 30% of the total training steps, as they consistently achieved the strongest validation performance. Model selection was guided by a heuristic that maximizes the average performance across the three tasks, ADI2 for MonoGen and XGen, and chrF++ for MT, which provides a simple summary of cross-task behavior but may mask trade-offs between generation quality and translation performance.

We observe substantial performance variation across dialects. Moroccan and Egyptian outperform Syrian, Saudi, and Palestinian in MonoGen and XGen, while MT performance remains consistent. The highest MT score is achieved for Saudi using the Aya IT model selected at the 30% checkpoint. Notably, despite having more IT data, Syrian underperforms Egyptian, suggesting that continual pretraining data volume plays a more critical role than IT data in dialectal differentiation.

Dialect	Sys.	Gen	DA→EN	EN→DA	DA→MSA	MSA→DA
EGY	C1	0.39	53.4	34.3	50.6	43.7
	C2	0.45	43.3	31.2	39.1	36.4
	P	0.45	47.7	33.6	43.9	39.5
MAR	C1	0.54	51.0	26.8	44.1	35.3
	C2	0.58	44.6	25.8	37.9	33.1
	P	0.57	44.9	22.7	39.3	34.9
PSE	C1	0.09	58.0	34.0	42.9	40.1
	C2	0.10	50.0	30.5	41.8	42.4
	P	0.10	50.4	36.9	41.6	40.5
SAU	C1	0.09	57.9	36.2	66.3	56.9
	C2	0.14	52.2	32.6	53.2	42.8
	P	0.10	58.4	34.1	65.3	55.3
SYR	C1	0.17	52.7	31.0	44.4	36.8
	C2	0.21	48.2	25.7	34.6	37.4
	P	0.18	54.1	31.0	40.3	37.6

Table 6: Official shared task results of our Contrastive-1 (C1), Contrastive-2 (C2), and Primary (P) systems across dialects. Gen denotes overall generation performance (monolingual and cross-lingual) evaluated using ADI2. MT results are reported using chrF++. Best results per dialect and task are in **bold**.

Dialect	Adequacy	Fluency
EGY	2.1	3.6
MAR	2.0	3.2
PSE	1.6	3.5
SAU	2.1	3.4
SYR	1.6	3.4

Table 7: Human evaluation in terms of adequacy and fluency of our Primary system across dialects.

6.2 Official Results

Table 6 presents the official shared task results using automatic metrics. No single system dominates all tasks across dialects: our Contrastive-1 (C1) and Primary (P) systems are the best on several MT directions, whereas Contrastive-2 (2) consistently yields the highest generation scores. Compared to other participants, our systems perform particularly strongly on MT across dialects, whereas generation remains more competitive across teams.

Table 7 presents human evaluation results in terms of adequacy and fluency of our primary systems across dialects. Compared to other participants, our system achieves the best scores across all dialects on both dimensions, with the exception of fluency on Moroccan.

7 Error Analysis

Figure 1 in Appendix C shows t-SNE projections of average sentence-level hidden representations for the five target dialects and MSA. Each plot is constructed from 600 parallel samples per dialect from the MADAR Dev set, using representations produced by the corresponding model. The visualizations align with the findings in Section 6.1: Egyptian and Moroccan dialects form more compact and well-separated clusters, consistent with their stronger MonoGen and XGen performance, while Syrian, Saudi, and Palestinian dialects exhibit more diffuse and overlapping representations, indicating greater representational ambiguity.

8 Conclusion

In this paper, we evaluated multiple strategies for improving DA performance in LLMs as part of our participation in the AMIYA Shared Task (closed data track). Across training setups, instruction-tuned models consistently outperformed base models on the AL-QASIDA benchmark, often achieving peak performance using only 30% of the task data. These results suggest that the volume of data during continual pretraining may have a stronger effect on the performance than IT data on DA tasks.

Limitations

In this work, we explored several model training approaches on a large list of models, but we acknowledge the limitations that should be considered when interpreting the results. Although we rely exclusively on the datasets provided by the organizers, the amount of data varies across dialects, with some dialects (e.g., Moroccan) dominating the corpus, while others are underrepresented. We try to mitigate this imbalance through down-sampling and data synthesis, but these methods do not fully resolve the issue, and models continue to have uneven performance across dialects.

Additionally, the use of synthetically generated data for instruction tuning introduces the risk of noise and hallucinations, and may not accurately reflect real-world usage of these dialects. We also note that our findings may not generalize to dialects not represented in the training data. Moreover, in some analyses, we rely on average performance across tasks to summarize cross-task behavior; while this provides a simple aggregate view, it may obscure trade-offs between generation quality and translation performance.

Finally, we limit our experiments to mid-sized models (7–9B parameters) and, due to time constraints, do not conduct a more in-depth analysis of the differences between the curriculum-based and mixed pretraining strategies explored in this work.

References

- Gheith A. Abandah, Moath R. Khaleel, Iyad F. Jafar, Mohammad R. Abdel-Majeed, Yousef H. Hamdan, Ashraf E. Suyyagh, Asma A. Abdel-Karim, and Shorouq M. AlAwawdeh. 2025. [Jordanian arabic to modern standard arabic translation using a large model tuned on a purpose-built dataset and synthetic error injection](#). *Jordanian Journal of Computers and Information Technology*, 11(3):319–335.
- Muhammad Abdul-Mageed, Hassan Alhuzali, and Mohamed Elaraby. 2018. [You tweet what you speak: A city-level dataset of Arabic dialects](#). In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan. European Language Resource Association.
- Muhammad Abdul-Mageed, AbdelRahim Elmadany, Chiyu Zhang, El Moatez Billah Nagoudi, Houda Bouamor, and Nizar Habash. 2023. [NADI 2023: The fourth nuanced Arabic dialect identification shared task](#). In *Proceedings of ArabicNLP 2023*, pages 600–613, Singapore (Hybrid). Association for Computational Linguistics.
- Muhammad Abdul-Mageed, Amr Keleg, AbdelRahim Elmadany, Chiyu Zhang, Injy Hamed, Walid Magdy, Houda Bouamor, and Nizar Habash. 2024. [NADI 2024: The fifth nuanced Arabic dialect identification shared task](#). In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 709–728, Bangkok, Thailand. Association for Computational Linguistics.
- Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020. [NADI 2020: The first nuanced Arabic dialect identification shared task](#). In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 97–110, Barcelona, Spain (Online). Association for Computational Linguistics.
- Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2021. [NADI 2021: The second nuanced Arabic dialect identification shared task](#). In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 244–259, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.
- Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2022. [NADI 2022: The third nuanced Arabic dialect identification shared task](#). In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 85–97, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Kathrein Abu Kwaik, Motaz Saad, Stergios Chatzikyriakidis, and Simon Dobnik. 2018. [Shami: A corpus of Levantine Arabic dialects](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Mohamed Al-Badrashiny and Mona Diab. 2016. [LILI: A simple language independent approach for language identification](#). In *Proceedings of the International Conference on Computational Linguistics (COLING)*, Osaka, Japan.
- Salwa Saad Alahmari. 2025. [SADSLyC: A corpus for saudi arabian multi-dialect identification through song lyrics](#). In *Proceedings of the 4th Workshop on Arabic Corpus Linguistics (WACL-4)*, pages 38–43, Abu Dhabi, UAE. Association for Computational Linguistics.
- Naif Alanazi, Mohammed Al-Batineh, and Hussein Abu-Rayyash. 2025. [Saudial: The saudi arabic dialects game localization dataset](#). *Data in Brief*, 62:111906.
- Emad A. Alghamdi, Reem Masoud, Deema Alnuhait, Afnan Y. Alomairi, Ahmed Ashraf, and Mohamed Zaytoon. 2025. [AraTrust: An evaluation of trustworthiness for LLMs in Arabic](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8664–8679, Abu Dhabi, UAE. Association for Computational Linguistics.
- Bashar Alhafni, Sarah Al-Towaity, Ziyad Fawzy, Fatema Nassar, Fadhil Eryani, Houda Bouamor, and Nizar Habash. 2024. [Exploiting dialect identification in automatic dialectal text normalization](#). In

- Proceedings of the Second Arabic Natural Language Processing Conference*, pages 42–54, Bangkok, Thailand. Association for Computational Linguistics.
- Rawan Nasser Almatham, Kareem Mohamed Darwish, Raghad Al-Rasheed, Waad Thuwaini Alshammari, Muneera Alhoshan, Amal Almazrui, Asma Al Wazrah, Mais Alheraki, Firoj Alam, Preslav Nakov, Norah A. Alzahrani, Eman Albilali, Nizar Habash, Abdelrahman Mustafa El-Sheikh, Muhammad Elmallah, Hamdy Mubarak, Zaid Alyafeai, Mohamed Anwar, Haonan Li, and 24 others. 2025. [BAL-SAM: A platform for benchmarking Arabic large language models](#). In *Proceedings of The Third Arabic Natural Language Processing Conference*, pages 258–277, Suzhou, China. Association for Computational Linguistics.
- Yazeed Alnumay, Alexandre Barbet, Anna Bialas, William Darling, Shaan Desai, Joan Devassy, Kyle Duffy, Stephanie Howe, Olivia Lasche, Justin Lee, Anirudh Shrinivason, and Jennifer Tracey. 2025. [Command r7b arabic: A small, enterprise focused, multilingual, and culturally aware arabic llm](#). *Preprint*, arXiv:2503.14603.
- Meshrif Alruily. 2020. [Issues of dialectal saudi twitter corpus](#). *The International Arab Journal of Information Technology*, 17:367–374.
- Fakhraddin Alwajih, Abdellah El Mekki, Samar Mohamed Magdy, AbdelRahim A. Elmadany, Omer Nacar, El Moatez Billah Nagoudi, Reem Abdel-Salam, Hanin Atwany, Youssef Nafea, Abdulfattah Mohammed Yahya, Rahaf Alhamouri, Hamzah A. Alsayadi, Hiba Zayed, Sara Shatnawi, Serry Sibae, Yasir Ech-chammakhy, Walid Al-Dhabyani, Marwa Mohamed Ali, Imen Jarraya, and 25 others. 2025. [Palm: A culturally inclusive and linguistically diverse dataset for Arabic LLMs](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32871–32894, Vienna, Austria. Association for Computational Linguistics.
- Mohamed Anwar, Abdelhakim Freihat, George Ibrahim, Mostafa Awad, Abdelrahman Atef Mohamed Ali Sadallah, Gurpreet Gosal, Gokul Ramakrishnan, Sarath Chandran, Biswajit Mishra, Rituraj Joshi, Ahmed Frikha, Etienne Goffinet, Abhishek Maiti, Ali El Filali, Sarah Al Barri, Samujjwal Ghosh, Rahul Pal, Parvez Mullah, Awantika Shukla, and 41 others. 2025. [Jais 2: A family of Arabic-centric open large language models](#). Technical report, IFM.
- Yasser Ashraf, Yuxia Wang, Bin Gu, Preslav Nakov, and Timothy Baldwin. 2025. [Arabic dataset for LLM safeguard evaluation](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5529–5546, Albuquerque, New Mexico. Association for Computational Linguistics.
- M Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham A. Alyahya, Sultan AlRashed, Faisal A. Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhatran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Majed Alrubaian, Ali Alammari, Zaki Alawami, Abdulmohsen Al-Thubaity, and 6 others. 2024. [ALLaM: Large language models for Arabic and English](#). *Preprint*, arXiv:2407.15390.
- M Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham Abdullah Alyahya, Sultan AlRashed, Faisal Abdulrahman Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhatran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Saad Amin Hassan, Dr. Majed Alrubaian, Ali Alammari, Zaki Alawami, and 7 others. 2025. [AL-Lam: Large language models for arabic and english](#). In *The Thirteenth International Conference on Learning Representations*.
- A. Bergman and Mona Diab. 2022. [Towards responsible natural language annotation for the varieties of Arabic](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 364–371, Dublin, Ireland. Association for Computational Linguistics.
- Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouni, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. [The MADAR Arabic dialect corpus and lexicon](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Houda Bouamor, Sabit Hassan, and Nizar Habash. 2019. [The MADAR Shared Task on Arabic Fine-Grained Dialect Identification](#). In *Proceedings of the Fourth Arabic Natural Language Processing Workshop (WANLP19)*, Florence, Italy.
- Abdelaziz Bounhar and Abdeljalil El Majjodi. 2025. [Atlaset dataset for moroccan darija: From data collection, analysis, to model trainings](#). *Hugging Face Blog*.
- Team Cohere, Aakanksha, Arash Ahmadian, Marwan Ahmed, Jay Alammari, Yazeed Alnumay, Sophia Althammer, Arkady Arkhangorodsky, Viraat Aryabumi, Dennis Aumiller, Raphael Avalos, Zahara Aviv, Sammie Bae, Saurabh Baji, Alexandre Barbet, Max Bartolo, Björn Bebensee, Neeral Beladia, Walter Beller-Morales, and 207 others. 2025. [Command a: An enterprise-ready large language model](#). *ArXiv*, abs/2504.00698.
- Ryan Cotterell and Chris Callison-Burch. 2014. [A Multi-Dialect, Multi-Genre Corpus of Informal Written Arabic](#). In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 241–245, Reykjavik, Iceland.
- John Dang, Shivalika Singh, Daniel D’souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom

- Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. [Aya Expand: Combining research breakthroughs for a new multilingual frontier](#). *Preprint*, arXiv:2412.04261.
- Shereen ElSayed and Mona Farouk. 2020. [Gender identification for egyptian arabic dialect in twitter using deep learning models](#). *Egyptian Informatics Journal*, 21.
- Fadhl Eryani, Nizar Habash, Houda Bouamor, and Salam Khalifa. 2020. [A spelling correction corpus for multiple Arabic dialects](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4130–4138, Marseille, France. European Language Resources Association.
- Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. [DIALECTBENCH: An NLP benchmark for dialects, varieties, and closely-related languages](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14412–14454, Bangkok, Thailand. Association for Computational Linguistics.
- Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Masoomali Fatehki, Anastasios Fragkopoulos, Maram Hasanain, and 23 others. 2025. [Fanar: An Arabic-centric multimodal generative AI platform](#).
- Charles F Ferguson. 1959. Diglossia. *Word*, 15(2):325–340.
- 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, and 542 others. 2024. [The Llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Nizar Habash, Mona Diab, and Owen Rambow. 2012. Conventional Orthography for Dialectal Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 711–718, Istanbul, Turkey.
- Nizar Habash, Salam Khalifa, Fadhl Eryani, Owen Rambow, Dana Abdulrahim, Alexander Erdmann, Reem Faraj, Wajdi Zaghouni, Houda Bouamor, Nasser Zalmout, Sara Hassan, Faisal Al shargi, Sakhar Alkhereyf, Basma Abdulkareem, Ramy Eskander, Mohammad Salameh, and Hind Saddiki. 2018. Unified Guidelines and Resources for Arabic Dialect Orthography. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.
- Faris Hijazi, Somayah AlHarbi, Abdulaziz AlHussein, Harethah Abu Shairah, Reem AlZahrani, Hebah Al-Shamlan, Omar Knio, and George Turkiyyah. 2024. [ArabLegalEval: A multitask benchmark for assessing Arabic legal knowledge in large language models](#). In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 225–249, Bangkok, Thailand. Association for Computational Linguistics.
- Mustafa Jarrar, Nizar Habash, Faeq Alrimawi, Diyam Akra, and Nasser Zalmout. 2016. Curras: an annotated corpus for the Palestinian Arabic dialect. *Language Resources and Evaluation*, pages 1–31.
- Serena Jeblee, Weston Feely, Houda Bouamor, Alon Lavie, Nizar Habash, and Kemal Oflazer. 2014. Domain and dialect adaptation for machine translation into egyptian Arabic. In *Proceedings of the Workshop for Arabic Natural Language Processing (WANLP)*, pages 196–206, Doha, Qatar.
- Amr Keleg, Sharon Goldwater, and Walid Magdy. 2023. [ALDi: Quantifying the Arabic level of dialectness of text](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10597–10611, Singapore. Association for Computational Linguistics.
- Salam Khalifa, Nizar Habash, Fadhl Eryani, Ossama Obeid, Dana Abdulrahim, and Meera Al Kaabi. 2018. A morphologically annotated corpus of emirati Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.
- Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Sadallah, Aisha Alraeesi, Khalid Al-mubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024. [ArabicMMLU: Assessing massive multitask language understanding in Arabic](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5622–5640, Bangkok, Thailand. Association for Computational Linguistics.
- Juhao Liang, Zhenyang Cai, Jianqing Zhu, Huang Huang, Kewei Zong, Bang An, Mosen Alharthi, Juncai He, Lian Zhang, Haizhou Li, Benyou Wang, and Jinchao Xu. 2024. [Alignment at pre-training! towards native alignment for arabic LLMs](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Karen McNeil and Miled Faiza. 2011. Tunisian Arabic Corpus : Creating a Written Corpus of an "Unwritten" Language. In *Proceedings of the Workshop on Arabic Corpus Linguistics (WACL)*.
- Basel Mousi, Nadir Durrani, Fatema Ahmad, Md. Arid Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur Absar Chowdhury, and Firoj Alam. 2025. [AraDiCE: Benchmarks for dialectal and cultural capabilities in LLMs](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4186–4218, Abu Dhabi, UAE. Association for Computational Linguistics.
- El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022. [AraT5: Text-to-text transformers for Arabic language generation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 628–647, Dublin, Ireland. Association for Computational Linguistics.

- Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. [CAMEL tools: An open source python toolkit for Arabic natural language processing](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association.
- Aissam Outchakoucht and Hamza Es-Samaali. 2021. [Moroccan dialect -darija- open dataset](#). *Preprint*, arXiv:2103.09687.
- Maja Popović. 2015. [chrF: character n-gram F-score for automatic MT evaluation](#). In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.
- Maja Popović. 2017. [chrF++: words helping character n-grams](#). In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.
- Nathaniel Robinson, Shahd Abdelmoneim, Anjali Kantharuban, Otba Alsoul, Salima Lamsiyah, Kelly Marchisio, and Kenton Murray. 2026. AMIYA shared task: Arabic Modeling In Your Accent at VarDial 2026. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, Rabat, Morocco. Association for Computational Linguistics.
- Nathaniel Romney Robinson, Shahd Abdelmoneim, Kelly Marchisio, and Sebastian Ruder. 2025. [AL-QASIDA: Analyzing LLM quality and accuracy systematically in dialectal Arabic](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22048–22065, Vienna, Austria. Association for Computational Linguistics.
- Abdelrahman Sadallah, Junior Cedric Tonga, Khalid Almubarak, Saeed Almheiri, Farah Atif, Chartrine Qwaider, Karima Kadaoui, Sara Shatnawi, Yaser Alesh, and Fajri Koto. 2025. [Commonsense reasoning in Arab culture](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7695–7710, Vienna, Austria. Association for Computational Linguistics.
- Hassan Sajjad, Ahmed Abdelali, Nadir Durrani, and Fahim Dalvi. 2020. [AraBench: Benchmarking dialectal Arabic-English machine translation](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5094–5107, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Ahmed Salama, Houda Bouamor, Behrang Mohit, and Kemal Oflazer. 2014. YouDACC: the Youtube Dialectal Arabic Comment Corpus. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 1246–1251, Reykjavik, Iceland.
- Hashem Sellat, Shadi Saleh, Mateusz Krubiński, Adam Pospíšil, Petr Zemánek, and Pavel Pecina. 2023. [UFAL parallel corpus of north levantine 1.0](#). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).
- Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Xudong Han, Sondos Mahmoud Bsharat, and 13 others. 2023. [Jais and Jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models](#). *Preprint*, arXiv:2308.16149.
- Guokan Shang, Hadi Abdine, Ahmad Chamma, Amr Mohamed, Mohamed Anwar, Abdelaziz Bounhar, Omar El Herraoui, Preslav Nakov, Michalis Vazirgiannis, and Eric P. Xing. 2025a. [Nile-chat: Egyptian language models for Arabic and Latin scripts](#). In *Proceedings of The Third Arabic Natural Language Processing Conference*, pages 306–322, Suzhou, China. Association for Computational Linguistics.
- Guokan Shang, Hadi Abdine, Yousef Khoubrane, Amr Mohamed, Yassine Abbahaddou, Sofiane Ennadir, Imane Momayiz, Xuguang Ren, Eric Moulines, Preslav Nakov, Michalis Vazirgiannis, and Eric Xing. 2025b. [Atlas-chat: Adapting large language models for low-resource Moroccan Arabic dialect](#). In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 9–30, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Bashar Talafha, Karima Kadaoui, Samar Mohamed Magdy, Mariem Habiboullah, Chafei Mohamed Chafei, Ahmed Oumar El-Shangiti, Hiba Zayed, Mohamedou Cheikh Tourad, Rahaf Alhamouri, Rwa Assi, Aisha Alraeesi, Hour Mohamed, Fakhreddin Alwajih, Abdelrahman Mohamed, Abdellah El Mekki, El Moatez Billah Nagoudi, Benelhadj Djelloul Mama Saadia, Hamzah A. Alsayadi, Walid Al-Dhabyani, and 8 others. 2024. [Casablanca: Data and models for multidialectal Arabic speech recognition](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21745–21758, Miami, Florida, USA. Association for Computational Linguistics.
- Taghreed Tarmom, William Teahan, Eric Atwell, and Mohammad Alsalka. 2019. [Compression vs traditional machine learning classifiers to detect code-switching in varieties and dialects: Arabic as a case study](#). *Natural Language Engineering*, 26.
- Gemma Team, 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, and 89 others. 2024. [Gemma: Open models based on gemini research and technology](#). *Preprint*, arXiv:2403.08295.
- Qwen Team. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.

BigScience Workshop. 2024. [bloom-7b1 \(revision 6232703\)](#).

Wajdi Zaghouani and Anis Charfi. 2018. ArapTweet: A Large Multi-Dialect Twitter Corpus for Gender, Age and Language Variety Identification. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

Omar F Zaidan and Chris Callison-Burch. 2011. The Arabic Online Commentary Dataset: an Annotated Dataset of Informal Arabic With High Dialectal Content. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, pages 37–41.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stalard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan, and Chris Callison-Burch. 2012. Machine Translation of Arabic Dialects. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 49–59, Montréal, Canada.

A Details Experimental Details

A.1 Continual Pretraining Setup

Across both CPT strategies, we use a batch size of 8 with gradient accumulation over 8 steps. Models are optimized using AdamW with a cosine learning rate schedule and a warm-up ratio of 0.03. Training is conducted in bfloat16 precision with gradient checkpointing enabled. The maximum sequence length is set to 256 tokens. We adopt a higher peak learning rate of 3×10^{-5} for Mixed CPT and for the initial curriculum CPT stage (training on EN and MSA), followed by a reduced learning rate of 1×10^{-5} during the dialect-only curriculum CPT stage to mitigate catastrophic forgetting.

To analyze intermediate adaptation behavior, additional checkpoints are saved at approximately 30% of training progress for both the Mixed and the dialect-only curriculum CPT strategies. Final checkpoints are saved at the end of each training run and used for subsequent instruction tuning and evaluation.

A.2 Instruction Tuning Setup

The instruction tuning setup was standardized across all experiments, using a maximum sequence length of 256, a learning rate of 3×10^{-5} , a batch size of 8 with gradient accumulation over 8 steps, a weight decay of 0.01, and training for a single epoch to avoid the risk of overfitting. Models were optimized using the AdamW optimizer. All instruction tuning experiments were implemented using Hugging Face’s Transformers.

B IT Synthetic Data Generation Prompts

MonoGen prompt Given the following question: **text**, generate a natural and contextually appropriate response written in **DA**.

- Only generate the response; do not generate anything else.
- The response must be a reasonable reply to the given prompt.
- Do NOT reply in MSA no matter what DA and only DA.

Where **DA** represents the dialect of the monolingual sample, and **text** refers to the monolingual sample embedded in the DA prompt from the Qasida repository.

XGen prompt Generate an English question where text is the direct answer.

Requirements:

- The question must be specific enough that text is the complete, natural response
- You **MUST** tell the user to answer in DA
- The question itself **MUST** be in English.
- Output the English question only while naturally requesting to respond in DA.
- Only say “Respond in DA” without saying please.

Where **DA** is the dialect of the sample, and **text** is the monolingual sample.

Model	EGY				MAR				PSE				SAU				SYR			
	MonoGen	XGen	MT	Avg																
ALLaM	0.33	0.03	44	0.27	0.16	0.00	44	0.20	0.06	0.00	43	0.16	0.07	0.00	60	0.22	0.07	0.00	45	0.17
Aya	0.43	0.03	31	0.26	0.39	0.04	27	0.23	0.07	0.02	28	0.12	0.12	0.02	41	0.18	0.24	0.04	29	0.19
Bloom	0.34	0.00	11	0.15	0.22	0.00	11	0.11	0.04	0.00	11	0.05	0.04	0.00	12	0.05	0.08	0.00	11	0.06
CommandR	0.48	0.05	30	0.28	0.42	0.05	29	0.25	0.08	0.02	31	0.14	0.10	0.02	45	0.19	0.28	0.03	31	0.21
Fanar-B	0.72	0.03	29	0.35	0.48	0.02	27	0.26	0.15	0.01	27	0.14	0.20	0.01	37	0.19	0.32	0.02	28	0.21
Fanar-I	0.16	0.05	41	0.21	0.21	0.07	36	0.21	0.03	0.05	39	0.16	0.04	0.04	46	0.18	0.09	0.05	39	0.18
Gemma	0.50	0.00	19	0.23	0.37	0.00	20	0.19	0.08	0.00	20	0.09	0.13	0.00	22	0.12	0.18	0.00	20	0.13
Jais-2	0.19	0.05	51	0.25	0.21	0.06	46	0.24	0.02	0.05	49	0.19	0.04	0.03	61	0.23	0.02	0.04	50	0.19
LLaMa-B	0.27	0.00	11	0.13	0.28	0.00	12	0.13	0.07	0.00	11	0.06	0.08	0.00	15	0.08	0.13	0.00	12	0.08
LLaMa-I	0.41	0.05	28	0.25	0.39	0.04	25	0.23	0.08	0.06	27	0.14	0.05	0.05	38	0.16	0.13	0.05	27	0.15
Qwen3-B	0.51	0.01	27	0.26	0.46	0.01	24	0.24	0.10	0.01	25	0.12	0.09	0.01	36	0.15	0.27	0.01	25	0.18
Qwen3-I	0.11	0.00	17	0.09	0.08	0.00	14	0.07	0.02	0.00	15	0.06	0.03	0.00	22	0.08	0.04	0.00	15	0.06

Table 8: Zero-shot baseline results across for all models across dialects in terms of ADI2 for MonoGen and XGen, chrF++ for MT. Avg denotes the macro-average performance across the three tasks after normalizing chrF++ to [0, 1]. B and I denote base and instruct models, respectively. Best scores per dialect are in **bold**.

Model	EGY				MAR				PSE				SAU				SYR			
	MonoGen	XGen	MT	Avg	MonoGen	XGen	MT	Avg	MonoGen	XGen	MT	Avg	MonoGen	XGen	MT	Avg	MonoGen	XGen	MT	Avg
ALLaM-IT	0.67	0.27	46	0.47	0.73	0.36	41	0.50	0.21	0.05	44	0.23	0.24	0.06	51	0.27	0.15	0.13	41	0.23
Aya-IT	0.62	0.16	51	0.43	0.64	0.39	46	0.50	0.15	0.01	46	0.21	0.19	0.03	61	0.28	0.25	0.10	46	0.27
Bloom-Curr-IT	0.64	0.14	41	0.40	0.65	0.20	34	0.40	0.17	0.02	38	0.19	0.08	0.01	51	0.20	0.20	0.09	33	0.21
CommandR-IT	0.53	0.3	16	0.33	0.69	0.52	13	0.45	0.14	0.05	22	0.14	0.14	0.07	31	0.17	0.22	0.15	11	0.16
Fanar-Curr-IT	0.58	0.15	37	0.37	0.66	0.27	33	0.42	0.15	0.02	39	0.19	0.06	0.02	45	0.18	0.22	0.07	35	0.21
Fanar-Mix-IT	0.67	0.13	32	0.37	0.71	0.34	31	0.45	0.17	0.02	32	0.17	0.17	0.03	42	0.21	0.21	0.09	30	0.20
Fanar-IT	0.57	0.12	18	0.29	0.62	0.32	15	0.36	0.18	0.05	19	0.14	0.05	0.03	24	0.11	0.17	0.08	16	0.14
Gemma-Curr-IT	0.51	0.17	22	0.30	0.68	0.28	18	0.38	0.17	0.04	27	0.16	0.06	0.05	27	0.13	0.18	0.10	28	0.19
Gemma-Mix-IT	0.64	0.41	33	0.46	0.67	0.30	29	0.42	0.15	0.04	35	0.18	0.22	0.05	42	0.23	0.21	0.05	29	0.18
Jais-2-IT	0.44	0.27	44	0.38	0.44	0.36	39	0.40	0.16	0.04	42	0.21	0.13	0.07	50	0.23	0.14	0.11	39	0.21
LLaMa-Curr-IT	0.27	0.12	39	0.26	0.41	0.29	35	0.35	0.10	0.07	40	0.19	0.04	0.05	48	0.19	0.13	0.07	38	0.19
LLaMa-Mix-IT	0.29	0.12	38	0.26	0.41	0.27	34	0.34	0.10	0.02	40	0.17	0.04	0.04	47	0.18	0.12	0.09	35	0.19
LLaMa-IT	0.49	0.25	43	0.39	0.72	0.51	38	0.54	0.14	0.07	40	0.20	0.04	0.03	51	0.19	0.23	0.20	37	0.27
Qwen3-Curr-IT	0.25	0.07	39	0.24	0.68	0.27	33	0.43	0.16	0.02	38	0.19	0.04	0.01	46	0.17	0.23	0.10	35	0.23
Qwen3-IT	0.25	0.05	39	0.23	0.67	0.27	33	0.42	0.13	0.02	38	0.18	0.06	0.01	46	0.18	0.23	0.09	36	0.23

Table 9: Evaluation results of partially trained models using 30% of the data across dialects in terms of ADI2 for MonoGen and XGen and chrF++ for MT. Avg denotes the macro-average performance across the three tasks after normalizing chrF++ to [0, 1]. Curr and Mix denote Curriculum and Mixed CPT, respectively, and IT denotes instruction tuning. Best scores per dialect are in **bold**.

Model	EGY				MAR				PSE				SAU				SYR			
	MonoGen	XGen	MT	Avg																
ALLaM-30%	0.67	0.27	46	0.47	0.73	0.36	41	0.50	0.21	0.05	44	0.23	0.24	0.06	51	0.27	0.15	0.13	41	0.23
ALLaM-50%	0.26	0.17	27	0.23	0.34	0.23	23	0.27	0.11	0.02	27	0.13	0.10	0.07	31	0.16	0.11	0.09	23	0.14
ALLaM-70%	0.24	0.16	23	0.21	0.33	0.19	21	0.24	0.11	0.03	25	0.13	0.08	0.05	29	0.14	0.09	0.08	22	0.13
ALLaM-100%	0.27	0.18	24	0.23	0.35	0.23	23	0.27	0.12	0.05	26	0.14	0.09	0.06	31	0.15	0.10	0.09	22	0.14
Aya-30%	0.62	0.16	51	0.43	0.64	0.39	46	0.50	0.15	0.01	46	0.21	0.19	0.03	61	0.28	0.25	0.10	46	0.27
Aya-50%	0.31	0.17	48	0.32	0.43	0.28	43	0.38	0.11	0.02	43	0.19	0.11	0.06	57	0.25	0.16	0.09	43	0.23
Aya-70%	0.31	0.17	47	0.32	0.41	0.27	43	0.37	0.10	0.03	43	0.19	0.11	0.06	56	0.24	0.14	0.11	42	0.22
Aya-100%	0.54	0.18	53	0.42	0.65	0.32	47	0.48	0.17	0.01	46	0.21	0.18	0.02	63	0.28	0.20	0.09	48	0.26
CommandR-30%	0.53	0.30	16	0.33	0.69	0.52	13	0.45	0.14	0.05	22	0.14	0.14	0.07	31	0.17	0.22	0.15	11	0.16
CommandR-50%	0.36	0.17	45	0.33	0.42	0.26	40	0.36	0.13	0.02	43	0.19	0.10	0.05	51	0.22	0.17	0.09	39	0.22
CommandR-70%	0.36	0.19	45	0.33	0.42	0.26	40	0.36	0.12	0.01	44	0.19	0.09	0.05	51	0.22	0.16	0.08	39	0.21
CommandR-100%	0.36	0.21	45	0.34	0.43	0.30	40	0.38	0.13	0.02	43	0.19	0.10	0.06	51	0.22	0.17	0.08	40	0.22
Fanar-Mix-30%	0.68	0.18	44	0.43	0.74	0.26	40	0.47	0.16	0.03	42	0.20	0.21	0.05	50	0.25	0.21	0.10	39	0.23
Fanar-Mix-50%	0.68	0.18	41	0.42	0.76	0.23	37	0.45	0.17	0.04	42	0.21	0.18	0.05	43	0.22	0.26	0.12	37	0.25
Fanar-Mix-70%	0.69	0.19	45	0.44	0.74	0.23	39	0.45	0.19	0.03	44	0.22	0.22	0.04	45	0.24	0.20	0.10	40	0.23
Fanar-Mix-100%	0.68	0.19	46	0.44	0.74	0.24	40	0.46	0.20	0.04	44	0.23	0.22	0.04	44	0.23	0.19	0.11	40	0.23
Gemma-Mix-30%	0.33	0.21	34	0.29	0.40	0.28	31	0.33	0.11	0.05	36	0.17	0.12	0.08	41	0.20	0.14	0.10	33	0.19
Gemma-Mix-50%	0.35	0.21	38	0.31	0.44	0.30	34	0.36	0.11	0.05	39	0.18	0.13	0.08	43	0.21	0.16	0.1	33	0.20
Gemma-Mix-70%	0.35	0.21	37	0.31	0.42	0.27	34	0.34	0.12	0.05	39	0.19	0.11	0.08	41	0.20	0.14	0.11	34	0.20
Gemma-Mix-100%	0.36	0.2	38	0.31	0.43	0.30	34	0.36	0.13	0.06	40	0.20	0.11	0.07	43	0.20	0.14	0.10	34	0.19

Table 10: Evaluation results of shortlisted models at instruction-tuning checkpoints (30%, 50%, 70%, and 100%) across dialects in terms of ADI2 for MonoGen and XGen and chrF++ for MT. Avg denotes the macro-average performance across the three tasks after normalizing chrF++ to [0, 1]. Best scores per dialect are in **bold**.

Submission	EGY	MAR	PSE	SAU	SYR
Primary (P)	ALLaM-30%	LLaMa-30%	ALLaM-30%	Aya-30%	Aya-30%
Contrastive-1 (C1)	Aya-30%	Aya-30%	Aya-100%	Aya-100%	Aya-100%
Contrastive-2 (C2)	Fanar-Mix-100%	ALLaM-30%	Fanar-Mix-100%	ALLaM-30%	LLaMa-30%

Table 11: Submitted models per dialect. Percentages denote the training step at which each checkpoint was selected.

C Error Analysis



Figure 1: t-SNE projections of the average hidden representations for sentences across the five dialects, categorized by model.

A Closed-Track System for Palestinian Arabic in the AMIYA Shared Task

Khaleel Hamad
Brigham Young University
khamad@byu.edu

Ahmad Al-Najjar
University Of Utah
u1386803@utah.edu

Abstract

We describe a closed track system for modeling Palestinian Arabic that is developed for the AMIYA shared task using a parameter efficient fine-tuning strategy. A 1.5B instruction-tuned language model was adapted with LoRA (Hu et al., 2021), updating only .28% of the model parameters, and trained on an aggregated set of conversations between Palestinians and resources covering both translation and generation. Model selection was guided by a comparative benchmark that prioritized performance efficiency and its tradeoffs. At the same time the paper focuses on targeting error analysis as well as structured instruction following. These findings illustrate both the viability and shed light on the current limitations of efficient adaptation methods for low-resource Arabic dialects.

1 Introduction

This paper presents our submission to the AMIYA Shared Task (Arabic Modeling In Your Accent), which aims to advance dialectal Arabic language modeling. The competition includes three tracks (1) closed data, (2) closed model, and (3) closed. Each imposed different constraints on the allowed training data and/or model choice. We participated in the closed track as it aligns with the task constraints while still enabling informed system design and evaluation. Following the shared task guidelines, which restrict the use of external resources in the closed track, we carefully selected and processed only the permitted datasets relevant to our target variety. We eventually selected the Palestinian dialect track, motivated by our prior experience developing a dialect-focused model and our familiarity with Palestinian Arabic, which supported dialect-aware data selection and error analysis. This paper makes five primary contributions, each addressing a distinct stage of the system development process: (1) a multi-source Palestinian

Arabic dataset aggregation framework combining more than four dialectal corpora; (2) a parameter-efficient fine-tuning pipeline using LoRA adapters (rank=16, alpha=32), updating only 0.28% of model parameters; (3) a structured model comparison framework enabling systematic evaluation of multiple LLM architectures using standardized diagnostic metrics; (4) an instruction-tuning strategy targeting both generation and translation tasks through unified prompt formatting; and (5) a flexible data preparation pipeline supporting reproducible experimentation. Together, these contributions highlight both the scarcity of Palestinian dialect resources and the importance of shared-task initiatives such as AMIYA in advancing inclusive Arabic language technologies.

2 Database

The dataset is a Palestinian Arabic corpus prepared for the AMIYA shared task at VarDial 2026 and is used to train and evaluate LLMs for dialect translation and generation. Examples are combined from several existing dialect resources into a single, constantly formatted JSONL dataset with pairs of instruction-response.

Our data is constructed from four underlying Palestinian dialect sources while the Combined Dialect Dataset is a reference to their union.(Abdul-Mageed et al., 2024; Talafha et al., 2024).We extract the Palestinian subset, which adds conversational, spoken-style utterances to the corpus. The dataset is constructed from the following dialectal resources:

1. **Maknuune Corpus** (Dibas et al., 2022): A Palestinian Arabic lexicon containing dialect words and phrases paired with Modern Standard Arabic (MSA) and English glosses. The corpus provides lexical entries and example sentences, enabling PalestinianMSA and PalestinianEnglish translation tasks.

2. **Shami Dataset** (Kwaik et al., 2018): A Levantine Arabic corpus derived from Twitter data. We retain only samples explicitly annotated as Palestinian dialect to capture user-generated, informal language typical of social media.
3. **Casablanca Corpus** (Abdul-Mageed et al., 2024; Talafha et al., 2024): A speech transcription dataset annotated for dialect, gender, and code-switching. We extract the Palestinian subset to incorporate spoken and conversational utterances.
4. **MASC Corpus**: A Jordanian Arabic review dataset consisting of 1,414 sentiment-annotated customer reviews. This corpus is used exclusively for auxiliary diagnostic experiments and is not included in AMIYA training or evaluation.
5. **JODA Corpus**: A large Jordanian Arabic corpus comprising social media text, film transcripts, and aligned MSA translations. Similar to MASC, this dataset is used only for diagnostic analysis and is excluded from AMIYA system training.
6. **Combined Dialect Dataset (ours)**: A unified JSONL corpus created by aggregating Palestinian Arabic instances from Maknuune, Shami, and Casablanca. Each example preserves source metadata and is formatted into instructionresponse pairs for generation and translation tasks.

2.1 Data Statistics

The Combined Dialect Dataset (unified corpus) contains 155,299 entries drawn from the five Arabic dialect datasets, the average text length of about 50 characters and a median of 35 characters. At the dialect level Palestinian contributes 48,278 instances, Jordanian contributes 69,262, and Syrian 37,759, giving a reasonably balanced Levantine pool from which the Palestinian subset can be extracted and used for AMIYA. In terms of source datasets, JODA corpus is the largest component with 59,135 sentences, followed by the Shami twitter corpus (55,418), the Maknuune lexicon (36,302 entries), Casablanca speech transcriptions (3,030 segments), and the smaller MASC review corpus (1,414 documents) (Figure 1).

The Maknuune lexicon contains primarily of very short lexicon items with a mean length of

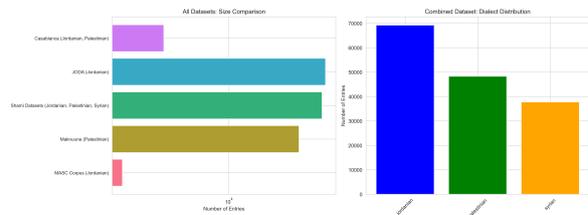


Figure 1: Overall comparison

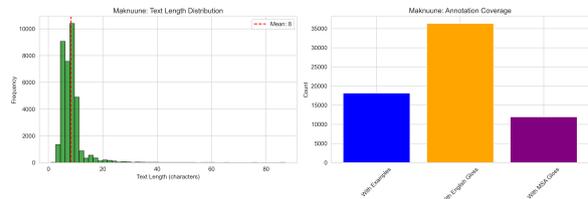


Figure 2: Maknuune analysis

roughly 8 characters ($\bar{1}.2$ words), and entries offer additional supervision: 18,140 entries include example sentences, and all entries have their English glosses paired, and 11,905 (32.8%) provide MSA translations (Figure 2). The Shami datasets contain sentence-level conversational text in Palestinian(10,642), Jordanian(7,017), and Syrian dialects(37,759), with average lengths between 67 and 86 characters with noticeable long tail of longer tweets (Figure 3). JODA corpus adds 59,135 Jordanian sentences with a mean length of 49 characters ($\bar{9}.5$ words); the split assigns about 91.5% of examples for training and 4.2% each to validation and tests as illustrated by the length histogram and split pie chart below (Figure 4).

The remaining resources contribute more domain-specific or spoken material. The MASC corpus provides 1,414 Jordanian customer reviews with an average length of 245 characters ($\bar{4}6.2$ words) annotated with positive or negative sentiment labels (Figure 5). Casablanca adds 3,030 speech transcripts including 1,334 Palestinian and 1,696 segments whose length distributions speak around 50-70 characters, with Palestinian utterance is slightly longer on average against the Jor-

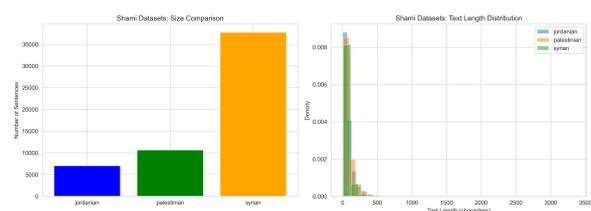


Figure 3: Shami Analysis

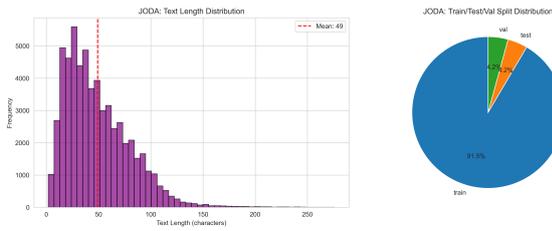


Figure 4: Joda Analysis

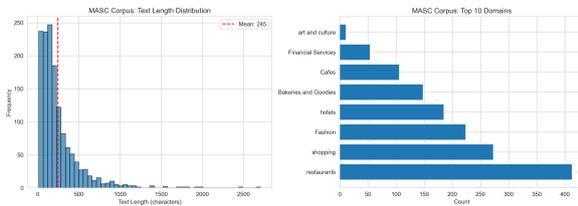


Figure 5: Masc Corpus

danian segment (Figure 6).

Overall, these statistics and their visualizations represent the pool from which we curate our Palestinian AMIYA dataset spans short lexical entities, sentence-level social media and translations pairs, and longer narrative or conversational passages, providing a diverse foundation for dialect generation and translation experiments.

Based on these statistics, we determined that sufficient Palestinian dialect data is available to support closed-track training without incorporating non-Palestinian dialects. Accordingly, only Palestinian-labeled instances from Maknuune, Shami, and Casablanca were used for AMIYA training and evaluation. Jordanian and Syrian data were retained solely for auxiliary diagnostic analyses and were excluded from all shared-task experiments.

2.2 Data Processing

For pre-processing the data we apply a uniform pre-processing pipeline that (1) loads Palestinian examples from all available sources, (2) filters and normalizes the raw text (3) converts examples into task specific input-output pairs, and (4) wrap them

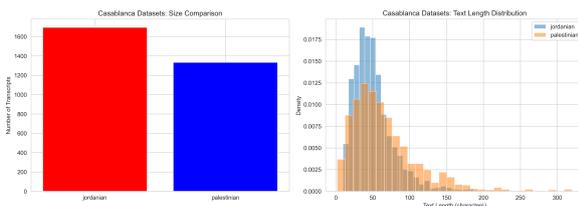


Figure 6: Casablanca

into an instruction format that is compatible with Llama-3 (Grattafiori et al., 2024). Dialect labels and encodings are inherited from the combined corpus pre-processing (section 3.b), where text has been cleaned, normalized, and dialect tags standardized to lowercase.

For generation tasks, a function is created to construct monolingual Palestinian examples by identifying the appropriate text column in each source (e.g. <Text>, < Sentences> , <example_usage>) and discarding entries with missing fields or length below a configured minimum threshold. Valid texts are used as outputs and conditions on the same content (for example, Write in Palestinian dialect: or Generate Palestinian dialect text about:), which makes the task instruction following rather than pure language modeling.

For translation tasks, a function is created that exclusively operates on the Maknuune lexicon, where it builds a Palestinian side by preferring examples sentences and falling back to the base dialect forms, splitting into multi-translations, it filters out very short or empty segments, and then creates aligned pairs for 4 directions (MSADA, DAMSA, ENDA, DAEN).

The resulting examples are converted into instruction style records, which puts each input-output pair into a task-specific template from <AMIYA_PROMPT_TEMPLATES> and stores the rendered prompt in an <instruction> field while preserving the raw input, output, task_type, and source metadata.

We partition the final instruction-formatted corpus into training, validation, and test sets using the <split_data> routine function from all examples to a pandas DataFrame, then applies a two-stage random split with 80/10/10 proportions; 80% for training, 10% for validation, and 10% for testing using <train_test_split> (Pedregosa et al., 2011) with a fixed random seed, ensuring that results are reproducible and that no example appears in more than one split.

The split is performed jointly over all task types so that each subset contains a representative mixture of generation and translation examples, while preserving the original task type and source metadata for downstream analysis.

Listing 1: Pseudocode: data preparation and model selection

```
DATASETS = [Maknuune, Shami, Casablanca]
```

```

ALL = []

for ds in DATASETS:
    rows = load(ds)
    rows = keep_label(rows, "Palestinian")
    rows = normalize_text(rows)
    rows = drop_missing(rows)

    gen = build_generation_pairs(rows)
    trn = build_translation_pairs(rows)

    ALL = ALL + gen + trn

TRAIN, DEV, TEST = split(ALL, 0.8, 0.1, 0.1,
                        seed=42)

CANDIDATES = [TinyLlama_1p1B,
              Qwen2_1p5B,
              Qwen2_7B,
              Llama3p1_8B,
              Dallah_3B,
              Jais_13B]

SCORES = {}
for m in CANDIDATES:
    subset = sample(DEV, n=500, seed=42)
    y_true = get_labels(subset)
    y_pred = infer_labels(m, subset)

    acc = accuracy(y_true, y_pred)
    f1m = macro_f1(y_true, y_pred)
    tps = throughput(m, subset)

    SCORES[m] = (acc, f1m, tps)

BEST = argmax(SCORES, key="f1m_then_tps")

```

3 Systems

Based on these statistics, we determined that sufficient Palestinian dialect data is available to support closed-track training without incorporating non-Palestinian dialects. Accordingly, only Palestinian-labeled instances from Maknune, Shami, and Casablanca were used for AMIYA training and evaluation. Jordanian and Syrian data were retained solely for auxiliary diagnostic analyses and were excluded from all shared-task experiments.

- TinyLlama-1.1B (Zhang et al., 2024)
- Qwen2-1.5B-Instruct (Yang et al., 2024)
- Qwen2-7B-Instruct (Yang et al., 2024)
- Meta-Llama-3.1-8B-Instruct (Grattafiori et al., 2024)
- Dallah-3B-v1
- Jais-13B-chat (Sengupta et al., 2023)

Metric	Value
TinyLlama-1.1B	
Accuracy	44.00%
Macro F1	20.37%
Weighted F1	26.89%
Avg Inference Time (s)	0.9161
Samples/sec	1.09
Qwen2-1.5B-Instruct	
Accuracy	44.00%
Macro F1	24.76%
Weighted F1	29.83%
Avg Inference Time (s)	0.0975
Samples/sec	10.25

Table 1: Diagnostic model performance and inference speed on a 500-sample subset. Detailed diagnostic scores were retained only for the two top-performing models, which were selected for further experimentation.

The first thing was to compare between them on a small dataset of around 500 rows, mainly due to limited resources. Next was to filter out all except the 2 highest: TinyLlama-1.1B and Qwen2-1.5B-Instruct (Table 1).

For completeness, Table ?? reports the same diagnostic metrics for the remaining candidate models evaluated on the same 500-sample subset. These results are included only to contextualize model selection and were not part of the AMIYA official evaluation.

During preliminary screening, all candidate models were evaluated using the same diagnostic setup. However, detailed per-model metrics were retained only for the two top-performing models that were selected for fine-tuning and downstream analysis. For the remaining candidates, evaluation was used solely to inform relative ranking and early filtering decisions, and detailed diagnostic outputs were not preserved.

Accuracy, predicted labels, and true labels reported in the confusion matrices correspond to an internal diagnostic dialect classification task used during model comparison. These metrics are not part of the AMIYA evaluation framework, and ADI2 or similar shared-task metrics are not applicable at this stage. The purpose of this analysis is to inspect error patterns rather than competitive task scores.

Although Qwen was the winner in most of these results, deeper analysis was required to ensure that it will be the best model to work with (Figures 7

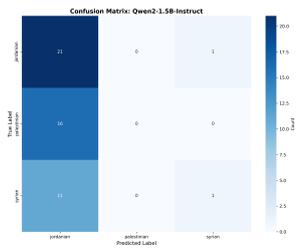


Figure 7: Confusion matrix: Qwen2-1.5B-Instruct

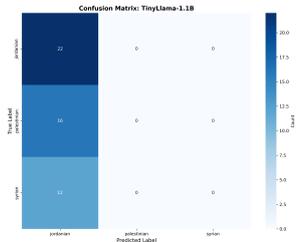


Figure 8: Confusion matrix: TinyLlama-1.1B

Metric	Value
Jordanian	
Precision	43.75%
Recall	95.45%
F1 score	60.0%
Support	22
Palestinian	
Precision	0.00%
Recall	0.00%
F1 score	0.00%
Support	16
Syrian	
Precision	50%
Recall	8.33%
F1 score	14.23%
Support	12

Table 2: Dialect classification metrics.

and 8). From these truth tables the final decision to choose Qwen was taken. Next, we examined the models understanding of the different dialects, where a new analysis was run for the models where Qwen came out with the highest scores of (Table 2). Dialect classification was used as an exclusively diagnostic tool during model selection and error analysis. It wasn't created as part of the AMIYA shared task evaluation and is not used at inference time for the submitted system. The purpose of this analysis was to assess the models' sensitivity to the varied dialects we are working on prior to fine-tuning, rather than using it to optimize performance for a standalone dialect identification task.

It is also important to note that dialect classification was conducted solely as a diagnostic tool for model selection and error analysis. It is not part of the AMIYA shared task evaluation nor used at inference time in the submitted system. In this case the results seemed to be too far apart to be used as the main decision making tool, which is why there was a restoration for other methods. So overall the decision to go with Qwen as the model to work on is a good decision because it has the better f1 score, has a much faster inference (10 samples per sec), as well as lower latency (.9161 per sample).

The final system that was built uses *Qwen/Qwen2.5-1.5B-Instruct* (Yang et al., 2024) as the base model, a 1.5 billion - parameter decoder-only transformer optimized for instruction following. The architecture follows

a standard GPT style decoder following the multihead self attention techniques, feed-forward networks, layer normalization, and residual connections. The model is loaded with mixed precision for memory efficiency, as shown in the model initialization code. As a simple baseline, we evaluated the base Qwen2.5-1.5B-Instruct model in a zero-shot setting using the same instruction templates and prompts as our fine-tuned system. Relative to this zero-shot baseline, LoRA fine-tuning improved instruction adherence and dialectal consistency, with the clearest gains observed in generation outputs.

Listing 2: Model loading code

```
model = AutoModelForCausalLM.from_pretrained(
    FINETUNING_MODEL_ID,
    dtype=torch.bfloat16 if device == "cuda"
    else torch.float32,
    device_map="auto" if device == "cuda" else
    None,
    trust_remote_code=True,
    low_cpu_mem_usage=True
)
```

This base model includes around 1.5 billion parameters, all of which remain frozen during fine-tuning. Gradient checkpointing is enabled to reduce memory consumption during training, allowing the model to really trade computation for memory by basically recomputing activations during backpropagation rather than storing them.

The fine-tuning process was done using LoRA (Hu et al., 2021) to adapt to only attention projection layers. LoRA injects trainable low-rank met-

rics into the attention mechanism, enabling efficient adaptation with minimal parameter overhead. This would target the four projection layers in each transformer block:

Listing 3: LoRA configuration and model wrapping

```
lora_config = LoraConfig(
    r=LORA_R, # Rank: 16
    lora_alpha=LORA_ALPHA, # Alpha: 32
    target_modules=LORA_TARGET_MODULES, #
        ["q_proj", "k_proj", "v_proj", "o_proj"]
    lora_dropout=LORA_DROPOUT, # Dropout: 0.1
    bias="none",
    task_type=TaskType.CAUSAL_LM,
)

model = get_peft_model(model, lora_config)
```

Here the rank parameter ($r = 16$) is controlling the dimensionality of the low-rank decomposition, while the alpha parameter ($\alpha = 32$) acts as a scaling factor. The configuration has resulted in 4.2 million trainable parameters, representing around only .28% of the base models total parameters. This allowed training in a much smaller time window and became much less resource depleting.

As for the training process it follows a structured pipeline beginning with data preparation, where Palestinian dialect examples are loaded from the JSON files and then converted to HuggingFace Dataset format (Lhoest et al., 2021). The training data here is just limited to 15,000 examples for computational efficiency, with task type distribution monitored to ensure balanced representation across the generation and translation tasks. For the final presentation model further training will be performed using the entire dataset. As for training, the **Trainer** class from huggingface transformers (Wolf et al., 2020) orchestrates the forward pass, loss computation and backpropagation as well as a couple of other tasks. The process generates comprehensive training statistics, including trainable parameter counts.

Training was conducted by using HuggingFace’s Trainer framework with AdamW optimizer, a learning rate of $2e-4$, batch size of 8 per device, and gradient accumulation steps of 4 for a total of 3 epochs. A fixed random seed was used for data splitting and training. Experiments were run on different GPUs with mixed precision enabled. These settings were kept constant across all compared models to ensure fair comparison.

4 Results

For the generation task, qualitative analysis of model output was conducted revealing both strengths and limitations. For example, when prompted to generate a greeting the model outputs natural Palestinian dialect phrases that are able to preserve regional linguistic features. Still, translation tasks are revealing limitations when translating from Modern Standard Arabic MSA to Palestinian Dialect. The model also seems to sometimes repeat the input MSA text rather than producing actual dialect output, indicating insufficient translation training data. Sample outputs have also demonstrated that the model is handling longer, contextually rich inputs better than short phrases or single words; this suggests that contextual information aids dialect identification.

Listing 4: LLM-as-judge prompt template used for qualitative audit

```
Given:
(1) Task instruction
(2) Model output
(3) Reference output

Decide one label: {Correct, PartiallyCorrect, Incorrect}

Criteria:
- For translation: meaning preservation and dialect appropriateness
- For generation: Palestinian dialect usage and instruction adherence

Return:
- label
- 1-2 sentence justification
```

4.1 LLM-based qualitative audit (error analysis)

Based on the 200-sample qualitative audit, we categorized model errors into a small number of recurring patterns. Table ?? summarizes the most frequent error types observed in the sampled outputs, along with their relative frequency and representative examples. These categories were derived inductively by inspecting incorrect and partially correct outputs and are intended to highlight systematic weaknesses rather than provide an exhaustive error taxonomy.

The most frequent errors are closely tied to data sparsity and prompt ambiguity, suggesting that richer contextual prompts and additional dialectal supervision may improve performance in future iterations.

In addition to the observed weaknesses, the model also demonstrates consistent strengths in

Error type	Freq.
Dialect drift to MSA	31%
Instruction violation	24%
Input copying	18%
Semantic shift	15%
Hallucination	12%

Table 3: Distribution of error types in the 200-sample qualitative audit.

Error type	Typical manifestation
Dialect drift	Output rendered in MSA
Instruction violation	Wrong task format
Input copying	Source text unchanged
Semantic shift	Meaning partially altered
Hallucination	Added content

Table 4: Qualitative descriptions of error types observed during audit.

Phenomenon	Success observed
Dialectal greetings	Natural informal salutations
Colloquial particles	Correct discourse markers
Verb morphology	Dialect-appropriate forms
Contextual coherence	Consistent dialect in context
Idiomatic usage	Region-appropriate expressions

Table 5: Representative strengths observed in Palestinian Arabic generation during the qualitative audit.

common dialectal constructions. Table 5 highlights representative cases where the model successfully produced natural Palestinian Arabic outputs.

To assess whether dialectal expressions were memorized or generalized, we inspected outputs corresponding to common idiomatic phrases. While some frequent expressions are likely present verbatim in the training data, the model also produces structurally similar variants that were not observed directly in the source corpora. This suggests that the model learns productive dialectal patterns rather than relying solely on memorization, although data sparsity limits the robustness of this generalization for less frequent constructions.

5 Discussion

Overall, parameter-efficient fine-tuning using LoRA proves effective for adapting a compact instruction-tuned model to Palestinian Arabic under constrained data and computational budgets. While the fine-tuned system demonstrates strengths in dialectal generation and idiomatic usage, limitations persist in translation robustness and instruction adherence. The qualitative error analysis highlights data sparsity and prompt ambiguity as primary sources of failure, indicating di-

rections for future improvements.

6 Conclusion

This paper presented our submission to the AMIYA Shared Task closed track, focusing on Palestinian Arabic modeling through parameter-efficient fine-tuning. By aggregating multiple dialectal resources and applying LoRA adaptation to a 1.5B instruction-tuned model, we demonstrate the feasibility of efficient adaptation for low-resource Arabic dialects. At the same time, persistent challenges in translation accuracy and instruction adherence underscore the need for richer, task-aligned datasets and combined qualitative-quantitative evaluation. We hope this work contributes practical insights toward more inclusive Arabic language technologies.

7 References

References

- Muhammad Abdul-Mageed, Basel Talafha, Motaz Saad, and 1 others. 2024. Casablanca: Data and models for multidialectal arabic speech recognition. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- S. Dibas, C. Khairallah, N. Habash, O. F. Sadi, and 1 others. 2022. Maknuune: A large open palestinian arabic lexicon. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*.
- A. Grattafiori, A. Dubey, Y. Jernite, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zhiqing Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- K. A. Kwaik, M. Saad, S. Chatzikyriakidis, and S. Dobnik. 2018. Shami: A corpus of levantine arabic dialects. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*.
- Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, and 1 others. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.
- F. Pedregosa, G. Varoquaux, A. Gramfort, and 1 others. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

- N. Sengupta, S. K. Sahu, B. Jia, and 1 others. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.
- B. Talafha, M. Abdul-Mageed, and 1 others. 2024. Casablanca: A multi-dialect arabic speech dataset with code-switching. In *Proceedings of EMNLP 2024*.
- Thomas Wolf, Lysandre Debut, Victor Sanh, and 1 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.
- A. Yang, B. Chen, B. Zhang, and 1 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.
- P. Zhang, G. Zeng, T. Wang, and W. Lu. 2024. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*.

Author Index

- ., Mamta, 172, 293
- Abboush, Mohammad, 359
- Abdelmoneim, Shahd, 1
- Afanasev, Iliia, 304
- Aksüt, Firat Cem, 87
- Al Almaoui, Perla, 339
- Al-Najjar, Ahmad, 385
- Alabdullah, Abdullah, 157
- Alali, Abdulhai, 352
- Alhafni, Bashar, 373
- Alkhder, Hasan, 359
- Allam, Yara, 373
- Als Boul, Otba, 1
- Aly, Ranwa, 373
- Amin, Serag, 373
- Anastasopoulos, Antonios, 210
- Arefyev, Nikolay, 275
- Bassignana, Elisa, 317
- Bawden, Rachel, 45
- Bompolas, Stavros, 210
- Bouillon, Pierrette, 339
- Broselow, Ellen, 250
- Buljan, Maja, 275
- Cercas Curry, Amanda, 317
- Chatzikyriakidis, Stergios, 101
- Chiang, David, 145
- Clérice, Thibault, 45, 223
- Cocarascu, Oana, 293
- Dent, Rasul, 223
- Dhasmana, Akriti, 145
- Dilsiz, Deniz Kaya, 18
- Dinu, Liviu P., 139
- Du, Mingzhe, 365
- Ekbal, Asif, 172
- El Zeftawy, Mohamed, 261
- Fedorova, Mariia, 275
- Fischer, Hanna, 123
- Gaber, Rana, 373
- García-Miguel, José M., 58
- Garcia, Marcos, 58
- Georgescu, Simona, 139
- Gollapalli, Sujatha Das, 365
- Guzii, Zakhar, 186
- Hakam, Mouad, 365
- Hamad, Khaleel, 385
- Han, Lifeng, 157
- Hassan, Lara, 261
- Helcl, Jindřich, 275
- Hengchen, Simon, 339
- Henriksson, Erik, 101
- Hillmann, Stefan, 87
- Hovy, Dirk, 317
- Iordache, Ioan-Bogdan, 139
- Irastortza-Urbieta, Xabier, 58
- Issam, Abderrahmane, 196, 352
- Joshi, Aditya, 18
- Kantharuban, Anjali, 1
- Karatsareas, Petros, 210
- Karunakaran, Sumirtha, 234
- Keleg, Amr, 261
- Khalak, Abdulmuizz, 196
- Khalifa, Salam, 250
- Khomenko, Pavlo, 186
- Knoeferle, Pia, 87
- Kodner, Jordan, 250
- Kuparinen, Olli, 32
- Kyslyi, Roman, 186
- Laippala, Veronika, 101
- Lameli, Alfred, 123
- Lamsiyah, Salima, 1
- Libovický, Jindřich, 70
- Lin, Chenghua, 157
- Ljubešić, Nikola, 242
- Lutgen, Anne-Marie, 113
- Mahaganapathy, Ahrane, 234
- Maheshwari, Sanjh, 293
- Marchisio, Kelly, 1
- Marchitan, Teodor-George, 139

Margariti, Antigoni, 210
Markantonatou, Stella, 210
Mederake, Nathalie, 123
Mekky, Ali, 261
Möller, Sebastian, 87
Murray, Kenton, 1
Mutal, Jonathan, 339

Nakov, Preslav, 261
Navakulan, Kavitha, 234
Nédey, Oriane, 45
Ng, See-Kiong, 365

Oepen, Stephan, 275
Onyshchenko, Bohdan, 186
Orlovskiy, Artem, 186
Ortiz Suarez, Pedro, 223

Panagiotou, Konstantina, 210
Panchal, Mihir, 172
Papadakis, Dimitrios, 101
Plaiti, Maria Elisavet, 210
Plum, Alistair, 113
Psaltaki, Erofilii, 101
Purschke, Christoph, 113

Qaddoumi, Abdelrahim, 250

Rajpoot, Aniket Singh, 293
Ralli, Angela, 210
Rambow, Owen, 250
Robinson, Nathaniel R., 1
Rønningstad, Egil, 275
Rupnik, Peter, 242

Sagot, Benoît, 45, 223
Samardžić, Tanja, 242
Sarveswaran, Kengatharaiyer, 234
Scherrer, Yves, 275
Spanakis, Gerasimos, 196
Srirag, Dipankar, 18
Srivastava, Aarohi, 145

Tsoukala, Chara, 210
Tzanakaki, Nefeli, 210

Uban, Ana Sabina, 139
Urbach, Nico, 123

Varshney, Deeksha, 172
Vico, Gianluca, 70

Zhang, Mike, 317