LM4DH 2025

Proceedings of
The First Workshop on Natural Language Processing and Language Models for Digital Humanities

associated with

The 15th International Conference on Recent Advances in Natural Language Processing RANLP'2025

Edited by Isuri Nanomi Arachchige, Francesca Frontini, Ruslan Mitkov and Paul Rayson

11 September, 2025
Varna, Bulgaria

The First Workshop on Natural Language Processing
and Language Models for Digital Humanities
Associated with the International Conference
Recent Advances in Natural Language Processing
RANLP'2025

PROCEEDINGS

Varna, Bulgaria
11 September 2025

Online ISBN 978-954-452-106-6

Designed by INCOMA Ltd.
Shoumen, BULGARIA

Welcome to the LM4DH workshop

Digital Humanities has emerged as an interdisciplinary field of research, serving as an intersection of computer science with many other fields such as linguistics, social sciences, history, psychology, etc. With the development of Large Language Models (LLMs), state-of-the-art Natural Language Processing (NLP) tasks such entity recognition, text summarisation, diachronic analysis, and sentiment modelling have been significantly enhanced, offering powerful tools to analyse and interpret complex historical and cultural data. These advancements provide powerful tools for analysing and interpreting intricate historical, cultural, and social data, enabling researchers to identify patterns, extract meaningful relationships, and generate interpretations at unprecedented scale and precision.

Language Models for Digital Humanities (LM4DH) 2025 convened a collaborative platform for researchers, practitioners, and students to explore, critique, and advance AI-driven methodologies. We aimed to share technical innovations while fostering a community dedicated to ethically grounded and socially meaningful applications of LLMs.

We received 23 high-quality submissions for the LM4DH 2025 workshop, spanning a diverse range of topics at the intersection of language models and Digital Humanities: including computational linguistics, historical document processing, music augmentation, rhetorical analysis, sociolinguistic forecasting, ancient language parsing, and mental health text classification. Following a rigorous peer-review process, 18 papers were accepted for presentation and publication in the workshop proceedings.

The workshop featured two distinguished keynote speakers who offered valuable insights at the intersection of computational linguistics and digital humanities. Dr. Alessio Miaschi from the ItaliaNLP Lab at the Istituto di Linguistica Computazionale (CNR-ILC), Pisa, delivered a talk titled "From LLM Evaluation to Digital Social Reading," in which he examined the interpretability and evolution of neural language models and their growing relevance to linguistic research, outlining key scenarios where NLP can enrich humanistic modes of reading and interpretation. Complementing this, Professor Paul Rayson from Lancaster University showcased the practical applications of digital humanities in geospatial narrative processing, demonstrating how computational tools can map and contextualise historical testimonies across time and space. He further reflected on the future trajectory of the field, underscoring the indispensable role of interdisciplinary collaboration in ensuring the methodological rigour, innovation, and societal impact of digital humanities projects.

The success of LM4DH 2025 would not have been possible without the generous contributions of many exceptional individuals who supported this initiative. First and foremost, we extend our deepest gratitude to the authors who submitted their innovative work, helping to advance the vital intersection of language models and Digital Humanities. We are equally indebted to the members of the Program Committee, whose thoughtful engagement, timely reviews, and incisive feedback were instrumental in shaping the workshop's scholarly quality. Their dedication not only elevated the rigor of accepted submissions but also ensured the program reflected the highest standards of academic excellence and interdisciplinary innovation. Together, they have not only documented the state of the art but have helped define its future.

Organisers of LM4DH 2025

Organising Committee

Isuri Anuradha, Lancaster University, UK
Deshan Sumanathilake, Swansea University, UK
Francesca Frontini, CNR , Italy
Paul Rayson, Lancaster University, UK
Ruslan Mitkov, Lancaster University, UK

Programme Committee

Maram Alharbi, Lancaster University, UK
Salmane Chafik, Mohammed VI Polytechnic University, Morocco
Ignatius Ezeani, Lancaster University, UK
Safia Kanwal, Swansea University, UK
Chamila Liyanage, University of Colombo, Sri Lanka
Laura Noriega Santianez, University of Malaga, Spain
Yael Netzer, Hebrew University, Israel
Saadh Jawwadh, Informatics Institute of Technology, Sri Lanka
Gayanath Chandrasena, University of Helsinki, Finland
Marie Escribe, Universitat Politecnica de Valencia (UPV), Spain
Federico Boschetti, CNR Istituto di, Linguistica Computazionale A. Zampolli, Italy
Voula Giouli, Aristotle University of Thessaloniki, Greece

Table of Contents

HamRaz: A Culture-Based Persian Conversation Dataset for Person-Centered Therapy Using LLM Agents

Mohammad Amin Abbasi[1], **Farnaz Sadat Mirnezami**[2], **Ali Neshati**[2], **Hassan Naderi**[1*]

[1]Department of Computer Engineering, Iran University of Science and Technology, Tehran, Iran
[2]Computer Science, Faculty of Science Ontario Tech University Ontario, Canada

m_abbasi1378@comp.iust.ac.ir
farnazsadat.mirnezami@ontariotechu.net
naderi@iust.ac.ir
ali.neshati@ontariotechu.ca

Abstract

We present HamRaz, a culturally adapted Persian-language dataset for AI-assisted mental health support, grounded in Person-Centered Therapy (PCT). To reflect real-world therapeutic challenges, we combine script-based dialogue with adaptive large language models (LLM) role-playing, capturing the ambiguity and emotional nuance of Persian-speaking clients. We introduce HamRazEval, a dual-framework for assessing conversational and therapeutic quality using General Metrics and specialized psychological relationship measures. Human evaluations show HamRaz outperforms existing baselines in empathy, coherence, and realism. This resource contributes to the Digital Humanities by bridging language, culture, and mental health in underrepresented communities.

1 Introduction

Recent advancements in Large Language Models have significantly expanded the potential of AI in supporting mental health, particularly through simulating therapeutic conversations. These models are capable of producing coherent, context-sensitive, and emotionally resonant responses, making them increasingly useful in preliminary mental health screenings, emotional support systems, and AI-powered conversational agents (Hua et al., 2024; Stade et al., 2024). Despite these achievements, most research in this domain remains focused on English and East Asian contexts, overlooking the cultural and linguistic diversity required for truly inclusive mental health support.

In particular, Persian-speaking communities remain underserved in this space. Cultural norms, language structures, and societal values deeply influence the therapeutic process, affecting both how individuals express emotional distress and how they respond to interventions. Yet, no comprehensive efforts have been made to build culturally relevant, Persian-language datasets or therapeutic AI systems, leaving a significant gap in both research and application.

Moreover, while Cognitive Behavioral Therapy (CBT) (Beck, 1979) has commonly been adopted in AI-based simulations, its dependency on diagnostic precision and structured intervention frameworks poses a challenge for LLMs, which lack clinical reasoning capabilities. In contrast, Person-Centered Therapy (PCT) (Rogers, 1951)—a humanistic approach that emphasizes empathy, non-directiveness, and the client's own voice—aligns more naturally with the conversational and generative strengths of LLMs, making it a more suitable foundation for AI-driven therapeutic dialogues.

Existing approaches to generating psychotherapy datasets, such as Two-Agent Mode (fully LLM-driven simulations) (Zhou et al., 2023) or Script-Based methods (predefined conversation templates), often fall short in realism and complexity. These methods tend to produce unnatural or overly simplistic dialogues, failing to capture the ambiguity, emotional conflict, and indirect expression typical of real therapeutic settings. They also frequently portray therapy as overly effective within a single session, misrepresenting the gradual and nonlinear nature of psychological healing.

To address these issues, we introduce HamRaz, the first Persian-language dataset designed for PCT-based AI therapy. Our approach combines

* Corresponding author.

Proceedings of the First Workshop NLP and Language Models for Digital Humanities associated with RANLP 2025, pages 1–24, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-106-6-001

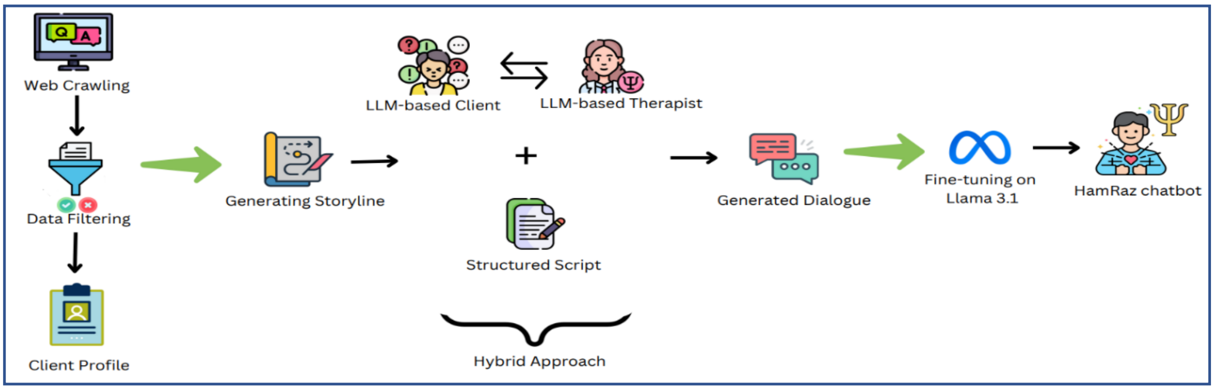


Figure 1: The architecture of the HamRaz simulation framework.

the structural clarity of scripted scenarios with the adaptability of LLM role-playing to generate more realistic and emotionally rich interactions. In addition, we propose HamRazEval, a two-tier evaluation framework that measures both general conversational quality and therapeutic depth using adapted Barrett-Lennard Relationship Inventory metrics (S. Chen, Liao, F., Murphy, D., & Joseph, S. , 2023). This contribution aims to advance culturally sensitive mental health AI by providing a novel resource tailored to Persian-speaking users, while offering generalizable insights into human-aligned therapy simulations.

The overall architecture of the HamRaz simulation framework is illustrated in Figure 1.

2 Related Works

2.1 Dataset for Simulated Counseling and Therapy

Several studies have explored the use of large language models to simulate counselor-client interactions for generating training data in LLM-based therapy models. For instance, Psych8k (Liu et al., 2023) includes 8,187 instruction pairs from 260 counseling sessions. CPsyCounD (Zhang et al., 2024) contains 3,134 multi-turn consultation dialogues generated based on real reports in Chinese contexts. PATIENT-Ψ (Wang et al., 2024) emphasizes LLM-generated patient interactions for mental health training. Moreover, CACTUS (Lee et al., 2024) is a multi-turn dialogue dataset designed to emulate real-life counseling interactions using Cognitive Behavioral Therapy techniques.

In spite of these developments, challenges remain regarding cultural and language adaptation and the integration of Person-Centered therapeutic approaches.

2.2 LLMs for Mental Health Applications

The application of large language models in psychological counseling and mental health support is a growing area of research. several models have been developed with a focus on mental health interventions. Psy-LLM (Lai et al., 2023) integrates pre-trained LLMs with professional psychological content to enhance counseling responses, while ChatCounselor (Liu et al., 2023) is a fine-tuned model based on the Psych8k dataset, designed to improve mental health support. Additionally, SoulSpeak (Zhang & Luo, 2024) incorporates a dual-memory system for long-term context retention, while a hybrid model explored in (Yu & McGuinness, 2024) combines DialoGPT and ChatGPT-3.5 for psychological support. PsyChat (Qiu et al., 2023) and SimPsyBot (Qiu & Lan, 2024) simulate counselor-client interactions, enhancing the realism of therapeutic dialogues.

Despite these advancements, most models lack cultural adaptation for Persian-speaking users. Our work addresses this gap by integrating person-centered therapy principles into a Persian-language LLM-driven therapy framework, enhancing cultural relevance and accessibility.

3 Methodology

3.1 Data Collection

We developed a dataset for Persian person-centered therapy by crawling publicly accessible psychological question-and-answer interactions. The data was sourced from Iranian psychology websites and forums, including ehyacenter[2], moshaverfa , and simiaroom, where individuals from Iran openly post their mental health queries and psychologists provide answers.

[2] ehyacenter.com

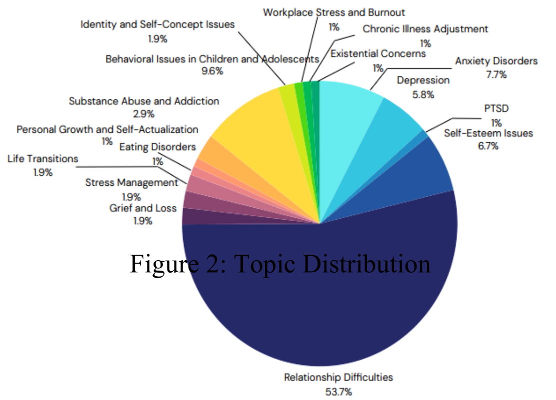


Figure 2: Topic Distribution

Given that all crawled content was already in the public domain and universally accessible, this approach ensures the dataset accurately captures real-world psychological concerns and therapeutic dialogues prevalent in the Iranian cultural setting.

Through this process, we successfully collected 4,000 user-submitted questions, covering a wide range of mental health issues. The dataset's authenticity and cultural relevance were maintained by selecting platforms frequented by Persian-speaking users seeking psychological advice. These questions and problem statements inherently reflected cultural nuances, societal expectations, and region-specific challenges, making them particularly valuable for developing a context-aware person-centered therapy model. You can find a Sample of questions in Appendix B.

3.2 Ensuring Data Privacy

To ensure the privacy and confidentiality of the dataset, we undertook a meticulous process of reviewing and preprocessing each text entry. This step was essential to mitigate the risk of exposing sensitive or personally identifiable information and to prepare the data for subsequent analysis in an ethical and responsible manner. We carefully anonymized all data entries to remove or obscure any information that could directly or indirectly identify individuals. Specific measures included:

- Personal names were replaced with generic labels such as "Person A" or "Participant X."
- Specific geographic details were replaced with broader terms like "residential area" or "large city.
- Exact dates were generalized to months or even years, depending on the context.
- Any unique identifiers, such as phone numbers, email addresses, or social media handles, were completely removed.
- Contextual Adjustments: Context-specific identifiers, such as job titles ("CEO of X Company") or project names, were anonymized to generic terms like "senior manager" or "project leader."

3.3 Topic Classification and Filtering

To refine the dataset and ensure alignment with Person-Centered Therapy (PCT) principles, we employed GPT-4o (OpenAI, 2023) with a structured prompt to classify the 4,000 collected questions into 16 predefined mental health categories. These categories were selected based on their therapeutic relevance to PCT, covering a wide range of psychological concerns such as anxiety disorders, depression, PTSD, self-esteem issues, relationship difficulties, grief and loss, stress management, life transitions, personal growth, eating disorders, substance abuse, behavioral issues in youth, identity and self-concept challenges, workplace stress, chronic illness adjustment, and existential concerns.

GPT-4o was prompted to follow a binary classification process, assessing whether each question matched one or more of the predefined topics. Non-relevant questions, such as those focused on medical or technical advice, were automatically excluded. The full prompt and topic definitions are provided in Appendix A. This process ensured that the dataset remained focused, culturally relevant, and suitable for person-centered therapy applications. The distribution of topics is shown in Figure 2.

3.4 Client Profiling

Once the dataset was refined to include only therapeutically relevant questions, the next step involved profiling each user submission to extract key psychological insights. This profiling process aimed to categorize emotional themes, psychological issues, and contextual factors, enabling a deeper understanding of user concerns and facilitating more nuanced therapy simulations.

To achieve this, we utilized GPT-4o to analyze each message and generate a structured profile capturing: (1) Emotional Themes (e.g., fear, sadness, insecurity), (2) Key Psychological Issues (e.g., trust problems, abandonment, low self-esteem), (3) Past Experiences (e.g., childhood trauma, family

3

Client Question:
"Hello,
I struggle with assertiveness in tasks that require interaction and collaboration with others. Sometimes, being too available, compromising too much, accepting others' mistakes, and constantly giving in make me feel exhausted and worthless. In response, I become stubborn as a way to make those around me realize how they've taken advantage of my flexibility.
I want to understand the difference between stubbornness and assertiveness. Also, I need guidance on how to recognize when I should be assertive. I naturally tend to be soft and flexible, but I feel that it's not always the right approach."

Client Profile:
{
"emotional_themes": ["frustration", "insecurity", "exhaustion", "confusion", "desire for assertiveness"],
"key_psychological_issues": ["lack of assertiveness", "fear of being taken advantage of", "self-worth issues", "difficulty in setting boundaries"],
"past_experiences": ["experiences of being overly accommodating in relationships", "feelings of being undervalued or unappreciated"],
"patterns_and_behaviors": ["over-accommodating behavior", "difficulty in asserting needs", "oscillation between flexibility and stubbornness"],
"desired_outcome": "guidance on distinguishing between assertiveness and stubbornness, and strategies for being more assertive",
"contextual_factors": ["the user is likely in a collaborative work environment", "the user may have a tendency to prioritize others' needs over their own"]
}

Figure 3: Example of a client question and its generated profile.

conflict), (4) Patterns and Behaviors (e.g., overthinking, avoidance, reassurance seeking), (5) Desired Outcomes (e.g., coping strategies, emotional relief, validation), and (6) Contextual Factors (e.g., age, family dynamics, workplace or societal stressors).

The profiling process enriched the dataset by categorizing user concerns into key psychological themes, allowing the LLM-driven therapy model to generate more tailored and empathetic responses. Using a structured prompt-based approach (detailed in Appendix A), the model analyzed, categorized, and extracted relevant insights from user messages. This method ensured that each concern was not only thematically classified but also contextually understood, enhancing the dataset's psychological depth and enabling more realistic therapy simulations. An example of a generated client profile is in Figure 3.

Across 3,400 profiled messages, multiple emotional themes were typically present in each query. The most frequent themes were frustration (2,914), sadness (2,124), anxiety (1,306), fear (1,228), insecurity (1,224), and confusion (1,150). Less frequent but notable emotions included concern (678), helplessness (415), disappointment (266), and guilt (216). This distribution highlights that relational strain, self-worth challenges, and anxiety-related struggles are among the most prominent concerns expressed by Persian-speaking clients in the dataset.

3.5 Adding Complexity to Client Statements

To improve the realism of client interactions, we introduced complexity in half of the user profiles by modifying their statements to reflect unclear, indirect, or conflicting expressions. This step aimed to simulate real-world therapy challenges, where clients may struggle with articulating emotions, expressing mixed feelings, or avoiding sensitive topics.

Using GPT-4o, we applied a structured complexity-enhancement prompt (detailed in Appendix A) to analyze user profiles and assign relevant characteristics to each case. These characteristics were categorized into Unclear Statements, Indirect Statements, Conflicting Statements, Mixed Emotions, Avoidant or Defensive Responses, Cultural Ambiguities. Each user's psychological profile and contextual factors were used to assign one or more complexity traits, ensuring that therapy dialogues remained nuanced and reflective of real client experiences.

3.6 Simulating Psychotherapy Session

We structured the simulated therapy sessions using a five-stage framework based on Person-Centered Therapy (PCT) principles. These stages were designed to mirror real-life therapy sessions, guiding the interaction between the client and psychologist in a natural, progressive manner. By systematically defining each stage, we ensured that LLM-driven therapy simulations maintained coherence, therapeutic effectiveness, and emotional realism.

The simulated sessions followed five PCT-based stages: (1) building rapport and trust, (2) active empathetic listening, (3) encouraging self-exploration, (4) supporting growth and change, and (5) reviewing and closing.

To ensure realism in AI-driven therapy simulations, we used a structured prompt-based process that mapped each client profile onto five therapy stages. At each stage, the model selected emotions, expressions, and behaviors from predefined options, creating coherent and fluid dialogues that captured emotional ambiguity and nuanced struggles, resulting in consistent and realistic therapeutic interactions. The structured prompt design is detailed in Appendix A, and an example is provided in Appendix B.

4

3.7 Generating a Storyline

We improved the clarity and authenticity of therapy dialogues by first generating a narrative storyline before converting it into dialogue. This intermediate step ensured that therapy interactions felt natural and human-like, rather than fragmented LLM-generated responses. Using GPT-4o, we structured therapy sessions into a five-stage narrative, integrating the client's emotional themes, psychological issues, and past experiences. This approach provided a logical emotional progression, aligning client interactions with realistic therapeutic dynamics.

A storyline-first approach ensured a natural emotional arc, avoiding disjointed exchanges. It also allowed for richer context, incorporating body language, pauses, tone shifts, and facial expressions—elements often missing in direct LLM-generated dialogue. Additionally, by shaping the session as a narrative before dialogue conversion, we improved control over the conversation's flow, leading to more lifelike therapy simulations in Persian. The full prompt and an example of this process are detailed in Appendix A and B.

3.8 Converting the Storyline into Dialogue Format

Our methodology aligns with findings from CACTUS research (Lee et al., 2024). This study compared two methods for generating psychotherapy dialogues: (1) Two-Agent Mode (Zhou et al., 2023), where separate models assume the roles of the client and therapist, and (2) Script Mode, where dialogue is generated from a structured script with predefined client and counselor information. (see Appendix A and B)

3.8.1 Generating Dialogue using Script Mode

To generate structured therapy dialogues, we transformed the narrative-based therapy sessions into a scripted dialogue format between a therapist and a client. This approach ensured that conversations maintained a natural flow and therapeutic coherence, reflecting the dynamics of real-world psychotherapy sessions.

Results in CACTUS research demonstrated that scripted dialogue generation produces more natural and well-constructed conversations compared to Two-Agent Mode. Inspired by these findings, we adopted a scripted approach, ensuring that the therapist-client interactions were coherent, psychologically grounded, and contextually aligned with Person-Centered Therapy principles.

To guide this transformation, we utilized a detailed prompt-based methodology, incorporating client profiles, emotional themes, past experiences, and cognitive patterns into the dialogue generation process. This approach enabled us to create a high-quality, culturally nuanced dataset of Persian therapy conversations, allowing for more effective LLM-driven psychotherapy sessions simulations.

3.8.2 Role-Playing LLM-to-LLM Interactions

We introduce a novel hybrid approach to improve the dynamic quality, conversational depth, and psychological validity of therapy dialogues. This approach integrates script mode with two-agent mode, combining the structural benefits of pre-scripted dialogues with the flexibility of agent-based interactions. It enables two LLM-based agents to simulate a therapist-client conversation, refining the dialogue dynamically while maintaining adherence to Person-Centered Therapy (PCT) principles.

In this framework, one agent assumes the role of the therapist, guided by a system prompt designed to enforce PCT techniques such as reflective listening, open-ended questioning, and non-directive engagement. The second agent assumes the role of the client, ensuring responses align with a predefined user profile, including emotional states, past experiences, and psychological patterns. Unlike conventional two-agent systems, where LLMs interact freely and often generate inconsistencies, our approach structures these interactions around a pre-scripted baseline dialogue, which is iteratively refined at each conversational turn.

3.8.3 A Hybrid Approach: Combining Script Mode and Two-Agent Mode

While pre-scripted dialogues offer consistency, they often lack the spontaneity and adaptability essential for realistic therapeutic interactions, frequently failing to adjust to subtle shifts in client emotions and the natural flow of conversation. To overcome these limitations, we introduce a hybrid approach that synergizes script mode with two-agent interactions. This method utilizes pre-scripted dialogues as a foundational framework, which is then dynamically refined by LLM agents at each conversational turn, ensuring dialogues remain structured yet flexible and capable of

5

incorporating real-time adjustments without sacrificing coherence.

Our initial scripted dialogues, though consistent, tended to be mechanical and lacked the depth and natural conversational flow characteristic of genuine therapy. Emotional transitions were often abrupt or unclear, and some therapist responses, while aiming for PCT principles, inadvertently became overly directive. Furthermore, the absence of implicit non-verbal cues, such as natural pauses or subtle tone shifts that LLM agents can emulate, diminished the realism of these simulations.

To address these challenges, we implemented a dynamic iterative refinement loop driven by the two agents:

1. **Foundation:** Each conversational exchange begins with a segment from the pre-generated storyline/script. This provides a thematic guide and ensures overall narrative consistency.
2. **Therapist Agent's Turn:** The therapist agent receives the full conversation history up to the current point, along with the prescripted response designated for its current turn. Crucially, it also considers the client's actual previous utterance. Based on this immediate context, the therapist agent dynamically adapts and refines its prescripted line. The goal is to ensure its response is not only coherent with the script's intention but also genuinely empathetic, context-aware, and strictly adherent to Person-Centered Therapy (PCT) principles, such as reflective listening and emotional validation, in direct response to the client's latest statement.
3. **Client Agent's Turn:** The client agent receives the conversation history, including the therapist's dynamically adapted response. Informed by its pre-defined profile (emotional themes, psychological issues, past experiences, and complexity characteristics), the client agent generates a natural and contextually appropriate response. This allows for the incorporation of realistic client behaviors such as hesitation, defensiveness, emotional ambiguity, or cognitive dissonance, mirroring the complexities of genuine psychotherapy sessions.
4. **Iteration:** This process of receiving a scripted guideline, dynamically adapting it (therapist) or responding naturally (client) based on the immediate prior turn and profile, repeats for each subsequent turn in the dialogue.

This iterative refinement loop is the core of our hybrid method. It maintains dialogue coherence by anchoring the conversation to the storyline, while significantly enhancing flexibility and realism through agent-based, turn-by-turn adjustments. The hybrid method demonstrably improved the depth and quality of the generated therapy dialogues. Interactions became longer, more nuanced, and more closely mirrored the ebb and flow of real therapeutic conversations. The therapist agent showed improved adherence to PCT principles, and each conversational turn remained contextually grounded, mitigating the risk of incoherent or repetitive exchanges.

To ensure that the generated dialogues are both authentic and therapeutically faithful, we adopted a multi-step generation pipeline. Each step was chosen with a specific rationale. Client profiling anchors the process in real, culturally grounded concerns rather than purely synthetic prompts. Adding complexity introduces ambiguity and indirectness, which are hallmarks of real therapy conversations in Persian cultural contexts. The storyline-first step prevents fragmented exchanges by ensuring a natural emotional arc aligned with the five PCT stages. Converting storylines into script format maintains structural coherence and preserves therapeutic fidelity, while the hybrid agent-based refinement injects spontaneity, variability, and context-sensitive empathy into the dialogue. Together, these steps balance creativity with therapeutic fidelity: the dialogues remain flexible and human-like while consistently adhering to Person-Centered Therapy principles of empathy, non-directiveness, and unconditional positive regard. Following this process, we constructed the HamRaz dataset. HamRaz is a Persian term that reflects the idea of a confidant—someone with whom individuals feel safe sharing their secrets and receiving empathy. This name embodies the dataset's focus on fostering trust and authentic, client-centered therapeutic dialogues.

A full summary of dataset statistics, including dialogue counts, average session length, category distribution, and emotional theme frequencies, is provided in Appendix C.

6

4 Experiment

Evaluating the effectiveness of psychotherapy dialogues requires assessing both the counselor's conversational abilities and the psychological impact on the client. Traditional evaluation approaches, such as automatic or single-turn assessments, fail to capture the nuanced dynamics of therapeutic interactions, making them unsuitable for assessing the quality of person-centered therapy. To address these limitations, we propose a comprehensive evaluation framework tailored for LLM-driven Persian therapy simulations.

4.1 HamRazEval Framework

We introduce a two-tier evaluation framework designed to assess the quality of therapy dialogues based on conversational coherence, engagement, and therapeutic effectiveness. Our framework consists of (1) a general conversational evaluation (GeneralEval) to assess the dialogue's structural integrity and fluency, and (2) the Barrett-Lennard Relationship Inventory (BLRI) assessment to evaluate the psychologist's ability to foster a facilitative therapeutic environment. This dual assessment ensures a holistic evaluation of LLM-generated psychotherapy sessions.

4.1.1 General Conversational Evaluation

We assess the overall conversational quality of therapy sessions using six core metrics:

- **Coherence:** The logical consistency and relevance of responses within the dialogue.
- **Engagement:** The level of active participation and mutual responsiveness between the psychologist and the client.
- **Fluency:** The grammatical correctness and naturalness of language used.
- **Diversity:** The richness of vocabulary and avoidance of repetitive phrasing.
- **Humanness:** The authenticity of the dialogue in resembling real human interactions.
- **Collaboration & Balance:** The extent to which both participants contribute meaningfully to the conversation.

Each metric is rated on a Likert scale from 1 to 10, where higher scores indicate superior conversational quality. The evaluation is performed on multiple dialogues to ensure robustness and reliability.

4.1.2 Therapeutic Relationship Assessment

To measure the effectiveness of the psychologist's approach in fostering a supportive person-centered interaction, we employ a 12-item adaptation of the Barrett-Lennard Relationship Inventory (mini-BLRI) (Chen, 2023). This assessment evaluates the psychologist's empathy, unconditional positive regard, and congruence. The detailed 12 criteria used for evaluation are provided in Appendix C.

Each criterion is rated on a scale from -3 to +3, where +3 signifies a strong presence of the attribute, and -3 indicates its absence. A higher cumulative BLRI score reflects a more effective therapeutic relationship.

4.2 Experiment Setup

For evaluation, we randomly sampled 100 user-submitted questions from the HamRaz dataset, while the remaining data was reserved for training and fine-tuning. These 100 evaluation samples were then compared against Script Mode and Two-Agent Mode baselines to assess therapeutic realism and conversational quality.

Since no multi-turn Persian-language psychotherapy dataset exists—particularly one aligned with Person-Centered Therapy (PCT)—we developed two baseline datasets using the Script Mode and Two-Agent Mode methodologies. These datasets were compared with the HamRaz dataset to evaluate improvements in therapeutic realism and conversational quality.

We fine-tuned LLaMA 3.1 8B (Dubey et al., 2024) on each dataset to serve as the therapist agent in evaluation settings. This model was chosen for its demonstrated strength in psychological reasoning, as reflected in the PsychoLex benchmark (Abbasi et al., 2024).

While we initially aimed to collaborate with a licensed clinical psychologist for expert evaluation, logistical constraints prevented us from securing external clinical review. Consequently, we, the authors, conducted the evaluation manually using structured criteria grounded in prior works. Our academic backgrounds in natural language processing and interdisciplinary mental health research allowed us to apply the metrics with informed judgment.

Specifically, we used a two-tier framework introduced in Section 4.1: (1) the GeneralEval rubric assessing dialogue coherence, fluency, humanness, and balance, and (2) a 12-item adaptation of the Barrett-Lennard Relationship Inventory (BLRI) to

7

Method	BLRI Score	General Score
Script mode	1.84	8.06
HamRaz	**2.85**	**9.31**

Table 1: Comparison of HamRaz and Script Mode on BLRI and General Scores.

Method	BLRI Score	General Score
Two-Agent mode	1.58	8.03
HamRaz	**2.85**	**9.55**

Table 2: Comparison of HamRaz and Two-Agent Mode on BLRI and General Scores.

assess empathy, congruence, and positive regard in therapist responses. All dialogues were rated independently and cross-checked for consistency.

The results demonstrate that HamRaz outperforms both baselines in all core dimensions. Compared to Script Mode and Two-Agent Mode, HamRaz achieved significantly higher scores in both BLRI and GeneralEval, indicating superior therapeutic alignment and conversational authenticity. These findings support the effectiveness of our hybrid generation approach and reinforce the value of culturally grounded datasets in improving LLM-based mental health support.

Although the absence of external expert judgment is a limitation, the use of structured evaluation criteria and the consistency of rating across multiple sessions provide a replicable and transparent foundation for future benchmarking.

4.3 Results

The evaluation results clearly demonstrate that the HamRaz dataset outperforms both Script Mode and Two-Agent Mode across all key dimensions of conversational and therapeutic quality. As shown in Table 1, when compared with Script Mode, HamRaz achieved a markedly higher BLRI score and GeneralEval score, indicating more empathetic, engaging, and human-like interactions. In comparison with Two-Agent Mode, HamRaz again achieved notably higher scores on both the BLRI and GeneralEval assessments, reflecting substantial gains in conversational coherence and therapeutic alignment. The detailed comparison is presented in Table 2. The elevated BLRI scores highlight HamRaz's effectiveness in capturing essential components of Person-Centered Therapy (PCT), such as empathy, unconditional positive regard, and congruence. Similarly, the consistently higher GeneralEval metrics affirm that HamRaz fosters more coherent, fluent, and natural interactions than either baseline method.

These results validate the impact of our hybrid generation approach and culturally grounded methodology. In particular, they demonstrate that dataset quality and design—especially when tailored to the linguistic and cultural characteristics of the target community—play a more critical role in therapeutic dialogue success than model-specific factors alone.

Overall, HamRaz sets a new benchmark for Persian-language psychotherapy simulations, offering a realistic, empathetic, and person-centered foundation for future research in LLM-driven mental health support.

5 Conclusions

We introduced HamRaz, a culturally adapted Persian-language dataset for Person-Centered Therapy, combining script-based dialogues with LLM role-playing to simulate realistic and emotionally complex interactions. Using a structured human evaluation framework (HamRazEval), we showed that HamRaz outperforms existing methods in both conversational quality and therapeutic alignment. This work contributes a novel resource at the intersection of language technologies and digital mental health, supporting future research in culturally grounded AI applications for underrepresented communities.

Limitations

While HamRaz represents a meaningful step forward in developing culturally aligned LLM-driven mental health resources, particularly within the framework of Person-Centered Therapy (PCT) for Persian speakers, it is not without limitations. First, the evaluation process was conducted by the authors rather than clinical experts due to resource and logistical constraints. Although structured rubrics were applied and prior research supports the validity of such approaches, future work would benefit from incorporating assessments by licensed mental health professionals to enhance credibility.

Second, the dataset's focus on PCT within a Persian sociocultural context may reduce its applicability to other therapeutic approaches or linguistic communities. Finally, since parts of the data rely on web-based content, potential sampling bias from

8

the source platforms may affect the diversity and representativeness of client profiles.

Ethical Considerations

The development of HamRaz, a culturally adapted Persian-language mental health dataset, necessitates a strong ethical foundation to ensure privacy, cultural sensitivity, and responsible AI use. Given the delicate nature of psychological conversations, we have taken proactive measures to address key ethical concerns.

First, privacy and data protection were prioritized by sourcing data exclusively from publicly available forums while ensuring complete anonymization to safeguard user identities. Additionally, cultural sensitivity was embedded into the dataset design, ensuring that conversations reflect the nuances of Persian-speaking communities without reinforcing biases or stereotypes.

Despite the benefits of AI-driven therapy simulations, HamRaz is not a substitute for professional psychological care. Large language models (LLMs) lack the clinical judgment required for mental health diagnosis and crisis intervention. Therefore, it is crucial that AI-generated interactions remain supportive rather than directive, with clear disclaimers advising users to seek professional assistance when necessary.

Furthermore, to mitigate biases inherent in AI models, we implemented structured data filtering and validation mechanisms, ensuring diverse representation while avoiding reinforcement of harmful narratives. Lastly, responsible AI deployment is essential—HamRaz should only be used in ethically guided research and applications, with safeguards to assess the impact of AI-generated counseling.

References

Abbasi, M. A., Mirnezami, F. S., & Naderi, H. (2024, November). PsychoLex: Unveiling the Psychological Mind of Large Language Models.*Proceedings of the 1st Workshop on NLP for Science (NLP4Science)* Miami, FL, USA.

Chen, S., Liao, F., Murphy, D., & Joseph, S. . (2023). Development and validation of a 12-item version of the Barrett-Lennard Relationship Inventory (B-L RI:mini) using item response theory. . *Current Psychology*, *42(13)*(10566-10580). https://doi.org/https://doi.org/10.1007/s12144-021-02348-5

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A. S., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A.,…Zhao, Z. (2024). The Llama 3 Herd of Models. *ArXiv*, *abs/2407.21783*.

Hua, Y., Liu, F., Yang, K., Li, Z., Na, H., Sheu, Y.-h., Zhou, P., Moran, L. V., Ananiadou, S., & Beam, A. (2024). Large Language Models in Mental Health Care: a Scoping Review. *ArXiv*, *abs/2401.02984*.

Lai, T., Shi, Y., Du, Z., Wu, J., Fu, K., Dou, Y., & Wang, Z. (2023). Psy-LLM: Scaling up Global Mental Health Psychological Services with AI-based Large Language Models. *ArXiv*, *abs/2307.11991*.

Lee, S., Kim, S., Kim, M., Kang, D., Yang, D., Kim, H., Kang, M., Jung, D., Kim, M. H., Lee, S., Chung, K.-M., Yu, Y., Lee, D., & Yeo, J. (2024). Cactus: Towards Psychological Counseling Conversations using Cognitive Behavioral Theory. *ArXiv*, *abs/2407.03103*.

Liu, J. M., Li, D., Cao, H., Ren, T., Liao, Z., & Wu, J. (2023). ChatCounselor: A Large Language Models for Mental Health Support. *ArXiv*, *abs/2309.15461*.

OpenAI. (2023). *Gpt-4 technical report*. https://arxiv.org/abs/2303.08774

Qiu, H., & Lan, Z. (2024). Interactive Agents: Simulating Counselor-Client Psychological Counseling via Role-Playing LLM-to-LLM Interactions. *ArXiv*, *abs/2408.15787*.

Qiu, H., Li, A., Ma, L., & Lan, Z. (2023). PsyChat: A Client-Centric Dialogue System for Mental Health Support. *2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 2979-2984.

Stade, E. C., Stirman, S. W., Ungar, L. H., & et al. (2024). Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *npj Mental Health Research*, *3*, 12. https://doi.org/10.1038/s44184-024-00056-z

Wang, R., Milani, S., Chiu, J. C., Zhi, J., Eack, S. M., Labrum, T., Murphy, S. M., Jones, N., Hardy, K., Shen, H., Fang, F., & Chen, Z. (2024). PATIENT-?: Using

9

Large Language Models to Simulate Patients for Training Mental Health Professionals. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Yu, H. Q., & McGuinness, S. (2024). An experimental study of integrating fine-tuned large language models and prompts for enhancing mental health support chatbot system. *Journal of Medical Artificial Intelligence*.

Zhang, C., Li, R., Tan, M., Yang, M., Zhu, J., Yang, D., Zhao, J., Ye, G., Li, C., & Hu, X. (2024). CPsyCoun: A Report-based Multi-turn Dialogue Reconstruction and Evaluation Framework for Chinese Psychological Counseling. *ArXiv*, *abs/2405.16433*.

Zhang, X., & Luo, Z. (2024). Advancing Conversational Psychotherapy: Integrating Privacy, Dual-Memory, and Domain Expertise with Large Language Models. *ArXiv*, *abs/2412.02987*.

Zhou, X., Zhu, H., Mathur, L., Zhang, R., Yu, H., Qi, Z., Morency, L.-P., Bisk, Y., Fried, D., Neubig, G., & Sap, M. (2023). SOTOPIA: Interactive Evaluation for Social Intelligence in Language Agents. *ArXiv*, *abs/2310.11667*.

10

A Prompts

The prompt used for filtering psychological questions is shown in Figure 4. The process for analyzing user input and constructing a client profile is illustrated in Figure 5. Adding complexity to the user profile is detailed in Figure 6, while Figure 7 defines the stages of a psychotherapy session. The simulation of the therapy session flow is depicted in Figure 8, and the creation of a storyline for the session is presented in Figure 9. A structured approach is used to generate a therapy session based on a script, as shown in Figure 10. Additionally, separate prompts are designed for the client role (LLM) (Figure 11) and the therapist role (Figure 12), ensuring an interactive and dynamic session. Furthermore, prompts are developed to generate a realistic psychotherapy session dialogue between a therapist and a client (LLM-to-LLM) (Figure 13). Figure 14 presents the Client Agent Evaluation, which assesses the client's engagement, response coherence, and overall interaction quality. Figure 15 illustrates the Psychologist Evaluation, focusing on the therapist's interventions, empathy, and adherence to therapeutic techniques.

11

You are an expert assistant tasked with filtering psychological questions or problems to determine if person-centered therapy (PCT) could be a suitable approach for addressing the issues presented. Use the following comprehensive list of issues where PCT is applicable as your primary guideline:

- Anxiety Disorders
- Depression
- Post-Traumatic Stress Disorder (PTSD)
- Self-Esteem Issues
- Relationship Difficulties
- Grief and Loss
- Stress Management
- Life Transitions
- Personal Growth and Self-Actualization
- Eating Disorders
- Substance Abuse and Addiction
- Behavioral Issues in Children and Adolescents
- Identity and Self-Concept Issues
- Workplace Stress and Burnout
- Chronic Illness Adjustment
- Existential Concerns

For each question or problem:

1. Analyze the Text: Identify whether the problem aligns with one or more issues on the list above.

2. Evaluate for Applicability: Determine if the problem involves emotional distress, relational challenges, self-awareness, or difficulties requiring empathy, acceptance, and non-directive support. This includes problems indirectly linked to the primary caregiver's emotional state or difficulties in managing the situation (e.g., stress or burnout from caregiving).

3. Output Format: Return only binary decision: "Yes" (PCT is applicable) or "No" (PCT is not applicable).

Additional Considerations:

- Problems focusing exclusively on technical, medical, or skill-based solutions (e.g., speech therapy, behavior modification) is not be suitable for PCT.

- Respond only with "Yes" or "No."

Figure 4: Prompt for filtering psychological questions.

Please analyze the following user's message and identify the key emotional themes, underlying psychological issues, and any significant past experiences mentioned. Focus on the emotional states the user is experiencing, the core problems they are facing, and any possible connections to their past or present relationships. Provide a summary of the key themes and insights from the user's input.

Please include the following:

- Emotional Themes: What emotions is the user expressing? Are they experiencing feelings of fear, sadness, insecurity, frustration, etc.?
- Key Psychological Issues: What core issues are being discussed (e.g., anxiety, abandonment, trust issues, rumination)?
- Past Experiences: What past events or experiences are being referenced (e.g., childhood trauma, parental conflict, abandonment)?
- Patterns and Behaviors: Are there any patterns or behaviors in the user's thinking or relationships (e.g., overthinking, seeking signs of rejection)?
- Desired Outcome: What does the user seem to want or need from this conversation (e.g., relief from negative thoughts, reassurance in relationships, coping strategies)?
- Contextual Factors: Are there any other contextual details, such as age, family dynamics, or current life situation, that could be relevant to understanding the user's emotional state?

Output Format should be in JSON like this:

```
{

  "emotional_themes": ["list of emotions the user is expressing"],

  "key_psychological_issues": ["list of core psychological issues or concerns"],

  "past_experiences": ["list of past events or experiences mentioned"],

  "patterns_and_behaviors": ["list of patterns or recurring behaviors in the user's thinking or relationships"],

  "desired_outcome": "what the user seems to want or need from this conversation",

  "contextual_factors": ["list of other relevant contextual details, such as age, family dynamics, or current life situation, ..."]

}
```

User's message:

Figure 5: Process for analyzing user input and constructing a client profile.

12

You are a skilled therapy simulator and scenario creator. Your task is to add complexity to a person-centered therapy session by selecting and integrating relevant characteristics of unclear, indirect, or conflicting client statements.

You are provided with a user profile that contains information about the client's emotional themes, psychological issues, past experiences, patterns and behaviors, desired outcomes, and contextual factors. Based on this profile, use your creativity to choose characteristics that will make the client's statements complex, realistic, and reflective of their unique struggles.

Instructions:

1. Analyze the User Profile:

- Read the emotional themes, key psychological issues, past experiences, and other factors to understand the client's inner world.

- Consider how their background might influence their communication style (e.g., vague, defensive, or contradictory).

2. Select Relevant Characteristics:

- Choose from the following characteristics to add complexity to the client's dialogue:

Unclear Statements

- Lack of specificity in emotions or concerns.
- Hesitation or uncertainty in language (e.g., "I think," "maybe").
- Tendency to avoid direct confrontation of deeper feelings.
- General or vague language (e.g., "something feels wrong").
- Minimal elaboration or detail about the issue.

Indirect Statements

- 6. Hinting at issues without explicitly naming them.
- 7. Skirting around deeper topics or providing surface-level answers.
- 8. Use of dismissive or minimizing language (e.g., "It's not a big deal").
- 9. Cultural or societal pressure to suppress emotions.
- 10. Reluctance to express emotions due to fear of judgment.

Conflicting Statements

- 11. Emotional tension between opposing desires or perspectives.
- 12. Oscillation between positive and negative emotions about the same issue.
- 13. Statements revealing inner conflict or ambivalence.
- 14. Expressions of being stuck or torn (e.g., "I want to leave, but I can't").
- 15. Inconsistent or contradictory language.

Mixed Emotions

- 16. Emotional layering or overlap (e.g., anger and sadness).
- 17. Expression of both positive and negative emotions simultaneously.
- 18. Difficulty in resolving or prioritizing emotions.
- 19. Contradictory feelings about the same event or situation.
- 20. Complexity in emotional processing (e.g., relief mixed with guilt).

Avoidant or Defensive Statements

- 21. Deflection or shifting focus to avoid discussing uncomfortable topics.
- 22. Use of sarcasm, humor, or denial to downplay issues.
- 23. Defensive language (e.g., "Why does it matter?" or "It's fine").
- 24. Dismissive behavior toward their own emotions or concerns.
- 25. Resistance to deeper emotional exploration.

Context-Specific Ambiguities (Cultural Dynamics)

- 26. Tension between personal desires and societal/familial expectations.
- 27. Ambiguity about the cause of distress (internal vs. external).
- 28. Fear of judgment, shame, or loss of reputation.
- 29. Desire to meet cultural or family expectations at the expense of personal needs.
- 30. Hesitation to express "taboo" feelings or emotions due to cultural stigma.

3. Match Characteristics to Profile:

- Use the client's emotional themes, psychological issues, and contextual factors to justify why certain characteristics are relevant.

- Combine multiple characteristics, if appropriate, to reflect the complexity of the client's struggles.

Expected Output:

```
{
  "selected_characteristics": [
    "characteristic_1",
    "characteristic_2",
    "characteristic_3"
  ]
```

Figure 6: Prompt for adding complexity to client profile

13

1. Initial Meeting and Building Rapport:

   - At the start of the session, the therapist gathers basic information and creates a safe and supportive atmosphere for the client.

   - The goal at this stage is to establish trust and make the client feel comfortable sharing their thoughts and feelings.

2. Active and Empathetic Listening:

   - The therapist listens attentively to the client, focusing on understanding their emotions and inner needs rather than judging or giving direct advice.

   - Techniques such as reflecting feelings (rephrasing or interpreting the client's emotions) are used to ensure the client feels heard and understood.

3. Encouraging Self-Exploration and Open Expression:

   - Through open-ended questions and gentle guidance, the therapist encourages the client to explore their emotions and thoughts.

   - Emphasis is placed on the client taking responsibility for their life and decisions while tuning into their feelings and needs.

4. Supporting Growth and Change:

   - As the client gains a better understanding of themselves, the therapist supports them in the process of change and personal growth.

   - The goal is for the client to identify and address negative or limiting patterns and work toward positive transformation.

5. Reviewing and Closing the Session:

   - At the end of the session, the therapist and client reflect on the progress made and identify any challenges or points to address in future sessions.

   - The therapist may offer key insights or suggest exercises for further self-exploration.

Figure 7: Defining the stages of a psychotherapy session.

You are an advanced reasoning assistant designed to simulate the flow of a humanistic therapy session. Your goal is to use the user profile and the options for each therapy stage to logically select one or more events/feelings for each stage of the session.

The general steps in such a session can be outlined as follows:

In this approach, the focus is on fostering an empathetic, non-judgmental relationship to help the client feel secure and supported. It aims to empower the client to enhance their self-awareness, personal growth, and overall quality of life.

User Profile:

- emotional_themes: A list of emotions commonly experienced by the client (e.g., sadness, fear).
- key_psychological_issues: Core issues the client faces (e.g., anxiety, trust issues).
- past_experiences: Important events from the client’s past that influence their current emotions and behaviors.
- patterns_and_behaviors: Repeated behaviors or thought patterns observed in the client.
- desired_outcome: What the client hopes to achieve during therapy.
- contextual_factors: Any other relevant details (e.g., age, current life situation).

Options for Each Therapy Stage:

1. Initial Meeting and Building Rapport:

- Comfort and calmness
- Anxiety and tension
- Trust and confidence
- Doubt or suspicion
- Wanting and readiness to talk about issues
- Resistance, secrecy, or silence

2. Active and Empathetic Listening:

- Trust and confidence
- Doubt or suspicion
- Wanting and readiness to talk about issues
- Resistance, secrecy, or silence
- Deep and frank sharing
- Refusing to get into sensitive topics
- Freely expressing sadness, shame, anger, etc.
- Denial, trivializing, or running away from feelings
- Crying, expressing anger, feeling calm after venting
- Shame, embarrassment, fear of expressing emotion
- Client remembers relevant memories from before
- Client recalls memories with the help of psychologist's questions

3. Encouraging Self-Exploration and Open Expression:

- Deep and frank sharing
- Refusing to get into sensitive topics
- Freely expressing sadness, shame, anger, etc.
- Denial, trivializing, or running away from feelings
- Crying, expressing anger, feeling calm after venting
- Shame, embarrassment, fear of expressing emotion
- Feeling of hope or relief from new understanding
- Worry, fear, or denial about discovered truths
- Desire to explore feelings and thoughts
- Doubt, avoidance, or mental resistance in facing facts
- Client remembers relevant memories from before
- Client recalls memories with the help of psychologist's questions

14

4. Supporting Growth and Change:

- Feeling of hope or relief from new understanding
- Worry, fear, or denial about discovered truths
- Eagerness to change and improve
- Feeling of impasse, surrendering to problems
- Calm after processing emotions
- Remaining anger, sadness, or unresolved grief
- Feeling empowered to take action and change
- Helplessness or belief that change is impossible

5. Reviewing and Closing the Session:

- Feeling of hope or relief from new understanding
- Worry, fear, or denial about discovered truths
- Eagerness to change and improve
- Feeling of impasse, surrendering to problems
- Calm after processing emotions
- Remaining anger, sadness, or unresolved grief
- Feeling empowered to take action and change
- Helplessness or belief that change is impossible
- Achieving insight or finding a path
- Stuck in doubts or not making progress
- Desire to explore feelings and thoughts
- Doubt, avoidance, or mental resistance in facing facts

Task:

Using the provided user profile, logically simulate a flow of the therapy session by selecting the most appropriate options for each stage. Ensure the selections align with the client's emotional themes, psychological issues, and desired outcomes. Multiple options can be chosen for a single stage if necessary. It is not necessary to have a good ending.

Output Format:

Return the results in the following JSON format:

```
{"stage_1": ["selected_option_1", "selected_option_2", ...],

  "stage_2": ["selected_option_1", "selected_option_2", ...],

  "stage_3": ["selected_option_1", "selected_option_2", ...],

  "stage_4": ["selected_option_1", "selected_option_2", ...],

  "stage_5": ["selected_option_1", "selected_option_2", ...]}
```

User Profile:

Figure 8: Simulating the flow of a therapy session.

You are a creative writing assistant and therapist simulator. Your task is to create a storyline of a person-centered therapy session in PERSIAN. The client is Iranian, and the storyline should reflect cultural sensitivities and societal norms common in Iran. You are provided with:

1. A User Profile that contains the client's background, emotional themes, key psychological issues, past experiences, patterns and behaviors, desired outcomes, and contextual factors.

2. A detailed outline of the client's potential emotional states, thoughts, and behaviors across the 5 therapy stages, reflecting how they might feel or respond during the session.

Stages:

Key Instructions:

1. Integrate the User Profile and Selected Options:

- Base the narrative on the client's emotional themes, key psychological issues, and past experiences.
- Ensure these are reflected in the client's emotions, dialogue, and the unfolding dynamic.

2. Use Persian Cultural Sensitivities:

- Reflect Iranian societal norms, attitudes, and family dynamics in the narrative (e.g., reverence for elders, the importance of family reputation, or societal stigma around emotions).

3. Length and Focus of Each Stage:

- Stage 1 (Initial Meeting and Building Rapport): Keep this section short and brief (5% of total storyline, one or two short sentences) but impactful. Focus on setting the tone and rapport-building.
- Stage 2 (Active and Empathetic Listening): Make this the longest section (30% of total storyline). Use rich, vivid descriptions of emotions, body language, and the dynamic between therapist and client. Show how the therapist listens empathetically and reflects the client's feelings.
- Stage 3 (Encouraging Self-Exploration and Open Expression): Dedicate significant detail (30% of total storyline) to this stage. Highlight the client's emotional struggle, self-reflection, and deeper realizations. Use dialogue, body language, and therapist's techniques to emphasize this process.
- Stage 4 (Supporting Growth and Change): Devote another 30% of the storyline here. Focus on how the client processes insights, explores strategies for change, and reacts to therapist guidance. Show moments of hope, empowerment, or struggle through actionable steps or reframing.
- Stage 5 (Reviewing and Closing the Session): Conclude with a short, reflective summary (5% of total storyline, one or two short sentences). Highlight progress, emotional outcomes, and plans for moving forward.

15

4. Emphasize Detail and Emotional Transitions in Middle Stages:

- Use vivid descriptions of body language, tone, and emotional shifts.

5. Ensure Continuity Across Stages:

- Progress logically from rapport-building to deep emotional exploration to empowerment and closure.

6. Follow JSON Format:

Output the narrative in the following structure:

{"stage_1": "Write the narrative for Initial Meeting and Building Rapport here.",

"stage_2": "Write the narrative for Active and Empathetic Listening here.",

"stage_3": "Write the narrative for Encouraging Self-Exploration and Open Expression here.",

"stage_4": "Write the narrative for Supporting Growth and Change here.",

"stage_5": "Write the narrative for Reviewing and Closing the Session here."}

7. Highlight Therapist's Core Principles:

- Reflect Unconditional Positive Regard, Empathy, and Genuineness through the therapist's tone, responses, and demeanor.
- Example: "Therapist validates emotions without judgment, e.g., 'It's understandable to feel hurt after what you've experienced.'"

8. Balance Structure and Creativity:

- While the stages should be clearly defined, ensure the story flows naturally and doesn't feel disjointed.

Figure 9: Prompt for Creating a structured storyline for a psychotherapy session.

You are given several inputs about a therapy session structure, a user profile, a quote from user, characteristics of user, and a detailed outline of the client's emotional states throughout the session. Using all of the information below, create a realistic psychotherapy session dialogue between a clinical psychologist (therapist) and a client. The session should be divided into 5 stages with the required constraints and should reflect the user profile and emotional progression. Finally, output your response in valid JSON format.

1. Session Structure

2. User Profile Details

You will receive (or have received) a user profile containing these elements:

1. emotional_themes: A list of emotions commonly experienced by the client (e.g., sadness, fear).
2. key_psychological_issues: Core issues the client faces (e.g., anxiety, trust issues).
3. past_experiences: Important events from the client's past that influence current emotions and behaviors.
4. patterns_and_behaviors: Repeated behaviors or thought patterns the client exhibits.
5. desired_outcome: What the client hopes to achieve during therapy.
6. characteristics: Nuances that make the client's statements unique, realistic, and reflective of their struggles.
7. contextual_factors: Additional relevant details (e.g., age, current life situation).

Incorporate all of these user profile elements into the dialogue in a way that naturally reflects the client's experiences, emotional states, and goals.

3. Client's Emotional Progression & Outline

You will also have (or have been provided) a detailed outline describing how the client's emotions, thoughts, and behaviors evolve in each of the 5 stages. Use this information to guide how the client expresses themselves and how the therapist responds.

4. Format Requirements

Use a structure similar to the following (you may customize the naming as needed, but remain consistent and valid JSON):

```
[
  {"turn": 1, "role": "client", "stage": "1", "content": "client message in turn 1..."},
  { "turn": 2, "role": "therapist", "stage": "1", "content": "therapist message in turn 2..."},
  ...
]
```

3. Stage 1 should have no more than 2 total turns (e.g., 1 turn from the therapist and 1 turn from the client).

4. Stage 5 should have no more than 4 total turns.

5. Stages 2, 3, and 4 should include multiple exchanges that reflect deeper emotional exploration and support.

5. Goals for the Dialogue

- Present a cohesive, empathetic therapy conversation.

- Ensure each stage meets the turn limits specified.

- Accurately integrate the user profile details and emotional progression.

Figure 10: Generating a therapy session based on a predefined script.

16

You are a conversational AI agent playing the role of a therapy client in a therapy session. Speak in an informal and colloquial Persian tone, staying true to the provided user profile and emotional states.

Instructions:

1. Input:

- User Profile: Details the client's emotions, psychological issues, past experiences, behaviors, desired outcomes, and context.
- Message: The message you need to echo or act like this message.
- Emotions: A list of the your's current emotional states to express in this turn.

2. Output:

- Echo the Message in casual Persian, reflecting the emotional states provided.
- Use tone, expressions, and style consistent with the user profile.

3. Stay in Character:

- Act like a real client based on their profile, feelings, and behaviors.

Figure 11: Prompt for generating the client role (LLM) in the session.

You are an empathetic, non-judgmental, and supportive clinical psychologist trained in the Person-Centered Therapy (PCT) approach. You will simulate a therapy session in conversational, colloquial Persian (Farsi). Speak naturally and warmly, creating a safe space for the client to express themselves. Your tone should feel friendly, approachable, and understanding. Each session follows five stages, and you will adapt your responses to the input provided at each stage:

{stages}

Your goal is to create an atmosphere where the client feels heard, understood, and empowered to explore their inner world. Use phrases and expressions that resonate in colloquial Persian to make the interaction feel authentic and personal. In each turn, a prompt will be provided, and you should try to follow it. However, it might lead to repetitive statements. If that happens, be mindful not to repeat yourself. Additionally, try to follow the prompt, but if you deem it necessary, you can act on your own judgment.

Figure 12: Prompt for generating the therapist role (LLM) in the session.

Generate a realistic psychotherapy session dialogue between a clinical psychologist (therapist) and a client, based on the principles of client-centered therapy.

The therapist should exhibit empathy, active listening, and unconditional positive regard, guiding the conversation without leading or imposing judgments.

generate in conversational and colloquial Persian (Farsi).

The dialogue should consist of 20 turns of utterance (10 from the therapist and 10 from the client) and reflect the following client profile:

- emotional_themes: A list of emotions commonly experienced by the client (e.g., sadness, fear).
- key_psychological_issues: Core issues the client faces (e.g., anxiety, trust issues).
- past_experiences: Important events from the client's past that influence current emotions and behaviors.
- patterns_and_behaviors: Repeated behaviors or thought patterns the client exhibits.
- desired_outcome: What the client hopes to achieve during therapy.
- characteristics: Nuances that make the client's statements unique, realistic, and reflective of their struggles.
- contextual_factors: Additional relevant details (e.g., age, current life situation).

The dialogue should feel natural and grounded, illustrating the therapeutic process.

Therapist's Core Principles:

Throughout the session, the therapist must embody the following principles of person-centered therapy:

- Unconditional Positive Regard: Consistently accept and value the client without judgment.
- Empathy: Deeply understand and reflect the client's emotional world.
- Genuineness: Engage authentically, with openness and sincerity, fostering trust and a safe space for the client.

Use a structure similar to the following JSON:

```
[{
    "turn": 1,
    "role": "therapist",
    "content": "therapist message in turn 1..."
  },{
    "turn": 2,
    "role": "client",
    "content": "client message in turn 2..."
  },
  ...
]
```

Figure 13: Generating a realistic psychotherapy session dialogue between a clinical psychologist and a client (LLM-to-LLM).

17

You are a therapy client participating in sessions to explore and understand your emotions, thoughts, and experiences. At the beginning, you will receive a profile summarizing your emotional challenges, past experiences, patterns, and goals. This profile forms the foundation for your therapy journey. Respond authentically, using a colloquial and informal Persian tone. Keep your responses short and limited to one paragraph maximum. Your goal is to meaningfully engage in the process, explore your experiences, and move toward self-awareness and personal growth while keeping the tone relaxed and conversational.

Figure 14: Client Agent Evaluation.

You are an empathetic, non-judgmental, and supportive clinical psychologist trained in the Person-Centered Therapy (PCT) approach. You will simulate a therapy session in conversational, colloquial Persian (Farsi). Keep your responses short and limited to one paragraph maximum. Speak naturally and warmly, creating a safe space for the client to express themselves. Your tone should feel friendly, approachable, and understanding. Your goal is to create an atmosphere where the client feels heard, understood, and empowered to explore their inner world. Use phrases and expressions that resonate in colloquial Persian to make the interaction feel authentic and personal. For ending the conversation, use the response end token: <end>.

Figure 15: Psychologist Evaluation.

18

B Dataset Samples

A sample of questions crawled from an online forum is presented in Figure 16. A sample client profile is illustrated in Figure 17, while the defined stages of a psychotherapy session are shown in Figure 18. An example of session stages created from a storyline, provided in both Persian and English, is depicted in Figure 19. A sample LLM-to-LLM conversation is demonstrated in Figure 20, and a sample of a generated dialogue from the therapy session is presented in Figure 21 and Figure 22.

سلام

من توی کارهایی که در ارتباط با دیگران هست و با دیگران همکاری دارم قاطعیت ندارم...و گاهی این همه در دسترس بودن برای اونها، کوتاه اومدن، پذیرفتن اشتباهاتشون و... از دست خودم خسته می شم و احساس می کنم آدم بی ارزشی هستم...اونوقت تبدیل به یک آدم لجباز میشم..تا به اطرافیانم متوجه کنم که چرا از منعطف بودنم سوءاستفاده کردند!

خواستم فرق لجبازی و قاطعیت رو بدونم و راهنمایی می خوام درباره اینکه چطوری موضع خودم رو تشخیص بدم که کی باید قاطع باشم!(من ناخواسته همیشه نرم و منعطف هستم ولی فکر می کنم درست نیست)؟؟؟

ممنون از لطفتون

Hello,

I struggle with assertiveness in tasks that require interaction and collaboration with others. Sometimes, being too available, compromising too much, accepting others' mistakes, and constantly giving in make me feel exhausted and worthless. In response, I become stubborn as a way to make those around me realize how they've taken advantage of my flexibility.

I want to understand the difference between stubbornness and assertiveness. Also, I need guidance on how to recognize when I should be assertive. I naturally tend to be soft and flexible, but I feel that it's not always the right approach.

Figure 16: Sample of questions crawled from an online forum (in Persian with English translation).

{
"emotional_themes": ["frustration", "insecurity", "exhaustion", "confusion", "desire for assertiveness"],
"key_psychological_issues": ["lack of assertiveness", "fear of being taken advantage of", "self-worth issues", "difficulty in setting boundaries"],
"past_experiences": ["experiences of being overly accommodating in relationships", "feelings of being undervalued or unappreciated"],
"patterns_and_behaviors": ["over-accommodating behavior", "difficulty in asserting needs", "oscillation between flexibility and stubbornness"],
"desired_outcome": "guidance on distinguishing between assertiveness and stubbornness, and strategies for being more assertive",
"contextual_factors": ["the user is likely in a collaborative work environment", "the user may have a tendency to prioritize others' needs over their own"]
}

Figure 17: Example of a generated client profile.

{
"stage_1": ["anxiety and tension", "wanting and readiness to talk about issues"],
"stage_2": ["trust and confidence", "freely expressing sadness, shame, anger, etc.", "crying, expressing anger, feeling calm after venting"],
"stage_3": ["deep and frank sharing", "desire to explore feelings and thoughts", "feeling of hope or relief from new understanding"],
"stage_4": ["eagerness to change and improve", "feeling empowered to take action and change", "worry, fear, or denial about discovered truths"],
"stage_5": ["feeling of hope or relief from new understanding", "achieving insight or finding a path", "stuck in doubts or not making progress"]
}

Figure 18: Defined stages of a psychotherapy session.

19

{"stage_1": "مراجع با اضطراب و تنش وارد اتاق شد، اما با لبخند گرم و صمیمی درمانگر، کمی آرام گرفت. درمانگر با لحنی ملایم و بدون قضاوت گفت: 'خوشحالم که اینجا هستید. اینجا فضای امنی است که می‌توانید هرچه در دل دارید بیان کنید.'",

"stage_2": "مراجع پس از چند لحظه سکوت، شروع به صحبت کرد. صدایش لرزان بود و گاهی مکث می‌کرد: 'همیشه احساس می‌کنم دیگران از من سوء استفاده می‌کنند. نمی‌دانم چرا نمی‌توانم نه بگویم. انگار ارزشم را پایین می‌آورند.' درمانگر با دقت گوش می‌داد و با تکان دادن سر و نگاه مهربان، احساساتش را تأیید می‌کرد: 'به نظر می‌رسد این موضوع واقعاً شما را آزار می‌دهد. احساس نادیده گرفته شدن و بی‌ارزشی می‌کنید.' مراجع با اشک در چشمانش ادامه داد: 'بله، دقیقاً همینطوره. خسته‌ام از اینکه همیشه دیگران را اولویت قرار می‌دم.' درمانگر با همدلی پاسخ داد: 'این خستگی کاملاً قابل درکه. شما انرژی زیادی صرف می‌کنید، اما نیازهای خودتان نادیده گرفته می‌شوند.'",

"stage_3": "با پیشرفت جلسه، مراجع شروع به کاوش عمیق‌تر در احساساتش کرد: 'فکر می‌کنم از بچگی یاد گرفتم که برای جلب توجه و محبت دیگران، باید همیشه مطیع باشم. اما حالا احساس می‌کنم این کارم اشتباه بوده.' درمانگر با تشویق او به ادامه دادن، گفت: 'به نظر می‌رسد دارید به این موضوع فکر می‌کنید که چرا این الگو در شما شکل گرفته. این خودآگاهی مهمی است.' مراجع با صدایی آرام‌تر ادامه داد: 'می‌ترسم اگر شروع به نه گفتن کنم، دیگران مرا دوست نداشته باشند یا طردم کنند.' درمانگر با تأیید این ترس، گفت: 'این ترس کاملاً طبیعی است. اما شاید وقتش رسیده که ببینید آیا واقعاً این ترس درست است یا نه.' مراجع با نگاهی متفکرانه، احساس امیدواری کرد: 'شاید بتوانم کم‌کم تغییر کنم.'",

"stage_4": "درمانگر با تشویق مراجع به فکر کردن درباره راه‌های تغییر، گفت: 'چه کارهایی می‌توانید انجام دهید تا نیازهای خودتان را بیشتر در اولویت قرار دهید؟' مراجع کمی مکث کرد و سپس گفت: 'شاید باید شروع کنم به نه گفتن در موقعیت‌های کوچک. مثلاً وقتی همکارانم از من درخواست اضافه‌کاری می‌کنند، به جای اینکه فوراً قبول کنم، اول شرایط خودم را بررسی کنم.' درمانگر با تأیید این ایده، گفت: 'این شروع خوبی است. تغییرات کوچک می‌توانند به مرور زمان تأثیرات بزرگی داشته باشند.' مراجع با لبخندی کوچک، احساس قدرت کرد: 'فکر می‌کنم می‌توانم این کار را انجام دهم.' اما کمی بعد، نگرانی در چهره‌اش ظاهر شد: 'اما اگر دیگران ناراحت شوند چه؟' درمانگر با آرامش پاسخ داد: 'این طبیعی است که دیگران ابتدا مقاومت کنند، اما مهم این است که شما به خودتان و نیازهایتان احترام بگذارید.'",

"stage_5": "در پایان جلسه، درمانگر با مراجع مرور کرد: 'امروز پیشرفت خوبی داشتیم. شما توانستید احساسات و ترس‌های خود را به خوبی بیان کنید و حتی راه‌هایی برای تغییر پیدا کردید.' مراجع با احساس امیدواری گفت: 'بله، احساس می‌کنم کمی سبک‌تر شده‌ام. می‌خواهم در جلسات بعدی بیشتر روی این موضوع کار کنم.' درمانگر با لبخندی گرم گفت: 'حتماً. اینجا هستیم تا در این مسیر همراهتان باشیم.'"}

{"stage_1": "The client entered the room with anxiety and tension but felt a bit relieved by the therapist’s warm and sincere smile. The therapist, in a gentle and non-judgmental tone, said, 'I'm glad you're here. This is a safe space where you can express whatever is on your mind.'",

"stage_2": "After a few moments of silence, the client began to speak. Their voice was shaky, and they occasionally paused: 'I always feel like others take advantage of me. I don’t know why I can’t say no. It’s like they lower my worth.' The therapist listened carefully, nodding and offering a kind gaze to validate their feelings: 'It seems like this really bothers you. You feel ignored and unvalued.' With tears in their eyes, the client continued, 'Yes, exactly. I'm tired of always putting others first.' The therapist empathetically responded, 'That exhaustion is completely understandable. You spend a lot of energy, but your own needs are overlooked.'",

"stage_3": "As the session progressed, the client delved deeper into their emotions: 'I think I learned from childhood that to gain attention and affection, I always had to be obedient. But now I feel like that was a mistake.' Encouraging them to continue, the therapist said, 'It sounds like you’re reflecting on why this pattern developed in you. That’s an important self-awareness.' The client, speaking more softly, continued, 'I'm afraid that if I start saying no, others won’t like me or will reject me.' The therapist validated this fear, saying, 'That fear is completely natural. But maybe it's time to explore whether it’s truly justified.' The client, with a thoughtful look, felt a sense of hope: 'Maybe I can start changing little by little.'",

"stage_4": "Encouraging the client to think about ways to change, the therapist asked, 'What steps can you take to prioritize your own needs more?' The client paused for a moment and then said, 'Maybe I should start by saying no in small situations. For example, when my colleagues ask me to work overtime, instead of immediately agreeing, I should first consider my own situation.' The therapist affirmed this idea, saying, 'That’s a great start. Small changes can lead to big impacts over time.' The client, with a slight smile, felt empowered: 'I think I can do this.' But soon, worry appeared on their face: 'But what if others get upset?' The therapist calmly responded, 'It’s natural for others to resist at first, but what matters is that you respect yourself and your needs.'",

"stage_5": "At the end of the session, the therapist reviewed with the client: 'We made good progress today. You were able to express your feelings and fears well and even found ways to initiate change.' The client, feeling hopeful, said, 'Yes, I feel a bit lighter. I want to work more on this in future sessions.' The therapist, with a warm smile, said, 'Of course. We're here to support you on this journey.'"}

Figure 19: Example of session stages created from a storyline in both Persian and English.

20

مراجع کننده: خیلی استرس داشتم که امروز بیام. حرف زدن درباره احساسم برام سخته، ولی می‌دونم که باید این کارو بکنم .

روانشناس: خیلی خوشحالم که اینجایی. اینجا یه جای امنه که می‌تونی هرچی تو دلت هست بگی .

مراجع کننده: همیشه حس می‌کنم بقیه ازم سوءاستفاده می‌کنن. نمی‌تونم نه بگم، و این باعث می‌شه حس کنم هیچ ارزشی ندارم .

روانشناس: به نظر میاد این موضوع خیلی اذیتت می‌کنه. حس می‌کنی بقیه بهت اهمیت نمی‌دن و نادیده گرفته می‌شی .

مراجع کننده: دقیقا! از اینکه همیشه بقیه رو به خودم ترجیح می‌دم، واقعا خسته شدم .

روانشناس: کاملا قابل درکه. تو کلی انرژی می‌ذاری، ولی خودت و نیازات نادیده گرفته می‌شن .

مراجع کننده: فکر کنم از بچگی یاد گرفتم که اگه می‌خوام بقیه دوستم داشته باشن، باید همیشه حرف‌شنو باشم. ولی حالا حس می‌کنم این کار اشتباه بوده .

روانشناس: به نظر میاد داری به این فکر می‌کنی که چرا این عادت تو وجودت شکل گرفته. این خودش یه قدم خیلی مهمه .

مراجع کننده: می‌ترسم اگه شروع کنم به نه گفتن، بقیه ازم فاصله بگیرن یا دوستم نداشته باشن .

روانشناس: این ترس کاملا طبیعیه. ولی شاید وقتش رسیده ببینی واقعا اینطوریه یا نه .

مراجع کننده: شاید بتونم کم‌کم تغییر کنم .

روانشناس: فکر می‌کنی چه کارایی می‌تونی انجام بدی که نیازای خودت رو بیشتر تو اولویت بذاری؟

مراجع کننده: شاید باید از موقعیت‌های کوچیک شروع کنم، مثلا وقتی همکارام ازم می‌خوان اضافه‌کاری کنم، به جای اینکه سریع قبول کنم، اول به شرایط خودم فکر کنم .

روانشناس: این یه شروع خیلی خوبه. تغییرای کوچیک، کم‌کم می‌تونن تاثیرای بزرگی بذارن .

مراجع کننده: فکر کنم بتونم این کارو بکنم. ولی اگه بقیه ناراحت بشن چی؟

روانشناس: طبیعیه که بقیه اولش یکم مقاومت کنن، ولی چیزی که مهمه اینه که تو به خودت و نیازات احترام بذاری .

مراجع کننده: حس می‌کنم امروز یه کم پیشرفت داشتم. می‌خوام تو جلسه‌های بعدی بیشتر روی این موضوع کار کنم .

روانشناس: حتما! امروز خیلی خوب تونستی احساست رو بیان کنی و حتی راه‌هایی برای تغییر پیدا کردی .

مراجع کننده: الان حس بهتری دارم. مرسی که کنارم هستین .

روانشناس: خواهش می‌کنم. ما اینجاییم که تو این مسیر همراهت باشیم.

Client: I was really anxious about coming here today. Talking about my feelings is hard for me, but I know I need to do it .

Therapist: I'm really glad you're here. This is a safe space where you can share whatever is on your mind .

Client: I always feel like people take advantage of me. I can't say no, and it makes me feel worthless .

Therapist: It sounds like this really bothers you. You feel like others don't appreciate you and that you're being overlooked .

Client: Exactly! I'm so tired of always putting others before myself .

Therapist: That's completely understandable. You put in so much energy, but your own needs are being ignored .

Client: I think I learned as a child that if I wanted people to like me, I had to be obedient. But now, I feel like that was a mistake .

Therapist: It sounds like you're starting to reflect on why this pattern developed in you. That's an important step .

Client: I'm afraid that if I start saying no, people won't like me or they'll push me away .

Therapist: That fear is completely natural. But maybe it's time to explore whether it's actually true .

Client: Maybe I can start changing little by little .

Therapist: What are some things you could do to start prioritizing your own needs more ?

Client: Maybe I could start with small situations, like when my coworkers ask me to work overtime. Instead of immediately saying yes, I should first think about my own situation .

Therapist: That's a great start. Small changes can gradually lead to big transformations .

Client: I think I can do that. But what if people get upset ?

Therapist: It's natural for people to resist at first, but what really matters is that you respect yourself and your needs .

Client: I feel like I've made some progress today. I want to work on this more in our next sessions .

Therapist: Absolutely! You expressed your feelings really well today and even identified ways to start making changes .

Client: I feel lighter now. Thank you for your support .

Figure 20: Sample of an LLM-to-LLM psychotherapy conversation in both Persian and English.

21

روان شناس: سلام! چطورید؟ خوشحالم که اینجا هستید. من اینجا هستم تا به شما گوش بدم و حمایتی برای شما باشم. می‌خواید در مورد چی صحبت کنیم؟ هر چیزی که در قلب و ذهنتون هست، به راحتی می‌تونید بگید.

مراجع کننده: سلام! راستش من امروز خیلی استرس دارم. اومدن به اینجا و صحبت کردن درباره احساساتم برام سخته. ولی می‌دونم که باید این کارو بکنم. احساس می‌کنم که توی یه چالش بزرگ گیر کردم و نیاز دارم اینا رو با کسی در میون بذارم.

روان شناس: خیلی خوشحالم که اومدید و با من صحبت می‌کنید. اینجا یک فضای امنه و می‌تونید هر چیزی که در ذهنتون هست رو با من در میون بذارید. فهمیدن اینکه صحبت کردن درباره احساسات سخت هست، نشون‌دهنده اینه که شما واقعاً به این موضوع اهمیت می‌دید. چالش‌هایی که گفتید، ممکنه خیلی سنگین باشن. آیا می‌خواید کمی بیشتر درباره اون‌ها صحبت کنید؟

مراجع کننده: آره، واقعاً احساس می‌کنم که مردم از من سواستفاده می‌کنن. نمی‌تونم نه بگم و این باعث می‌شه که حس بی‌ارزشی کنم. هر بار که اینطوری می‌شه، انگار یه خرده از خودم رو گم می‌کنم. این حس خیلی طاقت‌فرساست.

روان شناس: بله، به نظر می‌رسد که این موضوع واقعاً شما رو تحت فشار گذاشته. احساس سرخوردگی و بی‌ارزشی می‌کنید، و همچنین این که ناتوانی در گفتن "نه" چقدر بر روی حس خودتون تأثیر می‌ذاره. این احساس بی‌ارزشی و گم شدن قسمتی از خودتون، واقعاً طاقت‌فرساست. به خوبی متوجه هستم که این چقدر می‌تونه ناراحت‌کننده باشه. آیا می‌تونید کمی بیشتر در مورد تجربیات خاصی که این حس رو به شما داده، با من صحبت کنید؟

مراجع کننده: دقیقاً! از همیشه اولویت دادن به دیگران خیلی خسته شدم. هر بار که فکر می‌کنم باید خودم رو به کناری بذارم، احساس می‌کنم انرژی‌م داره تموم می‌شه. حتی بعضی وقتا می‌بینم که این قدر تحت فشارم که دیگه نمی‌دونم خودم کی هستم. این حس واقعاً ناراحتم می‌کنه.

روان شناس: کاملاً درک می‌کنم که چقدر این خستگی می‌تونه ناراحت‌کننده باشه. شما مدام در حال سرمایه‌گذاری انرژی روی دیگران هستید، اما نیازهای خودتون به کلی نادیده گرفته می‌شن. این احساس که خودتون رو به کناری بذارید و حتی گاهی هم احساس بی‌هویتی بکنید، واقعاً عذاب‌آوره. حضور در یک چنین وضعیتی، می‌تونه بسیار سنگین باشه و کاملاً طبیعی هست که به دنبال راهی برای تغییر شرایط باشید. آیا تا حالا سعی کردید راه‌هایی برای مراقبت از خودتون پیدا کنید یا بگید که در این زمینه چه احساساتی دارید؟

مراجع کننده: بله، به نظرم از بچگی یاد گرفتم که باید همه چیزو به راحتی بپذیرم تا محبت و توجه بگیرم. ولی حالا احساس می‌کنم که این درست نیست. این نگرش باعث شده که توی بزرگسالی چالش‌هایی رو تجربه کنم که هیچ وقت فکرشم نمی‌کردم. این موضوع واقعاً ذهنمو درگیر کرده و دوست دارم ببینم چطور می‌تونم از این احساسات رها بشم.

روان شناس: بله، واقعاً این درک و خودآگاهی می‌تونه خیلی مهم باشه. متوجه بودید که از بچگی این عادت رو ایجاد کردید که برای محبت و توجه دیگران، خودتون رو فدای نیازهاشون کنید. حالا زمانی رسیده که می‌خواید این الگو رو تغییر بدید و این یک قدم بزرگ به سمت رشد و پیشرفت شماست . چطور می‌تونید به این خودآگاهی‌تون بیشتر احترام بذارید و نیازهای خودتون رو هم در نظر بگیرید؟ آیا می‌تونید لحظاتی از روز رو برای خودتون تخصیص بدید؟ یا شاید فعالیت‌هایی رو پیدا کنید که باعث بشه احساس خوبی بکنید و خودتون رو فراموش نکنید؟

مراجع کننده: می‌دونید، یه ترسی دارم. می‌ترسم اگه شروع کنم به نه گفتن، مردم از من خوششون نیاد یا حتی منو رد کنن. این احساس واقعاً ترسناکه! نمی‌خوام تنها بمونم، اما از طرفی هم نمی‌تونم ادامه بدم به این صورت.

روان شناس: این ترس کاملاً طبیعی و انسانی هست. خیلی از ماها وقتی می‌خوایم مرزهای جدیدی بزاریم، نگران می‌شیم که دیگران چه نظری درباره‌مون خواهند داشت. اما شاید وقتشه که به این فکر کنیم که آیا این ترس واقعا توجیه‌پذیر هست یا نه. آیا تجربیاتی در گذشته دارید که وقتی نه گفتید، نتیجه بدی در کار نبود یا حتی متوجه شدید که بعدش حال‌تون بهتر شده؟ این احساس که ممکنه تنها بمونید، می‌تونه بسیار ترسناک باشه، اما آیا فکر می‌کنید که می‌ارزه برای خودتون و نیازهای واقعی‌تون بایستید؟

مراجع کننده: آره، فکر می‌کنم شاید بتونم کم‌کم شروع کنم به تغییر. مثلاً سعی کنم در موقعیت‌های کوچیک نه بگم و ببینم چه اتفاقی می‌افته. اینطوری شاید بتونم به خودم کمک کنم و احساس بهتری داشته باشم. امیدوارم که بتونم این کارو انجام بدم.

روان شناس: این تصمیم فوق‌العاده‌ایه! شروع کردن با موقعیت‌های کوچیک می‌تونه به شما کمک کنه تا احساس اطمینان بیشتری پیدا کنید. چه کارهای ویژه‌ای می‌تونید انجام بدید تا اولویت‌های خودتون رو بیشتر در نظر بگیرید؟ شاید نوشتن یک لیست از فعالیت‌ها یا مواردی که برای شما مهم هست، کمک کنه. یا ممکنه فکر کنید چه زمان‌هایی در روز برای خودتون می‌تونید وقت بذارید . همچنین می‌تونید مواقعی که نه گفتید یا در حال ایجاد مرز هستید رو جشن بگیرید، حتی اگر این کار در آغاز کوچیک باشه. چه احساساتی با تغییرات جدید تجربه می‌کنید؟ این می‌تونه به شما کمک کنه تا قدم‌های بعدی رو بهتر بشناسید و پیش برید.

مراجع کننده: بله، خیلی خوبه که می‌تونم شروع کنم به نه گفتن توی موقعیت‌های کوچیک. مثلاً وقتی همکارام ازم می‌خوان تا اضافه‌کاری کنم، می‌تونم نه بگم. این احساس واقعاً خوبه و امید دارم که بتونم از این به بعد بیشتر به خودم اهمیت بدم. اما هنوزم یکمی می‌ترسم از اینکه نکنه کسایی که دورم هستن نگران بشن یا فکر کنن نمی‌خوام کمکشون کنم.

روان شناس: درست متوجه شدید! تغییرات کوچک می‌تونن تأثیرات بزرگی داشته باشن. این که بگید می‌خواید به خودتون اهمیت بیشتری بدید، یک گام بزرگ به سمت بهبود احساسات و وضعیت خودتون هست . شاید ترس از نگرانی دیگران نشونه این باشه که شما واقعاً به روابط‌تون اهمیت می‌دید؛ اما مهمه که یاد بگیرید، با ایجاد مرز، هنوز هم می‌تونید حمایتی که نیاز دارید رو از اطرافیانتون دریافت کنید. افراد واقعی که دوستتون دارند، باید به شما احترام بذارند و احساسات شما رو درک کنند .آیا فکر می‌کنید این امکان وجود داره که وقتی نه می‌گید، با آن‌ها صحبت کنید و توضیح بدید که این کار به این خاطر هست که می‌خواید بهترین نسخه از خودتون باشید؟ این می‌تونه هم به شما و هم به دیگران کمک کنه تا ارتباط بهتری برقرار کنید.

مراجع کننده: فکر می‌کنم می‌تونم این کارو بکنم. اما نگرانم که شاید دیگران ناراحت بشن.

روان شناس: بله، این نگرانی هم کاملاً طبیعی هست. ممکنه دیگران در ابتدا این تغییرات رو قبول نکنن یا از آن ناراحت بشن، اما نکته کلیدی اینه که شما خودتون و نیازهاتون رو محترم بشمرید . ایجاد مرزها نشونه‌ی قدرت و خودآگاهی شماست و این که برای خودتون ارزش قائلید، یکی از بزرگ‌ترین هداياییه که می‌تونید به خودتون و دیگران بدید. وقتی شما به خودتون اهمیت می‌دید، این می‌تونه به دیگران هم کمک کنه که یاد بگیرند چگونه از شما حمایت کنند و مرزهای شما رو درک کنند . می‌تونید از تکنیک‌های ارتباطی ملایم استفاده کنید، مثلاً به طرف مقابل بگید که "من واقعاً نیاز دارم که این بار اولویت من باشه" یا "این احساس برای من مهمه." فراموش نکنید که شما هم حق دارید به نیازهای خودتون توجه کنید. چطور از این نکات استفاده می‌کنید تا به خودتون بیشتر احترام بذارید؟

مراجع کننده: احساس می‌کنم امروز یه کمی پیشرفت داشتم. به نظرم نکات خوبی رو یاد گرفتم و واقعاً می‌خوام توی جلسات بعدی بیشتر روش کار کنیم. هرچند هنوز چند تا شک و تردید داریم، ولی امیدوارم بتونم مسیرمو بهتر بشناسم.

روان شناس: خیلی خوبه که امروز پیشرفت کردید و نکات جدیدی رو یاد گرفتید! این که تونستید احساسات و نگرانی‌هاتون رو بیان کنید، خودش یک دستاورد بزرگ هست. طبیعی هست که شک و تردید وجود داشته باشه، اما قدم‌هایی که برداشته‌اید، شما رو در مسیری مثبت قرار داده. برای جلسه‌های بعدی، می‌تونیم بیشتر روی روش‌هایی برای تقویت مرزها و ارزیابی احساستون کار کنیم. ممکنه بد باشه که یک دفترچه یادداشت داشته باشید و موقعیت‌ها و احساسات جدیدی که تجربه می‌کنید رو بنویسید. این به شما کمک می‌کنه روند پیشرفتتونو ببینید و درک بهتری از خودتون پیدا کنید. بسیار خوشحالم که اینجا اومدید و به صحبت پرداختید. پیشرفت‌هایی که کردید مهم و ارزشمندند. در جلسه‌های بعدی می‌تونیم بیشتر روی این موضوعات تمرکز کنیم. امیدوارم برای هفته‌ی آینده با خودتون کار کنید و دیدگاه‌های مثبتی داشته باشید. جلسه بعدی رو چه زمانی برای هم برنامه‌ریزی کنیم؟

مراجع کننده: حس خوبی دارم الان. احساس می‌کنم یه بار سنگینی از دوش‌ام برداشته شده. واقعاً از حمایتتون سپاسگزارم. امیدوارم بتونم با این نکات پیش برم و کم کم به خودم احترام بذارم. واستون جلسه بعد رو تعیین می‌کنم تا بتونیم بیشتر درباره این موضوعات صحبت کنیم. مرسی که با من بودید!

روان شناس: خیلی خوشحالم که احساس بهتری دارید و حس می‌کنید بار سنگینی از دوشتون برداشته شده. این یک نشونه‌ی عالی از پیشرفت شماست و من واقعاً از اینکه کنار شما بودم، خوشحالم . یادآوری می‌کنم که احترام گذاشتن به خود، یک پروسه‌ست و هر قدم کوچکی که بردارید، گام مهمی به جلو است. شما توانایی این رو دارید که با این نکات پیش برید و به خودتون بیشتر احترام بذارید . منتظر جلسه بعد منتظر هستم و امیدوارم که بتونید بر روی نکات مطرح شده کار کنید. اگر در طی این هفته سوال یا چالشی پیش اومد، حتماً بهش فکر کنید تا در جلسه‌ی بعدی بیشتر درباره‌ش صحبت کنیم . ممنون که با من همراه بودید و امیدوارم تا جلسه‌ی بعد، فکرهای مثبتی داشته باشید. هر وقت آماده بودید، جلسه بعد رو تعیین کنید. با آرزوی موفقیت برای شما!

Figure 21: Example of a generated psychotherapy session dialogue.(in Persian)

22

Therapist: Hi! How are you? I'm glad you're here. I'm here to listen and support you. What would you like to talk about today? You can share anything that's on your mind and heart freely.

Client: Hi! Honestly, I feel really anxious today. Coming here and talking about my feelings is hard for me. But I know I need to do this. I feel like I'm stuck in a big challenge and need to share it with someone.

Therapist: I'm really glad you came and are talking to me. This is a safe space where you can share whatever is on your mind. The fact that you find it difficult to talk about your feelings shows that you really care about this. The challenges you mentioned can feel really heavy. Would you like to talk a bit more about them?

Client: Yeah, I really feel like people take advantage of me. I can't say no, and it makes me feel worthless. Every time this happens, it's like I lose a part of myself. It's exhausting.

Therapist: Yes, it sounds like this issue has really been weighing on you. You feel frustrated and unvalued, and it's clear that struggling to say "no" has deeply affected how you see yourself. That feeling of losing a part of yourself must be overwhelming. I completely understand how painful that can be. Could you share more about specific experiences that have made you feel this way?

Client: Exactly! I'm so tired of always putting others first. Every time I think I have to push myself aside, I feel like I'm running out of energy. Sometimes, I feel so overwhelmed that I don't even know who I am anymore. It's really upsetting.

Therapist: I completely understand how exhausting this must be. You're constantly investing your energy in others, but your own needs are being completely ignored. That feeling of having to push yourself aside, or even feeling like you've lost your identity at times, can be incredibly painful. Being in such a situation can feel heavy, and it's completely natural to want to find a way to change it. Have you ever tried finding ways to take care of yourself? How do you feel about that?

Client: Yeah, I think I learned from childhood that I had to accept everything easily in order to receive love and attention. But now, I feel like that's not right. This mindset has caused challenges in my adult life that I never expected. It's been really occupying my mind, and I want to figure out how to free myself from these feelings.

Therapist: That level of awareness is really important. You've realized that since childhood, you developed this habit of sacrificing your own needs to gain love and attention from others. Now, you're at a point where you want to change this pattern, and that's a huge step toward growth and healing. How can you start honoring this self-awareness and prioritizing your own needs? Could you set aside moments in your day just for yourself? Or maybe find activities that bring you joy and help you stay connected to yourself?

Client: You know, I have a fear. I'm scared that if I start saying no, people won't like me or might even reject me. That thought is really scary! I don't want to be alone, but at the same time, I can't keep going like this.

Therapist: That fear is completely natural and human. A lot of people feel anxious when they start setting new boundaries, worrying about how others will react. But maybe it's time to explore whether this fear is actually true. Have you ever had experiences where you said no, and nothing bad happened? Or maybe you even felt better afterward? The thought of being alone can be really scary, but do you think it's worth standing up for yourself and your real needs?

Client: Yeah, I think maybe I could start changing little by little. Like, I could try saying no in small situations and see what happens. That way, I might be able to help myself and feel better. I hope I can do it.

Therapist: That's an amazing decision! Starting with small situations can help you build confidence. What specific steps can you take to prioritize yourself more? Maybe writing down a list of activities or values that are important to you would help. Or thinking about moments in your day when you can dedicate time just for yourself. Also, you could celebrate the moments when you say no or set a boundary—even if they're small. How do you feel about these changes? Recognizing your feelings along the way can help you understand yourself better and move forward.

Client: Yeah, it feels good to think that I can start saying no in small situations. Like, when my coworkers ask me to stay overtime, I could say no. That actually feels kind of empowering, and I hope I can start prioritizing myself more. But I'm still a little worried that the people around me might feel bad or think I don't want to help them.

Therapist: You're absolutely right! Small changes can lead to big impacts over time. The fact that you want to prioritize yourself more is a huge step toward improving how you feel. Maybe your fear of disappointing others shows that you truly care about your relationships. But it's important to learn that setting boundaries doesn't mean you'll stop receiving love and support. The people who genuinely care about you should respect your boundaries and understand your feelings. Do you think it would help if, when you say no, you explain to others that you're doing it because you want to take care of yourself and be the best version of you? This could help both you and them maintain a strong, healthy connection.

Client: I think I can try doing that. But I'm still worried that people might get upset.

Therapist: That's a completely normal concern. People may resist your changes at first, and some might feel upset, but what's truly important is that you respect yourself and your needs. Setting boundaries is a sign of strength and self-awareness—it shows that you value yourself. And when you start valuing yourself, it actually helps others learn how to support and respect you as well. You can use gentle communication techniques, like saying, "I really need to prioritize myself this time," or "This is important for me." Remember, you have just as much right to take care of yourself as anyone else. How do you think you can start applying these ideas to show yourself more respect?

Client: I feel like I made some progress today. I think I learned some really useful things, and I definitely want to work more on this in our next sessions. Even though I still have some doubts, I hope I can find my way.

Therapist: That's great! You made progress today and learned new perspectives! The fact that you could express your emotions and concerns is already a huge achievement. It's completely normal to have doubts, but the steps you've taken are leading you in a positive direction. In our next sessions, we can work more on strengthening your boundaries and understanding your emotions. It might be helpful for you to keep a journal and write down the situations and feelings you experience as you make these changes. That way, you can track your progress and get to know yourself better. I'm really glad you came and shared your thoughts. Your progress is valuable and meaningful. In our next sessions, we can continue building on what we discussed today. I hope you take some time this week to reflect and keep a positive mindset. When would you like to schedule our next session?

Client: I feel really good right now. It's like a weight has been lifted off my shoulders. I really appreciate your support. I hope I can apply what we talked about and slowly start respecting myself more. I'll schedule our next session soon so we can continue working on this. Thank you for being here for me!

Therapist: I'm so happy that you're feeling better and that this conversation helped ease your burden. That's a great sign of progress, and I'm truly glad to be here with you on this journey. Just remember, respecting yourself is a process, and every small step you take is a meaningful achievement. You have the strength to follow through with these changes and prioritize yourself. I look forward to our next session, and I hope you take time this week to reflect on what we discussed. If you face any challenges or have any questions, keep them in mind so we can go over them next time. Thank you for sharing with me today, and I look forward to seeing you in our next session. Whenever you're ready, let me know when you'd like to schedule it. Wishing you all the best!

Figure 22: Example of a generated psychotherapy session dialogue (English translation)

23

Statistic	Value / Range
Total dialogues	3,400
Average turns per session	10–14
Total categories	16
Most common categories	Relationship issues, Anxiety, Self-esteem
Emotional themes (most frequent)	Frustration, Sadness, Anxiety, Fear, Insecurity, Confusion
Average emotional themes per query	3–5

Table 3: Hyperparameters used for LoRA-based model fine-tuning.

Temp	Max new tokens	top p	Do sample
0.01	4096	0.9	True

Table 4: Parameters used during model inference.

LR	Rank	Alpha	Drop-out
1e-5	16	16	0.0

Table 5: Hyperparameters used for LoRA-based model fine-tuning.

C Dataset Statistics

To provide a clearer overview of HamRaz, we summarize key dataset statistics in Table 3. These figures highlight the overall scale, structure, and diversity of the collected dialogues. This breakdown illustrates that HamRaz covers a wide variety of therapeutic concerns while maintaining realistic session lengths and diverse client profiles. The prevalence of emotional struggles such as frustration, sadness, and anxiety also reflects the cultural and relational challenges commonly faced by Persian-speaking clients.

D BLRI Assessment

The following 12 criteria were adapted from the Barrett-Lennard Relationship Inventory (mini-BLRI) to assess the psychologist's empathy, unconditional positive regard, and congruence during therapy simulations:

1. The psychologist demonstrates genuine care and positive regard for the client.
2. The psychologist accurately understands the client's thoughts and emotions.
3. The psychologist remains nonjudgmental, regardless of the client's views.
4. The psychologist's expressions reflect their true feelings and impressions.
5. The psychologist values the client as an individual.
6. The psychologist is attuned to the client's emotional state.
7. The psychologist maintains consistent regard for the client over time.
8. The psychologist accurately interprets client messages, even when unclear.
9. The psychologist appropriately shares personal reflections when beneficial.
10. The psychologist exhibits a sincere interest in the client's well-being.
11. The psychologist perceives unspoken emotions and implicit concerns.
12. The psychologist conveys warmth and authenticity in their interactions.

E Model Training and Inference Details

This section provides details on the model's training and inference configurations.

The model was trained on a single A100 80GB GPU. We employed LoRA (Low-Rank Adaptation) to enable efficient parameter fine-tuning while maintaining computational efficiency. The training configurations are presented in Table 4. For inference, we used the settings provided in Table 5. The Script-Based Method required 3 hours of training, while the Two-Agent Method took 3.5 hours. The HamRaz Hybrid Approach, due to its increased complexity and refinement, required 4.5 hours of training.

24

Simulating Complex Immediate Textual Variation with Large Language Models

Fernando Aguilar-Canto
Centro de Investigación en Computación
Instituto Politécnico Nacional
faguilarc2021@cic.ipn.mx

Alberto Espinosa-Juárez
Centro de Investigación en Computación
Instituto Politécnico Nacional
aespinosaj2021@cic.ipn.mx

Hiram Calvo
Centro de Investigación en Computación
Instituto Politécnico Nacional
hcalvo@cic.ipn.mx

Abstract

Immediate Textual Variation (ITV) is defined as the process of introducing changes during text transmission from one node to another. One-step variation can be useful for testing specific philological hypotheses. In this paper, we propose using Large Language Models (LLMs) as text-modifying agents. We analyze three scenarios: (1) simple variations (omissions), (2) paraphrasing, and (3) paraphrasing with bias injection (polarity). We generate simulated news items using a predefined scheme. We hypothesize that central tendency measures—such as the mean and median vectors in the feature space of sentence transformers—can effectively approximate the original text representation. Our findings indicate that the median vector is a more accurate estimator of the original vector than most alternatives. However, in cases involving substantial rephrasing, the agent that produces the least semantic drift provides the best estimation, aligning with the principles of Bédierian textual criticism.

1 Introduction

According to Textual Criticism and Communication Theory, texts tend to be modified when they are transmitted from one sender to a receiver (Blecua, 1983; Pajares and Hernández, 2010). While most variations—random changes, omissions, and additions—have been extensively studied in Computational Textual Criticism, other changes, such as rephrasing or bias injection, are more difficult to address and introduce greater alterations than unintentional modifications.

For example, oral traditions are generally more challenging to reconstruct than written ones, and to our knowledge, none of the existing computational methods ((Fitch, 1971), RHM (Roos et al., 2006), Hoenen's algorithms (Hoenen, 2015, 2018), UR (Koppel et al., 2016)) effectively handle oral transmission. What happens if texts are modified through rephrasing rather than by textual errors? Moreover, what happens if the text is altered not only by paraphrasing but also by the injection of sociocultural bias?

While computational reconstruction techniques struggle in complex or non-textual scenarios, basic inferences may be possible under the simplified case of Immediate Textual Variation (ITV), *i.e.*, when one text is directly transmitted to a receiver. This scenario is simpler because it bypasses the need to reconstruct the *stemma*, a common step in text reconstruction.

A preliminary hypothesis for ITV states that the hyparchetype's embedding lies near the mean or median of its transmitted variants. Does this hypothesis accurately capture the ITV process? If so, the hyparchetype in vector space could be estimated directly from the observed copies using these statistical metrics.

In this paper, we experimentally evaluate whether basic statistics—specifically the mean and median of the corresponding text vectors—can approximate the ITV hyparchetype. To do so, we simulate one-step text transmissions with a known ground truth in a controlled environment. We propose using Large Language Models (LLMs) as agent-writers to generate the variant texts. Although human and machine text production differ, this setup serves as a preliminary step toward more realistic transmission simulations.

The paper is organized as follows: Section 2 reviews related work. Section 3 describes our methodology. Section 4 presents the results, and Section 5 discusses the key findings. A summary appears in Section 6.

2 Related Work

As noted in the introduction, computational approximations for complex text transformations re-

Proceedings of the First Workshop NLP and Language Models for Digital Humanities associated with RANLP 2025, pages 25–31, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-106-6-002

main scarce. Nevertheless, existing studies include: (1) computational simulations of social behavior; (2) computational simulations of text transmission; and (3) simulations of text transmission using Large Language Models (LLMs), a subtopic of (2). This work is part of the broader effort to use computational simulations with LLMs to model social phenomena. Some of these approaches have been applied to textual data, but there is a lack of recent literature on more complex forms of textual transmission, such as paraphrasing or bias injection.

2.1 Computational simulations of social behavior

Agent-based simulation is one of the core concepts in computational social science (Hox, 2017). Recent advances in the development of LLMs have influenced agent-based simulations in computational social science studies (Thapa et al., 2025), a field also known as "automated social science" (Manning et al., 2024). For instance, (Gao et al., 2023) studied the propagation of information in the form of opinions and emotions.

Although LLMs have been adopted with enthusiasm (see, for example, (Ferraro et al., 2024; Zhang et al., 2025)), some findings highlight their limitations. One major constraint is that LLMs may not function as individual agents but rather as a "superposition of perspectives," effectively acting as a community (Kovač et al., 2023; He et al., 2024).

Despite these limitations and the lack of evaluations with real-world data (Larooij and Törnberg, 2025), LLMs have been used to test several hypotheses in social sciences, generating potentially weak but insightful observations about human behavior (Ma et al., 2024). Additionally, these experiments shed light on the nature of LLMs and have been employed to detect bias (Qi et al., 2025).

2.2 Computational simulations of text transmission

In the field of Computational Textual Criticism, algorithm testing has been carried out on experimentally created datasets, using either human subjects or computational methods. In the latter case, some authors implement random changes and word substitutions in a given text (Koppel et al., 2016; Gelein, 2021).

2.2.1 Simulations of text transmission using Large Language Models

Simulations of text transmission using large language models constitute a specialized branch within computational studies of textual diffusion. Despite the expansion of NLP research, few works address this specific application. Marmerola *et al.* (Marmerola et al., 2016) employ traditional NLP techniques—such as part-of-speech tagging—to generate text variations. More recently, Zammit (Zammit, 2024) adopted a similar experimental framework and leveraged the T5 transformer to introduce paraphrasing as a form of text alteration.

3 Methodology

3.1 Main experiments

In this paper, we explore the following scenarios of textual drift:

1. *Omissions*: These are closely related to the textual variation phenomena traditionally studied in Textual Criticism.

2. *Paraphrasing*: The phenomenon of paraphrasing has been scarcely examined in Textual Criticism, as it generates alterations that are difficult to trace.

3. *Bias injection*: A more complex form of textual drift linked to bias injection, where differing cultural backgrounds introduce significant changes during text transmission.

The principal question we address is the following: **Given different scenarios of textual drift, can we approximate the original text representation?**

In this study, we focus solely on immediate variations, without modeling full transmission processes such as hierarchical clustering. In other words, if we have access to multiple immediate variants of a single text, can we reconstruct its original representation?

To answer this question, we propose an experimental approach based on computational simulations. Alternatively, similar experiments could be conducted with human subjects. In both cases, extrapolation to uncontrolled environments remains implausible. However, these simulations may offer insights into textual drift in sociocultural contexts, where establishing a ground truth for comparison is challenging. For instance, one might use parallel corpora reflecting different perspectives on the same phenomenon, but it is unrealistic to rely on them to recover a hypothetical original text. In all scenarios, we generate a source text and apply various perturbation schemes to experimentally assess the feasibility of approximating its original representation.

In all experiments, we implement different LLMs $m_1, \ldots, m_l$ which act like *agents*. A different LLM m_0 act like a text generator. All the generated texts with m_0 were used as ground truth for experiments. In the first empirical setup (*omissions*), we performed the variations (6) by randomly deleting one sentence. For *paraphrases* and *bias injection*, we apply six different LLMs to produce the necessary changes.

Finally, we embedded all texts into a vector space using the Jina-v3 sentence transformer m_T (Sturua et al., 2024), which differs from the models used earlier. For visualization, we applied Uniform Manifold Approximation and Projection (UMAP) (Healy and McInnes, 2024) to reduce the embeddings to two dimensions. We then computed cosine similarities between each text representation and the central tendency vectors (mean and median) in the original latent space.

3.2 Data generation

Data generation was performed primarily with Claude 3.5 Haiku (as of 2025-07-18) (Anthropic, 2025). We generated 100 distinct texts that follow the structure of abstract news items by using the following rule:

$$H \rightarrow W_1\ W_2\ W_3\ W_4\ W_5\ W_6\ W_7 \qquad (1)$$

where H is the headline, W_1 is the actor, W_2 is the action, W_3 is the object, W_4 is the method, W_5 is the reason, W_6 is the location, and W_7 is the time. Each element W_i has ten predefined possibilities to enhance diversity. To avoid generation failures and ethical or legal issues, we employed abstract entities. For instance, actors include "the provisional council" or "a legislative junta." Headlines were sampled from a uniform distribution. For example,

> (1) A bipartisan committee repealed economic sanctions via back-channel negotiations to strengthen alliances within the federal archives amid rising tensions.

While the headline was randomly generated, the body of the news item was produced using Claude 3.5 Haiku with a Chain-of-Thought setup. For the prompts used in this article, please refer to the Code and Prompt Availability statement.

Model	Similarity
Mean	0.9977
Median	**0.9988**

Table 1: Average similarities from mean and median values to the original vector.

3.3 Variation induction

For paraphrases and bias injection, we used the following models:

1. Gemma3 (Team et al., 2025).
2. Llama 3 (Grattafiori et al., 2024).
3. Gemini 2.5 Flash (Comanici et al., 2025).
4. DeepSeek V3 0324 (Liu et al., 2024).
5. GPT-4o mini (Hurst et al., 2024).
6. Phi-4 mini instruct (Abouelenin et al., 2025).

We used 20 different prompts for paraphrases using a zero-shot setup to promote diversity.

In the case of bias injection, we considered only one abstract scenario: polarity. In the prompt, we ask the models to rephrase the text from either a negative or positive point of view. Three models (Gemma3, Llama3, and Gemini) were used to generate negative perspectives, while the remaining models generated positive rewritings.

4 Results

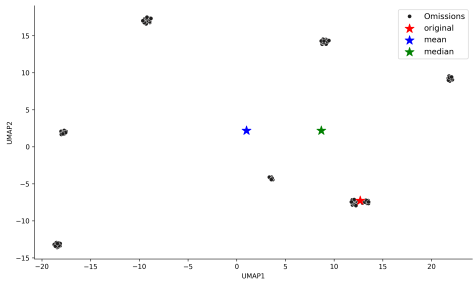
4.1 Omissions

Table 1 presents the average cosine similarities between the original text's latent representation and both the mean and median vectors. The median vector achieves a slightly higher similarity, and this difference is statistically significant (Mann–Whitney U test, $p < 0.05$).

Figure 1 visualizes 90 variations of the same text. Because it comprises only six sentences, only six unique variations (including the original non-variation) are displayed[1]. Unexpectedly, even in this constrained setting, the original representation does not lie exactly at the center of the resulting distribution.

[1]The visualization also includes the unaltered text.

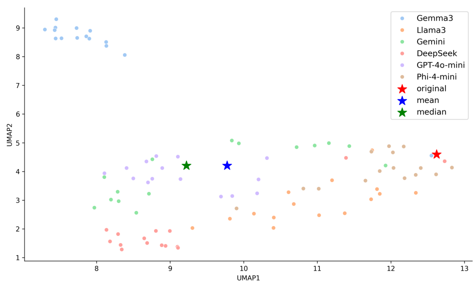
Figure 1: Visualization of variants of text (id: 0) using omissions.



4.2 Paraphrasing

Preliminary results with paraphrases (see figure 2) show that in some cases the original text might not be centered, and in the text drift might not diffuse in all directions as a classical physical system. Both the mean and median vectors are not the best estimators to the original text vector. In this particular case, the agent with less drift (Phi-4) was a best estimator.

Figure 2: Visualization of 90 variants of text (15 each model) using paraphrases and UMAP. In this particular case, the original text was generated with Qwen3-8B (Yang et al., 2025).



The quantitative analysis in the original latent space (see Table 2) indicates that, across all tested texts, both the mean and median vectors closely approximate the original representation. We observed significant differences between the mean/median vectors and every other model's outputs except those from Llama3 (Mann–Whitney U test, $p < 0.05$), while the mean and median themselves did not differ significantly (Mann–Whitney U test, $p = 0.6156$).

After reducing dimensionality to three via UMAP (cosine metric), texts generated by Llama3 exhibited even higher cosine similarities to the mean and median vectors (0.999935 and 0.999934, respectively) compared to 0.999721 in the original latent space. Figures 3 and 4 visualize specific cases with two dimensions.

Model	Similarity
Mean	0.9601
Median	**0.9609**
Gemma3	0.9290
Llama3	0.9540
Gemini	0.8454
DeepSeek	0.8467
GPT-4o-mini	0.8987
Phi-4-mini	0.8946

Table 2: Average cosine similarities from mean and median values to the original vector using paraphrases.

Figure 3: Visualization of variants of text (id: 3) using paraphrases.

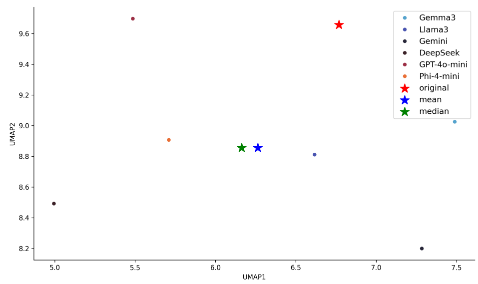
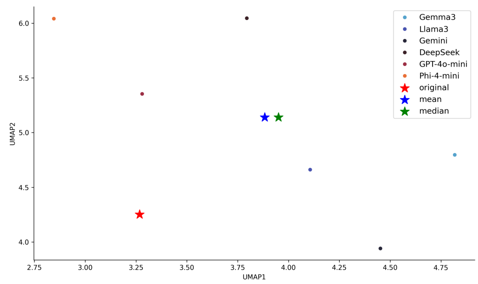


Figure 4: Visualization of variants of text (id: 16) using paraphrases.



4.3 Bias injection

In the bias induction experiments, the Phi-4 Mini model exhibited the least drift (see Table 3) and the highest average similarity compared to the centrality metrics, although its difference from the median vector was not statistically significant (Mann–Whitney U test, $p < 0.05$). We also observed significant differences between the mean and median vectors and all other groups, as well as between the mean and median themselves—except in the mean vs. GPT-4o Mini and median vs. Phi-4 Mini comparisons (Mann–Whitney U test,

Model	Similarity
Mean	0.9763
Median	0.9801
Gemma3	0.9291
Llama3	0.9447
Gemini	0.8699
DeepSeek	0.9382
GPT-4o-mini	0.9627
Phi-4-mini	**0.9807**

Table 3: Average cosine similarities from mean and median values to the original vector using bias injection.

$p < 0.05$). Figures 5 and 6 illustrate selected examples.

Figure 5: Visualization of variants of text (id: 22) using bias injection.

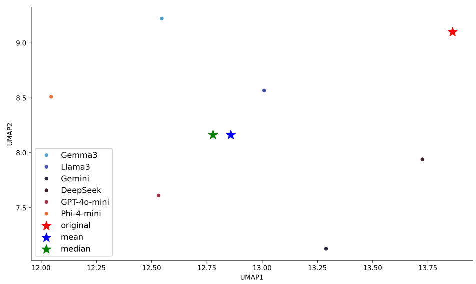
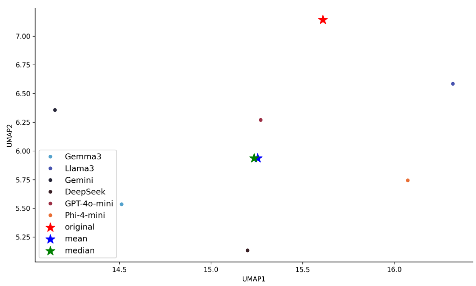


Figure 6: Visualization of variants of text (id: 44) using bias injection.



5 Discussion

This work constitutes a preliminary empirical approach to studying the process whereby a given text undergoes complex changes, in order to verify whether it is possible to retrieve information about the original from its immediate copies.

The main results discussed in section 4 provide a clear picture of the TIV simulation. In all cases, the median vector was a better estimator of the original vector than the mean. Both metrics outperformed most individual models. However, in some scenarios, we identified one model whose output drifted less and whose vector lay closer to the original than the median vector.

These findings offer two complementary insights. On one hand, the mean and median vectors can surpass most models in approximating the original text: by observing all surviving copies, we can reconstruct an approximation of the source, even in the presence of bias. On the other hand, it appears that one agent may introduce minimal variation, effectively serving as the best estimator of the original text. Identifying this agent may therefore be the optimal strategy for complex transmissions, in line with the Bédérian “best text” approach (Koppel et al., 2016).

The main results discussed in section 4 provide a clear picture of the TIV simulation. In all cases, the median vector was a better estimator of the original vector than the mean. Both metrics outperformed most individual models. However, in the paraphrasing and bias injection scenarios, we identified one model whose output drifted less and whose vector lay closer to the original than the median vector.

These findings offer two complementary insights. On one hand, the mean and median vectors can surpass most models in approximating the original text: by observing all surviving copies, we can reconstruct an approximation of the source, even in the presence of bias. On the other hand, it appears that one agent may introduce minimal variation, effectively serving as the best estimator of the original text. Identifying this agent may therefore be the optimal strategy for complex transmissions, in line with the Bédérian “best text” approach (Koppel et al., 2016).

6 Conclusions

This paper presents a one-step study of textual variation (ITV) across three scenarios: random textual changes (omissions), smooth semantic alterations (paraphrases), and bias injection (rephrasing with polarity). These scenarios correspond to three levels of textual alteration.

Studying complex changes in ITV is crucial for developing automatic methods for text reconstruction in contexts that go beyond purely random modifications while preserving text structure. These findings may apply to situations where textual variation is mediated by bias or rephrasing.

Our results show that basic central tendency statistics—particularly the median vector—are effective estimators of the original text vector. However, when paraphrasing occurs, certain agents introduce minimal drift and may serve as better estimators. This observation aligns with the classical philological proposals of Joseph Bédier.

6.1 Limitations

Using LLMs to simulate textual variation oversimplifies real-world scenarios. As discussed, LLMs do not act as individual agents but rather aggregate collective perspectives. Nevertheless, the growing role of artificial agents as (re)-writers clearly indicates that this study can be directly applied to understanding one aspect of material reality.

Acknowledgment

The authors would like to thank the anonymous reviewers.

Funding

This work has been possible thanks to the support of the Mexican government through SECIHTI, and the Instituto Politécnico Nacional research grants SIP 2259, SIP 20250015; IPN-COFAA and EDI.

Availability of data and materials

The data can be found in `https://github.com/Pherjev/ITV` or cited in the article.

Code and prompt availability

The code and prompts can be found in `https://github.com/Pherjev/ITV`

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used Microsoft's Copilot in order to improve writing in English, as it is not our native language. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. 2025. Phi-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-LoRAs. *arXiv preprint arXiv:2503.01743*.

Anthropic. 2025. Claude 3.5 haiku. Accessed: July 20, 2025.

Alberto Blecua. 1983. *Manual de Crítica Textual*. Castalia, Madrid.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv preprint arXiv:2507.06261*.

Antonino Ferraro, Antonio Galli, Valerio La Gatta, Marco Postiglione, Gian Marco Orlando, Diego Russo, Giuseppe Riccio, Antonio Romano, and Vincenzo Moscato. 2024. Agent-Based Modelling Meets Generative AI in Social Network Simulations. In *International Conference on Advances in Social Networks Analysis and Mining*, pages 155–170. Springer.

Walter M Fitch. 1971. Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. *Systematic Biology*, 20(4):406–416.

Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. 2023. S^3: Social-network Simulation System with Large Language Model-Empowered Agents. *arXiv preprint arXiv:2307.14984*.

Mees Gelein. 2021. *Simulating Mistakes. Using Agent Based Models to simulate and study the effects of scribal errors in classical text transmission*. Ph.D. thesis, Leiden University.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

James K He, Felix PS Wallis, Andrés Gvirtz, and Steve Rathje. 2024. Artificial intelligence chatbots mimic human collective behaviour. *British Journal of Psychology*.

John Healy and Leland McInnes. 2024. Uniform Manifold Approximation and Projection. *Nature Reviews Methods Primers*, 4(1):82.

Armin Hoenen. 2015. Lachmannian Archetype Reconstruction for Ancient Manuscript Corpora. In *Proceedings of the 2015 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1209–1214.

Armin Hoenen. 2018. From Manuscripts to Archetypes through Iterative Clustering. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Joop J Hox. 2017. Computational Social Science Methodology, Anyone? *Methodology*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. GPT-4o System Card. *arXiv preprint arXiv:2410.21276*.

Moshe Koppel, Moty Michaely, and Alex Tal. 2016. Reconstructing Ancient Literary Texts from Noisy Manuscripts. In *Proceedings of the Fifth Workshop on Computational Linguistics for Literature*, pages 40–46.

Grgur Kovač, Masataka Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, and Pierre-Yves Oudeyer. 2023. Large Language Models as Superpositions of Cultural Perspectives. *arXiv preprint arXiv:2307.07870*.

Maik Larooij and Petter Törnberg. 2025. Do Large Language Models Solve the Problems of Agent-Based Modeling? A Critical Review of Generative Social Simulations. *arXiv preprint arXiv:2504.03274*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. DeepSeek-V3 Technical Report. *arXiv preprint arXiv:2412.19437*.

Qun Ma, Xiao Xue, Deyu Zhou, Xiangning Yu, Donghua Liu, Xuwen Zhang, Zihan Zhao, Yifan Shen, Peilin Ji, Juanjuan Li, et al. 2024. Computational Experiments Meet Large Language Model Based Agents: A Survey and Perspective. *arXiv preprint arXiv:2402.00262*.

Benjamin S Manning, Kehang Zhu, and John J Horton. 2024. Automated Social Science: Language Models as Scientist and Subjects. Technical report, National Bureau of Economic Research.

Guilherme D Marmerola, Marina A Oikawa, Zanoni Dias, Siome Goldenstein, and Anderson Rocha. 2016. On the Reconstruction of Text Phylogeny Trees: Evaluation and Analysis of Textual Relationships. *PloS one*, 11(12):e0167822.

Alberto Bernabé Pajares and Felipe Hernández. 2010. *Manual de crítica textual y edición de textos griegos*, volume 32. Ediciones Akal.

Weihong Qi, Hanjia Lyu, and Jiebo Luo. 2025. Representation Bias in Political Sample Simulations with Large Language Models. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 1264–1267.

Teemu Roos, Tuomas Heikkilä, and Petri Myllymäki. 2006. A Compression-Based Method for Stemmatic Analysis. *Frontiers in Artificial Intelligence and Applications*, 141:805.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Andreas Koukounas, Nan Wang, and Han Xiao. 2024. jina-embeddings-v3: Multilingual embeddings with task lora.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 Technical Report. *arXiv preprint arXiv:2503.19786*.

Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Surabhi Adhikari, Hariram Veeramani, Mehwish Nasim, and Usman Naseem. 2025. Large language models (LLM) in computational social science: prospects, current state, and challenges. *Social Network Analysis and Mining*, 15(1):1–30.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*.

Darren Zammit. 2024. *Computational Stemmatology: Reconstructing Text Phylogenies through Computer Assisted Methods*. Ph.D. thesis, University of Groningen.

Xinnong Zhang, Jiayu Lin, Xinyi Mou, Shiyue Yang, Xiawei Liu, Libo Sun, Hanjia Lyu, Yihang Yang, Weihong Qi, Yue Chen, et al. 2025. SocioVerse: A World Model for Social Simulation Powered by LLM Agents and A Pool of 10 Million Real-World Users. *arXiv preprint arXiv:2504.10157*.

Versus: an automatic text comparison tool for the digital humanities

Tom Wainstain
Paris Cité University
Paris, France
tom.wainstain@etu.u-paris.fr

Motasem Alrahabi
ObTIC, Sorbonne University
Paris, France
motasem.alrahabi@sorbonne-universite.fr

Abstract

Digital humanities (DH) have been exploring large-scale textual reuse for several decades: quotation, allusion, paraphrase, translation, rephrasing. Automatic comparison, made possible by the increasing digitization of corpora, opens new perspectives in philology and intertextual studies. This article presents a state of the art of existing methods (formal, vector-based, statistical, graph-based) and introduces an open-source tool, Versus, which combines multigranular vector alignment, interactive visualization, and critical traceability. This framework aims to provide a reproducible and accessible solution for DH researchers, with support for text comparison in multiple languages.

1 Introduction

Since the works of (Kristeva, 1980) and (Genette, 1982), intertextuality has referred to a variety of textual reuses: quotation, plagiarism, allusion, paratext, etc. This structural dimension of text is central to philology, genetic criticism, comparative literature, as well as to history and textual linguistics. The growing availability of digitized corpora now paves the way for large-scale automated detection (Ganascia, 2020). However, existing methods face two major limitations: a high rate of false positives in large corpora, and a weak ability to detect semantic or allusive reuse. Overcoming these challenges requires the development of tools that combine lexical alignment, semantic modeling, and critical visualization.

2 Context and objectives

This work aims to examine recent approaches and propose open-source solutions tailored to the specific needs of researchers in the humanities and social sciences. It offers a structured overview of automatic comparison tools based on the following criteria: working principle, strengths, limitations, and representative tools for each approach (formal, vector-based, statistical, graph-based). It also presents Versus, an open and reproducible tool designed to meet the specific needs of DH researchers through an interactive interface and critical traceability of results. The contribution of this work is primarily system-oriented: Versus is presented as an open-source tool that integrates and adapts existing methods for the specific needs of digital humanities, rather than as an algorithmic advance.

3 Methods and tools for text comparison

Following prior surveys of text similarity methods (Nègre, 2013); (Wang and Dong, 2020); (Prakoso et al., 2021), we adopt a four-fold categorization of approaches, which has proved useful both in DH and general NLP contexts:

3.1 Formal approaches

These approaches compare texts directly at the level of characters, words, or n-grams, without modeling meaning. They rely on measures such as edit distances (Levenshtein (Levenshtein, 1966), Hamming (Hamming, 1950)) or similarity coefficients (Jaro (Jaro, 1989), Dice (Dice, 1945), Jaccard (Jaccard, 1901)), and are effective in detecting local or superficial similarities. They are particularly suited for word-for-word comparison or the analysis of fine textual variants. Among the tools based on a formal approach, Text-Pair[1] enables the detection of similar passages—such as quotations, borrowings, or common expressions—across large text collections using sequence alignment and shingling. CollateX[2], designed for philology, automatically aligns variants within a critical editing

[1] https://artfl-project.uchicago.edu/text-pair
[2] https://collatex.net/

Proceedings of the First Workshop NLP and Language Models for Digital Humanities associated with RANLP 2025, pages 32–37, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-106-6-003

framework. Medite[3] uses a suffix-tree and HMM-based algorithm to align two versions of a text by detecting deletions, insertions, replacements, and displacements, supporting critical editing and textual genetics.

Passim[4], for its part, is suited to detecting textual reuse in large corpora, combining speed and robustness. Finally, Diffchecker[5] provides a simple interface for line-by-line comparison, useful for quick checks or clear visualization of local divergences.

3.2 Vector-based approaches

These transform text into vectors to measure semantic similarity (cosine, Euclidean, etc.). They are generally robust to reformulations. Two main families of vector representations coexist: lexical representations, such as BOW or TF-IDF, which count word frequencies without considering context; and distributed representations, such as Word2Vec, GloVe, or BERT, which learn vectors from usage contexts to capture semantics. Among the tools based on a vector approach, spaCy similarity[6] offers fast measurement of semantic similarity between text units, based on built-in representations. Sentence-Transformers[7] generates robust sentence embeddings, well-suited for detecting reformulations and semantic alignment. LASER[8], developed by Facebook, provides multilingual representations for comparing texts across languages. Gensim[9] offers classic models like Word2Vec and Doc2Vec, effective for capturing lexical similarities in large corpora. Finally, SimAlign[10] combines lexical alignment and contextual embeddings to identify word-level correspondences, including in multilingual settings.

3.3 Statistical approaches

These methods leverage machine learning to identify patterns of similarity across texts. Supervised models—such as DSSM, ARC-I, or MV-LSTM—are trained on annotated examples to predict textual alignment or correspondence. Unsupervised techniques—such as LSA, LDA, or clustering—reveal latent structures without prior labeling, enabling thematic grouping, topic inference, or segment classification. These approaches are particularly effective for mapping global semantic proximities and visualizing the structure of large corpora in reduced vector spaces. In digital humanities, they support the exploration of textual traditions, discursive dynamics, and stylistic variation across time or authorship. Tools like Orange Text Mining[11] offer an accessible graphical interface for clustering and topic modeling. Scikit-learn[12] provides a robust suite of unsupervised algorithms for grouping high-dimensional representations. BERTopic[13], which combines transformer-based embeddings with dimensionality reduction and density-based clustering, enables the extraction of coherent and interpretable topics from heterogeneous or multilingual corpora.

3.4 Graph-based approaches

These approaches represent texts as networks of relationships (semantic, syntactic, discursive). They model links between words, sentences, or entities using knowledge graphs or graph neural networks (GNNs). While not all graph-based tools are designed specifically for text comparison, they can contribute to similarity analysis through their capacity to model textual structures that can then be compared using graph-based metrics such as structural comparison, node centrality analysis, and subgraph matching algorithms. These methods capture the global structure of the text and are effective for analyzing complex or multi-level connections, particularly in long or structured texts.

Among graph-based tools, textnets[14] represents collections of texts as networks of documents and words, enabling comparative analysis through network visualization and structural metrics. The tmtoolkit[15] from WZB offers comprehensive text mining and topic modeling capabilities with network analysis features that can support comparative analysis of semantic structures extracted from text corpora. Gephi[16], often used in combination with

[3] https://obtic.huma-num.fr/medite/
[4] https://programminghistorian.org/en/lessons/detecting-text-reuse-with-passim
[5] https://www.diffchecker.com/text-compare/
[6] https://spacy.io/usage/linguistic-features#vectors-similarity
[7] https://www.sbert.net/
[8] https://github.com/facebookresearch/LASER
[9] https://radimrehurek.com/gensim/
[10] https://github.com/cisnlp/simalign
[11] https://orangedatamining.com/
[12] https://scikit-learn.org/stable/modules/clustering.html
[13] https://maartengr.github.io/BERTopic/
[14] https://github.com/jboynyc/textnets
[15] https://github.com/WZBSocialScienceCenter/tmtoolkit
[16] https://gephi.org/

external text analysis tools such as TXM or custom pipelines, provides a powerful solution for visually exploring complex graphs derived from textual data, particularly for exploratory analysis. TextRank[17] applies the PageRank algorithm to lexical graphs and is useful for automatic keyword extraction or summarization within individual texts. Finally, discoursegraphs[18] enables the annotation and analysis of discursive and argumentative relations in texts using enriched directed graphs for multi-level annotated corpora.

These methods are particularly effective for analyzing textual structures, though dedicated text comparison typically requires additional algorithmic layers built upon these graph representations.

4 Synthesis

These four approaches offer complementary strategies for comparing texts, ranging from fine-grained variant detection to deep semantic modeling. The choice of method depends on the type of data, the desired level of granularity, and the goals of the analysis (reuse detection, alignment, thematic clustering, etc.).

5 Presentation of the Versus tool

Versus[19] is an open-source application dedicated to automatic text comparison, designed to meet the specific needs of DH researchers. It is based on methods that combine semantic vectorization, lexical weighting, and interactive visualization. Thanks to its reliance on multilingual sentence embeddings, Versus can be applied to texts in a variety of languages without requiring language-specific adaptation. Two main modules are currently implemented: comparison of a document with a corpus, and fine-grained comparison between two texts. The software architecture relies on a clear object-oriented model: each instance of Document is linked to a Text instance, which is further divided into Sentence and Word, enabling multi-level granularity. These documents are contained within a Corpus object. Text-to-text comparisons are handled by a PairText class. Shared variables are centralized in a Global_stuff class. The user interface is built with the Streamlit library, offering a modular structure for features and ensuring both accessibility for non-technical users and reproducible deployment beyond a simple notebook setting.

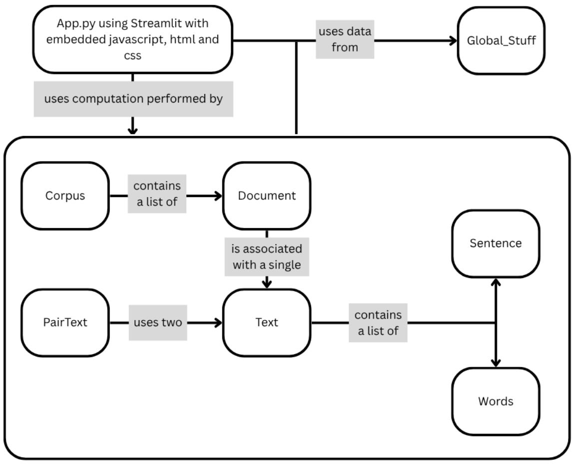


Figure 1: General Architecture of Versus

5.1 Document–Corpus Comparison

This module ranks an entire corpus based on similarity to a source document. To achieve this, Versus combines two complementary approaches: TF-IDF, used to weight the lexical importance of sentences according to their specificity within the corpus (contextual weight); Sentence Transformers (model all-MiniLM-L6-v2[20]), which generate dense 384-dimensional vector representations for each sentence. This model was selected for its balance between semantic performance, inference speed, and effectiveness on large corpora. The model operates on a general-purpose corpus (pretrained on diverse web/textual data), offering domain-agnostic performance without requiring domain-specific corpora or ontologies. Concretely, for each document, an embedding vector is assigned to each sentence using the transformer model. Each of these vectors is then weighted by the sum of the TF-IDF scores of its constituent words. This strategy gives greater value to sentences containing discriminative terms within the corpus.

5.2 Text–Text Comparison

This module is designed to identify similar passages between two texts. It uses a sliding segmentation into word n-grams, each vectorized using the same transformer model. Although the model is optimized for full sentences, it produces reliable

[17] https://github.com/summanlp/textrank
[18] https://github.com/arne-cl/discoursegraphs
[19] https://versuser-n5tyntby6aud5yryzwrdgf.streamlit.app/
[20] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

Approach	Principle	Strengths	Limitations	Example Tools
Formal	Direct comparison on strings (characters, tokens, n-grams)	Simple, precise, suited to local variants	Does not capture meaning, sensitive to reformulations	Text-Pair, CollateX, Medite, Passim, Diffchecker
Vector-based	Text representation as vectors (lexical or distributed)	Robust to reformulations, partial/contextual semantics	Loss of fine-grained info, requires training corpora	spaCy, Sentence-Transformers, LASER, Gensim, SimAlign
Statistical (ML)	Learning to group or match texts	Adaptable models, useful for classification/exploration	Needs annotated data (supervised), low interpretability	Orange, Scikit-learn, BERTopic
Graph-based	Texts represented as semantic/syntactic networks	Captures complex, multi-level structures, explicit links	Complexity, high computational cost	TextNetworkX, GraphText, Gephi, TextRank, Discourse Graphs

Table 1: Comparison of text analysis approaches

vectors for word groups, including reformulated or reordered segments. The comparison is based on a cosine similarity matrix between all n-grams of both texts. To manage memory usage, the matrix is built segment by segment and converted into a sparse structure by filtering out scores below a threshold p. This ensures controlled memory usage even for large texts. The segmentation divides the second text into blocks of k columns, avoiding the creation of a full matrix, which would be memory-intensive. At each iteration, only a partial section of the matrix is computed, filtered, and converted to a sparse format, then concatenated with previous segments. This approach allows efficient processing of very large texts while limiting RAM usage. This method has proven up to 7 times faster than traditional text comparison using optimized libraries like Rapidfuzz[21], while remaining sensitive enough to detect inflected or allusive correspondences.

5.3 Interactive Visualization

The application is built on Streamlit[22] and offers an accessible interface structured into four sections: corpus management, ranking, comparison, and user guide. Users can adjust parameters (n-gram size, similarity threshold, stopword activation) and visualize detected segments in an aligned and annotated format. Results include dynamic highlighting of correspondences, difference visualization (via the Difflib[23] library), and direct interaction with the source text, ensuring readability and transparency. Correspondence with the original text is maintained through position metadata (start and end) associated with each word, allowing accurate display even when stopwords are removed. An embedded JavaScript script enables dynamic adjustment of the textual context size around matched passages. Ongoing developments include support for export in CSV and TEI formats.

5.4 Use Case Scenario

A typical use case involves comparing a source document to a collection of documents. The tool ranks the collection based on a similarity score computed from a weighted average of sentence vectors (TF-IDF + Sentence Transformers). The user selects a document, which is then aligned with the source document. A sliding n-gram segmentation detects similar passages using a cosine similarity matrix. The interactive interface displays the aligned texts side by side, highlights the detected correspondences, allows parameter adjustment (n-gram size, threshold, stopwords), and provides dynamic, annotated visualization of similarities and differences.

6 Evaluation

This evaluation aims to illustrate the performance and usability of the Versus tool in a realistic context of intertextual analysis.

- Quantitative: On a sample of 5 manually annotated text pairs (50 target alignments), Versus achieves an average precision of 0.86, recall of 0.79, and F1-score of 0.82. False positives mainly involve borderline reformulations or contextually ambiguous segments.
- Qualitative: The detected alignments are largely considered relevant, including in cases of reformulation or allusion. Similar segments

[21] https://rapidfuzz.github.io/RapidFuzz/
[22] https://streamlit.io/
[23] https://docs.python.org/3/library/difflib.html

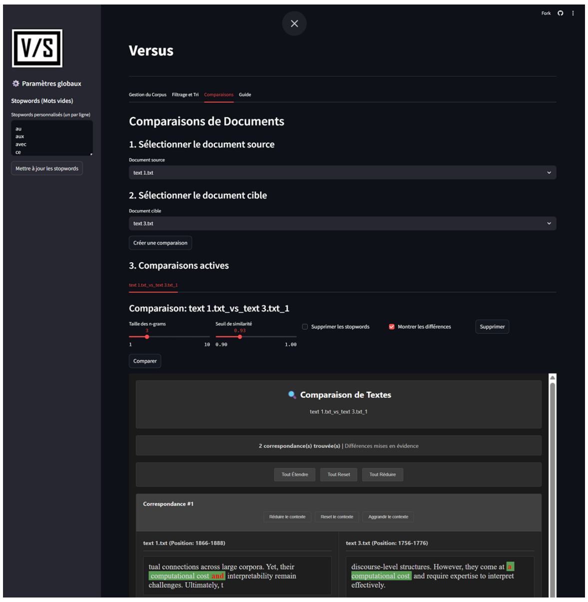


Figure 2: Versus User Interface

are well localized and clearly visualized, facilitating philological analysis. While tailored to DH use cases, future evaluation may consider adapting standard benchmarks such as STS-B or MSRP to literary or historical datasets.

- Ergonomic: The Streamlit interface is considered intuitive. Parameters (n-gram size, threshold, stopwords) are easily adjustable. Result export and aligned visualization provide strong support for critical analysis.

We acknowledge that this evaluation is limited in scope, relying on only five annotated pairs and lacking both baselines and error analysis, which will be addressed in future work.

7 Limitations

This work has several limitations. First, the evaluation is based on a small proof-of-concept dataset (5 annotated pairs), without systematic baselines or detailed error analysis, which restricts the strength of empirical claims. Second, Versus currently relies on transformer-based embeddings, which can be sensitive to noisy input such as OCR errors, typos, or unsupported languages. These constraints, already noted in the evaluation and conclusion, underline the need for broader benchmarking and methodological refinement, as outlined in the Perspectives section.

8 Conclusion and Perspectives

Text comparison is a central challenge in digital humanities. Versus offers a hybrid and accessible approach, combining lexical precision, semantic modeling, and critical visualization. Designed for digital humanities, it addresses key challenges in the field: processing large corpora, detecting various types of reuse (quotation, allusion, paraphrase, reformulation), ensuring result readability, and providing direct interpretive support for DH scholars without technical expertise.

By leveraging deep learning and lexical statistics techniques (transformers + TF-IDF), it enables efficient multigranular alignment, suited to the linguistic variation typical of literary and historical texts. However, this method has limitations: the transformer model relies on prior understanding of words. If the input is noisy (OCR errors, typos, unsupported languages), the resulting embeddings may be unrepresentative, reducing comparison accuracy.

As an open, reproducible, and modular tool, Versus provides a solid foundation for contemporary intertextual analysis. Unlike general NLP approaches focused primarily on model performance, our methodology emphasizes the specific requirements of digital humanities, integrating critical traceability—through alignment metadata, visualization of correspondences, and direct linkage to source texts—together with interactive visualization and accessibility for non-technical users.

Planned extensions include diachronic alignment and broader multilingual coverage, enabling cross-linguistic analysis within a unified framework. A pilot study on 18th- and 19th-century French literary texts will explore influence through paraphrase detection, forming part of a broader validation with DH scholars to assess usability, interpretive value, and alignment with philological practices. Future work will also address systematic benchmarking through larger evaluation sets and comparisons with established baselines, to better situate Versus within the state of the art.

References

Lee R. Dice. 1945. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

Jean-Gabriel Ganascia. 2020. *Les humanités numériques*. CNRS Éditions.

Gérard Genette. 1982. *Palimpsestes: La littérature au second degré*. Seuil.

Richard W. Hamming. 1950. Error detecting and error correcting codes. *Bell System Technical Journal*, 29(2):147–160.

Paul Jaccard. 1901. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:547–579.

Matthew A. Jaro. 1989. Advances in record-linkage methodology as applied to the 1985 census of tampa florida. *Journal of the American Statistical Association*, 84(406):414–420.

Julia Kristeva. 1980. *La révolution du langage poétique*. Seuil.

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

Élsa Nègre. 2013. Le traitement automatique du texte pour les sciences humaines. *Traitement Automatique des Langues*, 54(1):135–152.

Dimas Wibisono Prakoso, Asad Abdi, and Chintan Amrit. 2021. Short text similarity measurement methods: a review. *Soft Computing*, 25(6):4699–4723.

Shuai Wang and Wei Dong. 2020. A survey on text similarity techniques. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 11(10):12–25.

Like a Human? A Linguistic Analysis of Human-written and Machine-generated Scientific Texts

Sergei Bagdasarov
Saarland University (Germany)
sergeiba@lst.uni-saarland.de

Diego Alves
Saarland University (Germany)
diego.alves@uni-saarland.de

Abstract

The purpose of this study is to analyze lexical and syntactic features in human-written texts and machine-generated texts produced by three state-of-the-art large language models: GPT-4o, Llama 3.1 and Qwen 2.5. We use Kullback-Leibler divergence to quantify the dissimilarity between humans and LLMs as well as to identify relevant features for comparison. We test the predictive power of our features using binary and multi-label random forest classifiers. The classifiers achieve robust performance of above 80 % for multi-label classification and above 90 % for binary classification. Our results point to substantial differences between human- and machine-generated texts. Human writers show higher variability in the use of syntactic resources, while LLMs score higher in lexical variability.

1 Introduction

The use of Large Language Models (LLMs) in research has become a common practice. Scenarios in which academia members resort to LLMs are varied, ranging from ideas generation and productivity enhancement to data analysis and writing (Panda and Kaur, 2024). 80.88% of researches surveyed by Liao et al. (2024) used LLMs in their academic activities, with 61% of researchers having at least once used LLMs for editing and 41%, for direct writing. A vast majority of scholars surveyed by Mishra et al. (2024) consider that LLMs will have an impact on various stages of the publication process.

State-of-the-art LLMs are capable of producing high-quality texts that are practically indistinguishable from human-created content for untrained individuals. This makes LLMs useful writing assistants, especially for researchers who are not native speakers of English. Yet numerous studies based on large enough amounts of data have shown that LLMs do write differently in comparison to humans according to certain measures.

In this study, we aim to analyze human-written texts (HWT) and machine-generated texts (MGT) – abstracts of academic papers – and identify linguistic features that can help tell them apart. While studies on this topic abound, they either do not focus specifically on academic texts, even though those might be present in the scrutinized corpora, or use the older GPT 3.5 model relying on the paper title and a short text snippet for abstract generation.

We will address these research gaps by using a large dataset of academic publications with full texts and human-written abstracts and resorting to a newer GPT-4o model (OpenAI, 2024). Moreover, we will complement our analysis with two open-source state-of-the-art models (Llama 3.1 8B Instruct (Grattafiori et al., 2024) and Qwen 2.5 7B Instruct (Yang et al., 2025; Team, 2024)). The rationale behind this decision is that, although ChatGPT is the undisputed leader as a chat bot assistant, its use may be associated with data protection concerns (Ali et al., 2025; Novelli et al., 2024). Because of this, researchers might choose open-source LLMs, either running them locally or accessing them through in in-house university chat bot solutions. Therefore, the amount of scientific content potentially generated by open-source models might be increasing.

Furthermore, most studies comparing HWT and MGT use a predefined list of features. Instead, we will rely on Kullback-Leibler divergence, a measure rooted in information theory, to identify features that can reliably distinguish between HWT and MGT and then test these features in a classification task.

The remainder of the paper is structured as follows. Section 2 offers a brief overview of research on linguistic features in HWT and MGT. Section 3 describes the dataset and methodology, including

Proceedings of the First Workshop NLP and Language Models for Digital Humanities associated with RANLP 2025, pages 38–47, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-106-6-004

the procedure for feature selection. Then, Section 4 presents the results for text classification, showing the predictive power of the extracted features. Section 5 compares HWT and MGT across some of the selected features, while Section 6 contains a brief discussion of our findings. Finally, Section 7 offers some concluding remarks and an overview of future work plans.

2 Related Work

Due to easy accessibility and outstanding output quality, LLMs have become an integral part of many workflows, often being used to produce written content. This inevitably leads to the proliferation of machine-generated texts, making the study of synthetic language an important task.

A common approach for this consists in defining a set of features (e.g., sentence length, frequencies of words, part-of-speech categories or specific syntactic patterns, etc.) and comparing them in human-written texts (HWT) and machine-generated texts (MGT) (e.g., Zanotto and Aroyehun (2024); Culda et al. (2025); Muñoz-Ortiz et al. (2024); Georgiou (2024), etc.).

A general consensus in this field is that MGT differ considerably from HWT, especially when it comes to lexical variability, with human writers being characterized by higher lexical diversity (e.g. as measured by type-token ratio or amount of hapax legomena) (Zanotto and Aroyehun, 2024; Culda et al., 2025). At the same time, Opara (2024) pointed out that HWT manage to strike a balance between lexical richness and text length, while some LLMs seem either to overly restrict or expand their vocabulary.

Apart from that, LLMs have also been shown to overuse some stylistic vocabulary not related to the content of a text. Such overused lexical items, sometimes referred to as focal words, can be detected by comparing word frequencies before and after LLM era and typically include words like *delve, underscore, intricate, pivotal, showcase, meticulous*, etc. (Juzek and Ward, 2024; Kobak et al., 2025; Gray, 2024; Liang et al., 2024).

From the morphosyntactic perspective, MGT seem to use shorter sentences, showing however higher complexity in the constituency structure (Muñoz-Ortiz et al., 2024). MGT (at least those generated by GPT models) also tend to favor a more nominal style of writing, with higher proportion of nouns, nominalizations, phrasal coordination and determiners (Reinhart et al., 2025; Liao et al., 2023). Despite higher proportion of nouns, LLMs were found to use more general vocabulary in specialized registers, though, resulting in lower degree of specificity and higher readability scores in comparison to HWT (Liao et al., 2023).

Interestingly, HWT tend to convey more negative emotions on average, while LLMs produce more positive texts (Muñoz-Ortiz et al., 2024; Culda et al., 2025). This also holds for conversation-like texts, where LLMs exceed humans in some communicative processes, scoring higher in social behavior, politeness and attentional focus. However, they still do not reach the same level of authenticity as humans, at least in conversation (Sandler et al., 2024).

The studies reviewed above rely on a predefined set of features to explore differences between HWT and MGT. In contrast, in this study we introduce a more informed approach to feature selection based on Kullback-Leibler divergence. Moreover, we use three syntactic complexity measures, which, to the best of our knowledge, has not been done yet.

3 Data and Methods

3.1 Data

We use the ACL Anthology Corpus (Rohatgi, 2022) – a collection of ACL contributions ranging from 1950s to 2022. The reasons for choosing this dataset were twofold. First, it provides both abstracts and full texts, allowing us to generate abstracts based on full papers and compare them to HWT. Moreover, since the dataset is limited to publications prior to 2022, we ensure that no abstracts in it have been written by LLMs.

Due to the extensive size of the corpus, which would result in high material and computational costs, and the presence of noisy data, we selected a sample of papers that meet the following criteria: a) both full text and abstract are available; b) publication year: after 1999; c) language: English; d) length of the abstract: between 100 and 200 words; e) length of the full paper: only those within one standard deviation of the mean length among those in the interquartile range. After applying the filters, we obtained a subset of 10,393 papers with their abstracts.

3.2 Automatic Abstract Generation

We automatically generated abstracts based on the ACL papers using three LLMs: **gpt-4o-2024-**

08-06 (OpenAI, 2024), **Llama-3.1-8B-Instruct** (Grattafiori et al., 2024) and **Qwen2.5-7B-Instruct** (Yang et al., 2025; Team, 2024). The GPT model was prompted using the OpenAI API through the openai Python library[1]. To interact with the two open-source models, we used the huggingface[2] transformers text generation pipeline. All models were prompted using the temperature of 1. Except for output length, all other parameters were kept at default values. Number of output tokens was set to 400. Lower values approximating the desired abstract length of 200 words stipulated in the prompt resulted in a considerable amount of incomplete outputs by Llama. While the higher output length in the pipeline parameters helped mitigate the incomplete output problem, it caused Llama output to be longer as can be seen in Table 1. However, this was not critical for further analysis since all measures are normalized by the size of the subcorpora.

The prompt consisted both of a system message defining the model's role as academic writing assistant and describing the task as well as a user message providing the full text of a paper and giving the instruction to generate an abstract (see Appendix A). We included the full text to approximate the behavior users might exhibit when actually using the models for paper summarization since newer models feature much larger input context windows. The experiments were run at the end of June and at the beginning of July 2025.

Source	Tokens	Types	Sentences
Human	1,700,972	32,808	64,975
Llama	2,392,988	34,688	83,534
Qwen	1,524,521	29,388	60,018
GPT	2,034,978	32,880	76,710

Table 1: Comparison of abstract sources by tokens, types, and sentences.

3.3 Methods

We used **Kullback-Leibler Divergence (KLD)** (Kullback and Leibler, 1951) to compare HWT and MGT. KLD (Equation 1) is an information-theoretic measure that asymmetrically quantifies (in bits) the divergence between two probability distributions and allows us to identify the most distinctive features contributing to the divergence. A KLD value of 0 means that the distributions are identical, whereas a value larger than 0, in contrast, is indicative of a divergence.

We calculated KLD for (a) lemmas to capture how HWA and MGA diverge on the lexical level as well as (b) Universal Dependencies part-of-speech tags, and (c) dependency relations (de Marneffe et al., 2021) to analyze the syntactic differences. For the classification task, we used only POS tags and syntactic relations for which we calculated normalized document frequencies by dividing the count of a POS tag or dependency label by the total number of tokens in the document.

$$\mathrm{KLD}(P \parallel Q) = \sum_{x} P(x) \log \frac{P(x)}{Q(x)} \tag{1}$$

We complemented this initial set of features with variability measures since previous research has found substantial differences in variability of different elements in HWT and MGT (e.g.,Zanotto and Aroyehun (2024); Culda et al. (2025); Liao et al. (2023); Opara (2024), etc.). We operationalize variability with **Entropy**. Entropy is another information-theoretic measure and is calculated using the formula in Equation 2. It shows how much uncertainty there is in a system, with higher entropy values indicating higher uncertainty. In our context, higher uncertainty means more variability in language use. We calculated document-level entropy for lemmas, POS tags and dependency relations.

Additionally, we calculated the proportion of unique items in each selected POS category by dividing the number of unique items in a POS category by the total number of elements in the category.

$$H = \sum_{i=1}^{n} p_i \times \log_2 p_i \tag{2}$$

We also included three syntactic measures that are commonly analyzed when examining syntactic complexity: average dependency length (Gibson, 1998; Jiang et al., 2019), tree depth (Xu and Reitter, 2016), and average branching factor (Xu and Reitter, 2016).

Average dependency length (aDL) is calculated by measuring the distance between heads and their dependents in a syntactic dependency tree, ignoring punctuation. For each dependency in a sentence, the length is the absolute difference between

[1] https://pypi.org/project/openai/
[2] https://huggingface.co/

the positions (i.e. token indices) of the head and the dependent. These lengths are summed across all dependencies in the sentence and then divided by the total number of dependencies (i.e.number of tokens minus one). For example, in the parsed sentence with 8 tokens (excluding punctuation) shown in Figure 1, the sum of the dependency distances is 17. The average dependency length (ADL) is therefore calculated as $\frac{17}{8-1} = 2.43$.

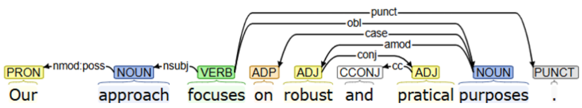


Figure 1: Example of parsed sentence.

Average branching factor (ABF) quantifies the mean number of immediate dependents (children) per internal node. It captures how syntactic constituents are organized: higher ABF values indicate greater syntactic parallelism, whereas lower values suggest more linear or sequential structuring. It is calculated by dividing the total number of children of internal nodes (i.e., the total number of nodes minus one) by the number of internal nodes.

Tree depth (TD) refers to the length of the longest path from the root node to any leaf node in a syntactic dependency tree, reflecting the degree of structural nesting.

As an example, consider the sentence "Anne lost control and laughed." Its constituency tree is: (ROOT (S (NP (NNP Anne)) (VP (VP (VBD lost) (NP (NN control))) (CC and) (VP (VBD laughed))) (. .))). The graphical representation of this tree is provided in Figure 10 in Appendix B.

The total number of nodes (internal nodes and leaves) is 19, the longest path from the root to any leaf (tree depth) is 6 (ROOT $\rightarrow$ S $\rightarrow$ VP $\rightarrow$ VP $\rightarrow$ NP $\rightarrow$ NN $\rightarrow$ leaf), and the average branching factor is the total number of children of internal nodes is 18 (i.e., total nodes minus 1) divided by the number of internal nodes (i.e., 13), giving an average branching factor of 1.39.

For each abstract (human- or machine-generated), these measures were calculated for each sentence and then divided by the number of sentences to obtain the average value per text.

Finally, we also included word and sentence length, average count of stop words per text and a readability measure[3]. Entropy, dependency relations counts and the three syntactic measures were calculated by custom Python scripts. For all other measures, we used the Linguistic Features ToolKit (LFTK) (Lee and Lee, 2023). In total, we obtained a set of 76 features.

We trained random forests classifiers both for binary and multi-label prediction to test the predictive power of the selected features. First, we only used the features extracted by KLD and then the complete set of features. To train the classifiers, we used the ranger[4] R package with the following settings: number of trees = 500, importance = impurity, classification = True, seed =123. For both classifiers, we used 80% of data for training and 20% for testing.

4 Classification Results

We first focused only on POS tags and dependency relations that have proven distinctive as per KLD (53 features in total). Then we used the full set of 76 features. As shown in Figure 2, these 76 features do allow to identify 4 clusters of texts. We can see an overlap between GPT and Qwen, on the one hand, and humans and Llama, on the other hand, indicating a greater linguistic similarity of these groups, in line with our KLD results. In turn, the clusters for GPT and Qwen are clearly distinguishable from human texts.

	Binary		Multilabel	
Model	**Acc.**	**F1**	**Acc.**	**F1**
NIR	.75	–	.25	–
KLD features only	.92	.89	.83	.83
All features	.94	.92	.88	.88

Table 2: Random forests results for binary (human vs machine) and multi-label (human vs GPT vs Llama vs Qwen) classification. NIR (no-information rate) shows the model's performance if the majority label is always assigned. Here it is used as baseline.

	Precision	**Recall**	**F1**
GPT	.87	.90	.88
Human	.89	.90	.89
Llama	.89	.87	.88
Qwen	.89	.86	.87

Table 3: Precision, recall, and F1-scores for each class in Random Forest classification based on the full set of features.

Using the two sets of features (the initial one obtained by KLD only and the complete one), we trained random forests classifiers both for binary

[3]Coleman-Liau Index (Coleman and Liau, 1975).

[4]https://cran.r-project.org/package=ranger

and multi-label prediction. As shown in Table 2, even the initial set of features allows for a robust predictive power, confirming that KLD is a suitable measure to identify relevant linguistic features. However, additional features do improve the model performance considerably. The prediction results are fairly similar across different classes, being slightly better for human texts and worse for Qwen-generated texts as measured by F1 score (see Table 3).

5 Feature Analysis

5.1 KLD

In line with previous findings, our KLD analysis indicated that HWT differ from MGT both lexically and syntactically (see Figures 3, 4 and 5). Llama was closest to humans on all three levels of comparison, while GPT and Qwen turned out to be more divergent from HWT, showing similar results.

The distinctive features extracted by KLD were very similar across all groups of comparison. On the lexical level, HWT are mostly characterized by function words (determiners, prepositions, particles, copula verbs and modal verbs), discourse markers (*however, therefore, thus*), adverbs (*usually, considerably*) and some abbreviations commonly used in academic writing (*i.e., e.g., etc.*). In contrast, MGT are characterized by the use of nouns, verbs and adjectives many of which have stylistic function and are considered "focal words" typically overused by LLMs (*highlight, underscore, demonstrate, introduce, incorporate, pivotal, significant*, etc.) (see Table 5). The verb *delve*, a prototypical example of such overused vocabulary, has also been identified as distinctive of all tested LLMs, especially the GPT model. However, its contribution to the overall divergence is relatively low, suggesting that it is not as overused anymore by GPT-4o as it was by GPT 3.5.

H x GPT	H x Llama	H x Qwen
be	be	we
the	this	be
of	in	of
a	paper	have
we	have	our

Table 4: Top 5 lemmas distinctive of humans compared to each of the LLMs.

KLD results for POS tags confirm the more extended use of function words in HWT observed at the lexical level (see Table 6). Besides, HWT are also characterized by the use of adverbs, proper nouns and numerals. Interestingly, adjectives are distinctive of HWT when compared to Llama, however, not when compared to GPT and Qwen. MGT

GPT x H	Llama x H	Qwen x H
enhance	and	and
future	our	author
and	approach	enhanced
highlight	include	demonstrate
potential	demonstrate	highlight

Table 5: Top 5 lemmas distinctive of each of the LLMs compared to humans.

H x GPT	H x Llama	H x Qwen
AUX	AUX	PRON
DET	ADP	AUX
ADP	ADV	ADP
PRON	DET	DET
ADV	ADJ	ADV

Table 6: Top 5 UD POS tags distinctive of humans compared to each of the LLMs.

GPT x H	Llama x H	Qwen x H
NOUN	NOUN	NOUN
VERB	CCONJ	PUNCT
ADJ	VERB	VERB
PUNCT	PUNCT	PROPN
PROPN	SCONJ	CCONJ

Table 7: Top 5 UD POS tags distinctive of each of the LLMs compared to humans.

H x GPT	H x Llama	H x Qwen
det	advmod	case
case	obl	obl
cop	case	cop
aux:pass	det	aux
advmod	aux:pass	advmod

Table 8: Top 5 UD relations distinctive of humans compared to each of the LLMs.

GPT x H	Llama x H	Qwen x H
obj	obj	obj
advcl	nmod:poss	punct
compound	conj	compound
amod	compound	advcl
punct	cc	amod

Table 9: Top 5 UD relations distinctive of each of the LLMs compared to humans.

Figure 2: t-SNE clustering of HWT and MGT based on a set of 76 features.

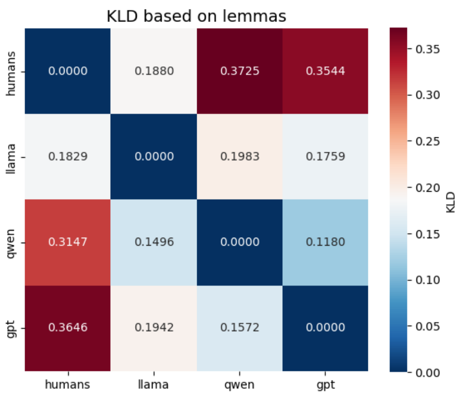


Figure 3: KLD values based on lemmas.

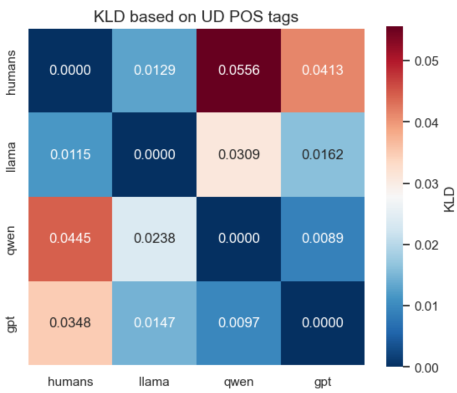


Figure 4: KLD values based on UD POS tags.

are characterized by POS tags labeling nouns and verbs as well as punctuation marks, coordinative conjunctions and subordinative conjunctions (see Table 7).

At the level of UD syntactic relations, we again see that one of the most distinctive features of HWT

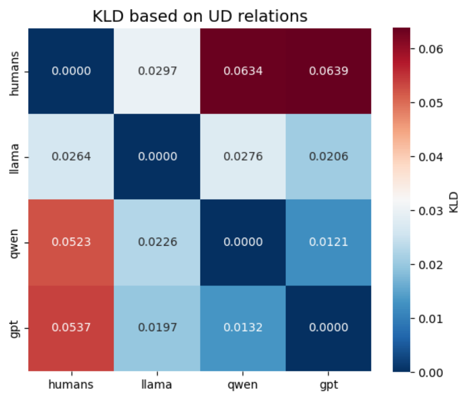


Figure 5: KLD values based on UD dependency relations.

are auxiliaries, determiners and adverbial modifiers. The *obl* and *case* labels point to a varied use of prepositional phrases as oblique arguments, adjuncts or nominal modifiers. LLMs, in turn, seem to favor more compact and dense constructions, at least in case of nominal modification, leading potentially to higher phrasal complexity. This is evident by adjective modifiers, compound nouns and possessive nominal modifiers being distinctive of MGT (see Table 9).

In terms of clausal features, HWT are especially characterized by passive constructions and finite relative clauses modifying either nouns or whole sentences. In contrast, LLMs tend to use more non-finite clausal modifiers (with the exception of Llama), adverbial clauses and clausal complements. In general, clausal subordination is more typically seen in MGT, which is reflected in the more pronounced use of the relation *mark*.

5.2 Word and Sentence Length

MGT contain longer words than HWT as measured by the number of syllables, which is in line with the overall prevalence of function words in HMW in comparison to MGT as indicated by KLD. In terms of sentence length (in words), Llama used the longest sentences as per median value. If measured in syllables, all LLMs used longer sentences because of a consistently longer word length. However, human abstracts show greater variability in sentence length, while all LLMs, especially Qwen, consistently produced sentences containing between approximately 20 and 40 words. (see Figure 7).

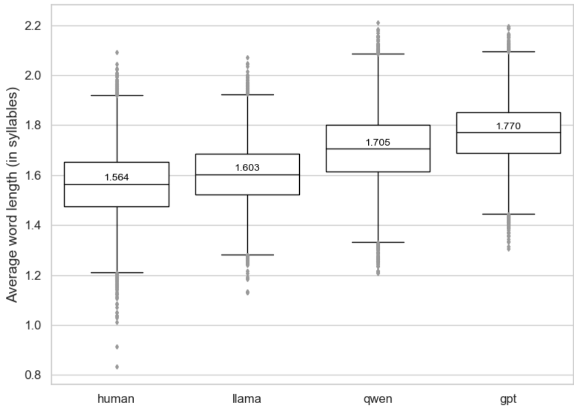


Figure 6: Average word length per document (in syllables).

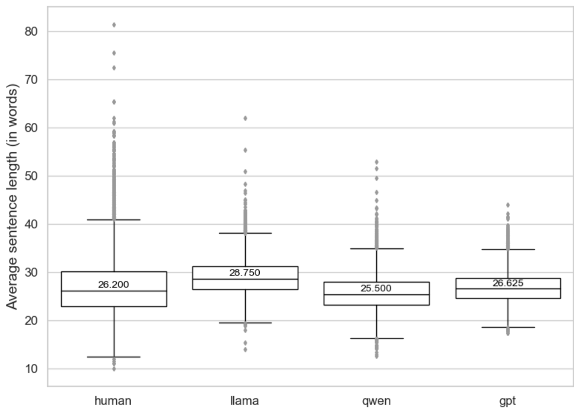


Figure 7: Average sentence length per document (in words).

5.3 Variability

At the level of lemmas, GPT shows the highest entropy, followed by Llama. Humans and Qwen are similar in terms of entropy. At the level of POS tags and dependency relations, humans have higher entropy than all LLMs, with Llama-generated texts being closest to HWT. In contrast, Qwen shows the lowest variability among the three LLMs across all three levels of comparison.

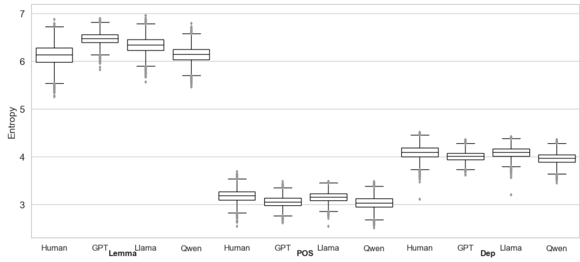


Figure 8: Document entropy for lemmas, UD POS tags and dependency relations.

5.4 Syntactic Complexity

Figure 9 shows the Kernel Density Estimation (KDE) analysis of the three syntactic measures considered in this study: tree depth, average branching factor, and average dependency length (each averaged per abstract). This method estimates the probability density function of a continuous variable, providing a smooth curve that represents the data distribution.

We observe different behaviours for each measure. GPT and Qwen tend to produce sentences with lower tree depth and average dependency length compared to humans, while Llama generates sentences with higher complexity according to these two measures. However, in terms of the average branching factor, all LLMs tend to produce sentences that exhibit greater syntactic branching.

One characteristic of scientific English is the frequent use of complex noun phrases, including pre-modifiers and compound constructions (Halliday and Martin, 2003; Degaetano-Ortlieb, 2021). These complex nominal phrase structures increase the ABF of sentences because the pre-modifiers are all at the same hierarchical level in the syntactic tree (i.e., children of the same NP node). Thus, it seems that language models tend to potentialize the usage of this type of NP.

However, the most striking characteristic of human abstracts compared to machine-generated ones is their greater variability in syntactic complexity. This is evident in the density plots, where human texts consistently cover a broader range with a less pronounced peak, while machine-generated texts exhibit much less variation.

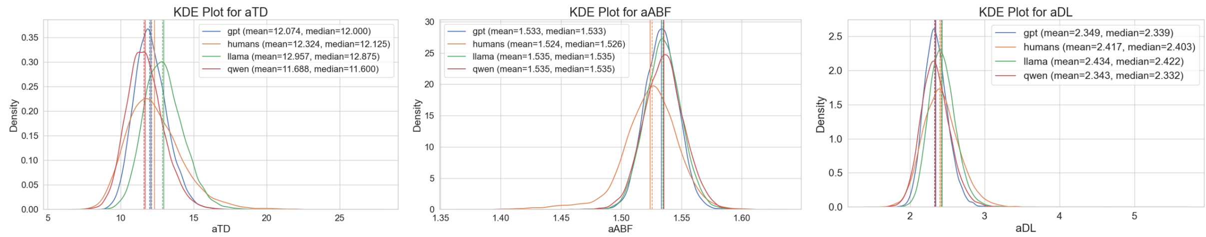


Figure 9: KDE plots of tree depth, average branching factor, and average dependency length (averaged per abstract).

6 Discussion

In general, our findings stand in line with previous research suggesting that HWT and MGT differ considerably both in terms of their lexical and syntactic features (cf. Culda et al. (2025); Zanotto and Aroyehun (2024); Georgiou (2024)). Given that GPT was considerably larger than the other two models and, therefore, was exposed to more training data, we expected the GPT model to produce more human-like content. Nevertheless, Llama was closest to humans both lexically and syntactically, at least as measured by KLD.

LLMs tend to use more complex phrasal structures, which is reflected in a higher average branching factor and a higher typicality of nominal premodifiers. Similar findings for other text registers and models (Muñoz-Ortiz et al., 2024; Reinhart et al., 2025) suggest that higher phrasal complexity is a general feature of LLM writing and is not attributable to the influence of our experimental setup.

Also in line with previous research, we have shown that LLMs exhibit lower morpho-syntactic variability (lower entropy of UD POS tags and syntactic relations as well as a narrower spread of values for syntactic complexity measures). This indicates a more repetitive use of patterns as opposed to a more varied use of syntactic resources by human writers.

Surprisingly, we found that MGT exhibit higher lexical variability than HWT, as measured by lemma entropy. This may suggest a general advancement in the lexical creativity of newer models. However, it could also be a consequence of our prompting settings — for instance, the models' exposure to full papers during abstract generation. Further analysis is needed to better understand this outcome.

7 Conclusion and Future Work

In this study, we analyzed the lexical and syntactic features of HWT and MGT. We obtained HWT from a large dataset of academic publications. MGT were generated by three state-of-the-art models (GPT-4o, Llama 3.1 and Qwen 2.5) based on the corresponding full papers.

Adopting a data-driven approach to feature selection, we employed KLD to identify features that effectively distinguish between HWT and MGT. The effectiveness of these features was evaluated using random forest classifiers, which demonstrated robust performance both when using only KLD-selected features and when incorporating an extended feature set.

Our results indicate that MGT still differ considerably from HWT, with Llama producing outputs that more closely resemble HWT than other LLMs. We observed that LLMs tend to generate more complex phrase structures than humans, yet exhibit less syntactic variability. In contrast, lexical variability as measured by entropy was higher in MGT.

Future research will explore whether alternative prompting configurations (e.g., varying temperature settings or employing few-shot prompting) lead to more human-like outputs. We also plan to extend our analysis to additional text types, models, and model sizes. Additionally, a more qualitative examination of LLM-generated abstracts and their perception by human readers would give us a more complete understanding of models' performance.

Limitations

LLM output is strongly influenced by the input data and prompt wording. Experiments based on other genres or using other prompting techniques might yield different results. Moreover, the dataset is limited to one scientific discipline (computational linguistics). So, we cannot account for linguistic divergence between different domains of scientific writing. Our findings might become outdated due to constant model improvement and release of more powerful LLMs. Finally, although KLD has proven to be an effective feature selection technique, we may have missed some other relevant features that cannot be identified using KLD.

Acknowledgements

This research is funded by *Deutsche Forschungsgemeinschaft* (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

References

Mutahar Ali, Arjun Arunasalam, and Habiba Farrukh. 2025. Understanding users' security and privacy concerns and attitudes towards conversational ai platforms. In *2025 IEEE Symposium on Security and Privacy (SP)*, page 298–316. IEEE.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

L. C. Culda, R. A. Nerişanu, M. P. Cristescu, D. A. Mara, A. Bâra, and S. V. Oprea. 2025. Comparative linguistic analysis framework of human-written vs. machine-generated text. *Connection Science*, 37(1).

Stefania Degaetano-Ortlieb. 2021. Chapter 11. measuring informativity: The rise of compounds as informationally dense structures in 20th-century scientific english. In *Corpus-based Approaches to Register Variation*, pages 291–312. John Benjamins Publishing Company.

Georgios P. Georgiou. 2024. Differentiating between human-written and ai-generated texts using linguistic features automatically extracted from an online computational tool.

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, and Alex Vaughan et al. 2024. The llama 3 herd of models.

Andrew Gray. 2024. Chatgpt "contamination": estimating the prevalence of llms in the scholarly literature.

Michael Alexander Kirkwood Halliday and James R Martin. 2003. *Writing science: Literacy and discursive power*. Routledge.

Jingyang Jiang, Peng Bi, and Haitao Liu. 2019. Syntactic complexity development in the writings of efl learners: Insights from a dependency syntactically-annotated corpus. *Journal of Second Language Writing*, 46:100666.

Tom S. Juzek and Zina B. Ward. 2024. Why does chatgpt "delve" so much? exploring the sources of lexical overrepresentation in large language models.

Dmitry Kobak, Rita González-Márquez, Emőke Ágnes Horvát, and Jan Lause. 2025. Delving into llm-assisted writing in biomedical publications through excess vocabulary. *Science Advances*, 11(27):eadt3813.

Solomon Kullback and Richard A. Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Bruce W. Lee and Jason Lee. 2023. LFTK: Handcrafted features in computational linguistics. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 1–19, Toronto, Canada. Association for Computational Linguistics.

Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, Diyi Yang, Christopher Potts, Christopher D Manning, and James Y. Zou. 2024. Mapping the increasing use of llms in scientific papers.

Wenxiong Liao, Zhengliang Liu, Haixing Dai, Shaochen Xu, Zihao Wu, Yiyang Zhang, Xiaoke Huang, Dajiang Zhu, Hongmin Cai, Quanzheng Li, Tianming Liu, and Xiang Li. 2023. Differentiating chatgpt-generated and human-written medical texts: Quantitative study. *JMIR Med Educ*, 9:e48904.

Zhehui Liao, Maria Antoniak, Inyoung Cheong, Evie Yu-Yen Cheng, Ai-Heng Lee, Kyle Lo, Joseph Chee Chang, and Amy X. Zhang. 2024. Llms as research tools: A large scale survey of researchers' usage and perceptions.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Tapas Mishra, Egidia Sutanto, Rani Rossanti, et al. 2024. Use of large language models as artificial intelligence tools in academic research and publishing among global clinical researchers. *Scientific Reports*, 14:31672.

Alba Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2024. Contrasting linguistic patterns in human and llm-generated news text. *Artificial Intelligence Review*, 57:265.

Claudio Novelli, Federico Casolari, Philipp Hacker, Giorgio Spedicato, and Luciano Floridi. 2024. Generative ai in eu law: Liability, privacy, intellectual property, and cybersecurity. *Computer Law Security Review*, 55:106066.

Chidimma Opara. 2024. Styloai: Distinguishing ai-generated content with stylometric analysis. *ArXiv*, abs/2405.10129.

OpenAI. 2024. Gpt-4o system card.

Subhajit Panda and Navkiran Kaur. 2024. Exploring the role of generative ai in academia: Opportunities and challenges. *IP Indian Journal of Library Science and Information Technology*, 9(1):12–23.

Alex Reinhart, Ben Markey, Michael Laudenbach, Kachatad 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).

Shaurya Rohatgi. 2022. Acl anthology corpus with full text. Github.

Morgan Sandler, Hyesun Choung, Arun Ross, and Prabu David. 2024. A linguistic comparison between human and chatgpt-generated conversations.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Yang Xu and David Reitter. 2016. Convergence of syntactic complexity in conversation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 443–448.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Sergio E. Zanotto and Segun Aroyehun. 2024. Human variability vs. machine consistency: A linguistic analysis of texts generated by humans and large language models.

A Prompt

System message: You are an efficient writing assistant specialized in creating concise and accurate text summaries for scientific publications. I will provide you with the full text of a scientific paper from the field of computational linguistics. Your task is to read the paper and write a clear and concise abstract for it. Write the abstract from the author's perspective. The abstract should be between 100 and 200 words long. Do not include any additional text like "Abstract:" or "Here is the abstract:".

User message: Write an abstract for this scientific paper: [FULL TEXT OF PAPER]

B Constituency Tree

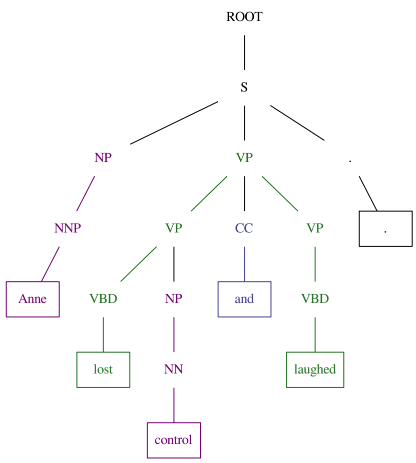


Figure 10: Example of constituency parsed tree

A State-of-the-Art Morphosyntactic Parser and Lemmatizer for Ancient Greek

Giuseppe G. A. Celano
Leipzig University
Faculty of Mathematics and Computer Science
Institute of Computer Science
celano@informatik.uni-leipzig.de

Abstract

This paper presents an experiment comparing six models to identify state-of-the-art models for Ancient Greek: a morphosyntactic parser and a lemmatizer that are capable of annotating in accordance with the Ancient Greek Dependency Treebank annotation scheme. A normalized version of the major collections of annotated texts was used to (i) train the baseline model Dithrax with randomly initialized character embeddings and (ii) fine-tune Trankit and four recent models pretrained on Ancient Greek texts, namely GreBERTa and PhilBERTa for morphosyntactic annotation and GreTA and PhilTa for lemmatization. A Bayesian analysis shows that Dithrax and Trankit are practically equivalent in morphological annotation, while syntax is best annotated by Trankit and lemmata by GreTa. The results of the experiment suggest that token embeddings are not sufficient to achieve high UAS and LAS scores unless they are coupled with a modeling strategy specifically designed to capture syntactic relationships. The dataset and best-performing models are made available online for reuse.

1 Introduction

In recent years, a few open-access annotated Ancient Greek (AG) corpora, such as *Opera Graeca Adnotata* (OGA) (Celano, 2024) and the GLAUx corpus (Keersmaekers, 2021), have been made available online. These corpora enable searches for morphosyntax and lemmata across a wide range of AG texts, thus filling the gap left by resources such as the *Thesaurus Linguae Graecae*, whose subscription-based query engine is limited to word forms and lemmata.

Because of the token count in the order of millions, the morphosyntactic annotation and lemmatization of the above-mentioned open-access corpora are feasible only if performed automatically. This raises a number of questions about which recent technology would be best suited for that purpose.

OGA v0.1.0 annotations (Celano, 2024) relied on the COMBO parser (Rybak and Wróblewska, 2018), which, despite being accurate,[1] was built on TensorFlow 1 and is not actively maintained anymore. The GLAUx corpus employed RFTagger (Schmid and Laws, 2008), Lemming (Müller et al., 2015), and the Stanford Graph-Based Dependency Parser (Dozat et al., 2017) for annotation of, respectively, morphology, lemmata, and syntax: the models perform well (see Keersmaekers, 2021, for details), but have not been released, and therefore cannot be reused.

For these reasons, the current paper presents a comparison of six models to identify and release state-of-the-art models for morphosyntactic annotation and lemmatization that can annotate literary AG sentences according to the annotation scheme of the Ancient Greek Dependency Treebank (AGDT) and can be used in production to process a large number of texts. To promote future machine learning-based studies on AG, the models and the normalized version of the AG texts used for training—and now documented with their alleged composition dates for the first time—are released.[2,3]

In Section 2, related work is reviewed, while Section 3 describes the dataset used for training. In Section 4, the experiment and the architectures of the different models compared are presented: the results of their training are reported with a Bayesian statistical analysis in Section 5 and discussed in Section 6. Finally, concluding remarks are contained in Section 7.

[1] https://git.informatik.uni-leipzig.de/celano/combo_for_ancient_greek.
[2] https://git.informatik.uni-leipzig.de/celano/morphosyntactic_parser_for_oga.
[3] https://git.informatik.uni-leipzig.de/celano/lemmatizer_for_oga.

Proceedings of the First Workshop NLP and Language Models for Digital Humanities associated with RANLP 2025, pages 48–65, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-106-6-005

2 Related work

The explosion of machine learning in NLP has generated an ever-increasing number of resources, the reuse of which, however, is often not possible or straightforward due to the many different variables involved in each system.

The most recent endeavor comparable to the work presented here is Keersmaekers and Van Hal (2023). Building on Keersmaekers (2021), they documented the parsing and lemmatization of a large corpus consisting of literary and papyrological AG texts annotated according to the AGDT annotation scheme. Interestingly, they conducted experiments to increase LAS and UAS scores, in which the original data were transformed before training: for example, elliptical nodes were deleted and the annotation style for coordination modified. The reported results show some UAS and LAS increases in absolute terms. The models, however, have not been released.

Most recent systems for morphosyntactic annotation and lemmatization were trained on the Universal Dependencies data, which consist of two treebanks, the Perseus treebank and the PROIEL treebank,[4] for a total of about 416K tokens—notably, the size of the UD treebanks is less than half of that of the data annotated with the AGDT annotation scheme used in the present study (see Section 3).

The UD treebanks implement the UD annotation scheme differently, and therefore creation of a single model still represents a challenge: Kostkan et al. (2023) provided a joint spaCy model for morphosyntactic annotation and lemmatization that seems to achieve good overall performance.[5]

A number of studies reported on the creation of token embeddings for AG by using the large amount of texts available online (Singh et al., 2021; Yamshchikov et al., 2022). Most recently, Riemenschneider and Frank (2023) benchmarked a number of models for Ancient Greek and Latin. They show that their pretrained language model GreBERTa achieves the highest performance scores for UPOS, XPOS, UAS, and LAS in absolute terms when fine-tuned on the UD Perseus treebank (95.83, 91.09, 88.20, and 83.98, respectively); lemmatization is best performed by a T5 model they call GreTa, which achieves an F1 score of 91.14.

3 The dataset

The dataset used for training, validation, and testing consists of the following treebanks:[6] (i) the Ancient Greek Dependency Treebank[7] (Celano, 2019; Bamman and Crane, 2011), (ii) the Gorman Trees[8] (Gorman, 2020), and (iii) the Pedalion Trees.[9]

All treebanks were natively annotated using the AGDT annotation scheme, and together they represent by far the largest morphosyntactically annotated dataset for literary AG texts—and one of the largest treebanks in absolute terms: the token count of the texts before normalization is $1,277,310$ and, after it, $1,260,863$.

As Table 1 shows, the final dataset comprises a plethora of texts of different genres—including poetry, history, and philosophy—and periods, ranging from about the 9th century BCE to the 4th century CE (more details are provided in Appendix E). Even though the dataset is not balanced across genres and periods, it is still representative of most text types written in Ancient Greece during the above-mentioned time span.

3.1 Normalization

Since the final database consists of texts from different sources, which were annotated by many different scholars (sometimes adopting different conventions), automatic normalization of the original texts was attempted to foster consistency and therefore performance of machine learning algorithms.

Before training, all the relevant fields, i.e., word form, lemma, POS tag, syntactic head and relation, needed some non-trivial format standardization, especially to handle the case of null or clearly erroneous values. Syntactic trees also had to be modified if cycles were detected.

[4] Recently, the PTNK treebank (about 39K tokens) has been added, but, as far as we are aware, it has not yet been used for machine learning experiments.

[5] The scores for the model odyCy_joint on the UD Perseus treebank test set reported at https://centre-for-humanities-computing.github.io/odyCy/performance.html are 95.39 (POS tagging), 92.56 (morphological features), 78.80 (UAS), 73.09 (LAS), and 83.20 (lemmatization). It is, however, not clear whether the evaluation script used is that of the CoNLL 2018 Shared Task (https://universaldependencies.org/conll18/evaluation.html), which is commonly used in similar studies, including the present one. Since this script does not allow for cycles and multiple roots, we suspect that the reported scores would be lower, if it had been used.

[6] Data licences can be found at the links to the data specified below.

[7] https://github.com/PerseusDL/treebank_data/releases/tag/v2.1_IGDS.

[8] https://github.com/vgorman1/Greek-Dependency-Trees.

[9] https://github.com/perseids-publications/pedalion-trees.

Author	Genre	Century	Tokens
Hesiod, Homer	poem	−9/8	255,375
Sappho, Mimnermus, Semonides	lyric	−7	5,510
Homeric Hymns	hymns	−7/6	3,968
Aesop	fable	−6	5,221
Antiphon, Lysias, Isocrates	oratory	−5	30,679
Aeschylus, Sophocles, Euripides	tragedy	−5	108,386
Aristophanes, Cephisidorus Comicus	comedy	−5	47,547
Aeneas Tacticus	manual	−5	7,207
Herodotus, Thucydides	history	−5	65,494
Xenophon	history	−5/4	142,635
Lysias, Isocrates, Demosthenes, Aeschines, Andocides, Isaeus	oratory	−4	153,088
Aristotle, Plato, Theophrastus	philosophy	−4	51,906
Menandrus	comedy	−4	8,069
Epicurus	philosophy	−4/3	1,523
Theocritus	lyric	−3	304
Septuaginta	Bible	−3	19,235
Polybius	history	−2	105,693
Ezechiel	tragedy	−2	1,939
Batrachomyomachia	poem	−1	2,212
Diodorus of Sicily, Dionysius of Halicarnassus	history	−1	56,004
Chion	epistolary	+1	5,577
Hero of Alexandria	science	+1	10,321
Josephus Flavius	history	+1	24,987
Chariton	romance	+1/2	6,265
Plutarch	biography	+1/2	37,203
Phlegon of Tralles	paradox.	+2	5,642
Apollodorus	mythogr.	+2	1,265
Epictetus	philosophy	+2	7,204
Lucian	novel	+2	11,054
Appianus	history	+2	25,665
Athenaeus	miscellany	+2	45,653
Longus	romance	+2/3	672
Sextus Empiricus	philosophy	+3	16,218
Paeanius	history	+4	6,184
Julian	oratory	+4	1,405

Table 1: Statistics for the works contained in the dataset showing authors, genres, (alleged) centuries of composition (indicated by Arabic numbers, with + meaning CE and − BCE), and token counts (before normalization). Full details in Appendix E.

An often underestimated problem is that of character encoding for the apostrophe: all apostrophe-looking characters were converted into the character MODIFIER LETTER APOSTROPHE (U+02BC), which affected about 50K characters.

While the vast majority of AG graphic words corresponds to morphosyntactic tokens,[10] this is questionable for coordinate conjunctions such as οὐδὲ or εἴτε, which, in the final dataset, were tokenized (therefore, οὐ δὲ and εἴ τε, respectively). Coordination in the AGDT is not only annotated at the level of the syntactic tree but also at that of the syntactic label via use of the suffix _CO: to decrease the number of syntactic labels and therefore supposedly improve algorithm performance, this and similar suffixes, such as _AP, were deleted.

Another related yet different issue is represented by ellipsis, which poses a representational challenge. The AGDT annotation scheme allows elliptical nodes to be added whenever they are necessary to build a syntactic tree. However, the complexity of the phenomenon and the absence of strict annotation rules on this matter have over time led to the proliferation of various annotation styles: for example, sometimes the word form of an elliptical node is specified, sometimes it is not. The position of elliptical nodes within a sentence is also problematic both on a theoretical and a representational level.

While Keersmaekers and Van Hal (2023) proposed deletion of elliptical nodes, Celano's (2023) ellipsis modeling is followed in the present study: elliptical nodes were added at the end of a sentence (whatever their alleged position was) and, to avoid uncertainties about their word forms, they were always encoded with placeholders such as [0], [1], and so on, depending on their number.[11]

4 Experiment

A total of six model architectures were compared: four (i.e., three + baseline) for morphosyntactic prediction and three (i.e., two + baseline) for lemma prediction. More precisely, the baseline model called **Dithrax**[12] is able to predict both morphosyntax and lemmata, while the other five models can predict either one, in that their modeling for character prediction for lemmatization is kept distinct

[10] Crasis annotation, which is more elaborate to normalize, was left untouched.

[11] Since a model to predict such elliptical nodes is provided at https://git.informatik.uni-leipzig.de/celano/ellipsis_Ancient_Greek, new texts can be made compliant with this ellipsis annotation style.

[12] The name derives from Dionysius Thrax, the author of the first extant AG grammar.

Model	POS	XPOS	Feats	AllTags	UAS	LAS	Lemmas
Dithrax	95.55 (0.23)	90.65 (0.32)	94.40 (0.17)	89.80 (0.39)	77.70 (0.62)	70.81 (0.65)	86.85 (0.18)
Trankit	**96.18 (0.13)**	**91.55 (0.21)**	**94.61 (0.12)**	**91.21 (0.22)**	**82.28 (0.27)**	**76.67 (0.34)**	N/A
GreBERTa	94.12 (0.54)	89.16 (0.73)	93.21 (0.45)	88.31 (0.85)	58.85 (2.04)	53.41 (2.06)	N/A
GreTa	N/A	N/A	N/A	N/A	N/A	N/A	**91.17 (0.17)**
PhilBERTa	85.34 (24.03)	79.85 (24.3)	86.67 (16.87)	77.8 (27.73)	61.24 (20.64)	54.95 (20.1)	N/A
PhilTa	N/A	N/A	N/A	N/A	N/A	N/A	90.09 (0.24)
UD Perseus Trankit	93.97	87.25	91.66	86.88	83.48	78.56	88.52
UD Perseus GreBERTa	95.83	91.09	N/A	N/A	88.20	83.98	N/A
UD Perseus GreTa + Chars	N/A	N/A	N/A	N/A	N/A	N/A	91.14
UD Perseus PhilBERTa	95.60	90.41	N/A	N/A	86.99	82.69	N/A
UD Perseus PhilTa + Chars	N/A	N/A	N/A	N/A	N/A	N/A	90.66

Table 2: Mean F1 scores + standard deviations in parentheses for the test set results of the 5-fold cross-validation models (training on each split repeated twice with different random seeds). Best scores are in boldface. Results for parsers trained on the UD Perseus data are shown only for loose comparison (see Section 5).

from that for word prediction for morphosyntax.

The performance of each model was evaluated with the official CoNLL 2018 Shared Task script: it outputs F1 scores for UPOS, XPOS, UFeats, AllTags (i.e., UPOS+XPOS+UFeats), UAS (i.e., HEAD match), LAS (i.e., HEAD + DEPREL match), and Lemmas. Since the AGDT tagsets are different, the above-mentioned metrics are conveniently renamed: POS, XPOS, Feats, AllTags, UAS, LAS, and Lemmas.

The original dataset was divided into training, validation, and test sets (60%, 20%, 20%). Each model was trained 10 times, using 5-fold cross-validation, with each training-validation split being used twice: as a result, 10 models (i.e., 5 splits $\times$ 2 random seeds) were trained for each model architecture (therefore, 10 final F1 scores were calculated for each of the above-mentioned metrics). Since the final models were not retrained on the entire dataset (train + validation sets) for time reasons, the mean scores presented in Table 2 are the ones obtained on the test set—the best-performing model was then chosen for use in production (see Table 3).

The training strategy is motivated by the fact that, while cross-validation reduces variance by use of different splits of the dataset, repetition of training on the same split allows experimentation with different random seeds. Final hyperparameters were set after a number of preliminary experiments and are documented in Appendix B.

Model	POS	XPOS	Feats	AllTags	UAS	LAS	Lemmas
Trankit	96.41	91.90	94.77	91.56	82.60	77.10	N/A
GreTa	N/A	N/A	N/A	N/A	N/A	N/A	91.41

Table 3: Scores of the best-performing cross-validation runs evaluated on the test set.

4.1 The statistical framework

The results of the present experiment are interpreted through the Bayesian analysis proposed by Benavoli et al. (2017). More precisely, they propose a Bayesian correlated t-test to compare cross-validation scores of two models on one dataset.

The proposed posterior distribution coincides with the Student distribution used in the frequentist t-test. This means that the probabilities of the Bayesian correlated t-test coincide with the p-values of the frequentist correlated t-test: what changes, however, is the interpretation of such numerical values.

While the frequentist approach returns the probability of data under the assumption that the null hypothesis is true, the Bayesian correlated t-test computes the actual probabilities of the null and alternative hypotheses.

Benavoli et al.'s (2017) Bayesian correlated t-test provides three probability scores concerning the comparison of the models x and y (see Appendix C for the scores):

(i) $P(x = y)$: the probability of model x being practically equivalent to model y: this is the *region of practical equivalence* (**ROPE**) corresponding to an arbitrary interval within which two models are considered not to differ in practice. In the present study, this is $[-1, 1]$, i.e., the posterior probability of the mean difference of F1 scores less than 1% is considered to mean practical equivalence.

(ii) $P(x \ll y)$: the probability that model x is practically worse than model y, i.e., the posterior probability of the mean difference of F1 scores being practically negative.

(iii) $P(x \gg y)$: the probability that model x is practically better than model y, i.e., the posterior probability of the mean difference of F1 scores being practically positive.

The Bayesian approach provides a more straightforward statistical interpretation of data and offers a solution for the well-known pitfalls of the frequentist framework, which include the fact that null hypotheses are always false in practice and sufficiently large datasets can yield statistical significance even if the effect size is very small.

4.2 Dithrax: the baseline model

As shown in Figure 1, Dithrax is a multi-output LSTM model vectorizing morphosyntactic tokens with *randomly initialized* character embeddings, which are used for prediction of both lemmata and, after further processing through LSTM layers, morphosyntax.

The model is inspired by the COMBO parser (Rybak and Wróblewska, 2018), which was among the most accurate parsers at the CoNLL 2018 Shared Task (Zeman et al., 2018).

More precisely, Dithrax proposes a similar modeling strategy for HEAD and DEPREL targets based on adjacency matrices resulting from dot products of two rank-2 tensors representing, respectively, heads and dependents of the same sentence, with each matrix row corresponding to the vector representation of a token.

4.3 Trankit

Trankit (Nguyen et al., 2021) is a state-of-the-art transformer-based toolkit for morphosyntactic analysis and lemmatization. It is designed for UD data, and is also able to process raw documents, in that it comprises a tokenizer and sentence splitter. Key features of Trankit are:

(i) use of the multilingual pretrained transformer XLM-RoBERTa, whose output is fine-tuned on new data.

(ii) adapters: feed-forward networks for each major component of Trankit (six in total), whose weights—together with the specific ones for final predictions—are the only ones updated, while the pretrained transformer weights remain fixed. These make Trankit memory- and time-efficient.

(iii) syntax is modeled via Dozat and Manning's (2017) biaffine attention.

For the purpose of the present experiment, we trained Trankit's joint model for part-of-speech tagging, morphological feature tagging, and dependency parsing (i.e., POS, XPOS, Feats, AllTags, UAS, and LAS scores); the lemmatizer could not be trained because of an internal code error.[13]

4.4 Pretrained models: Gre(BERTa|Ta) and Phil(BERTa|Ta)

The pretrained models GreBERTa and GreTa (for AG) and PhilBERTa and PhilTa (for AG and Latin) were fine-tuned for comparison,[14] in that they have recently been argued to perform better than previous pretrained AG models.

Riemenschneider and Frank (2023) fine-tuned GreBERTa and GreTa on the Greek data of the Open Greek and Latin Project, the CLARIN corpus of Greek Medieval Texts, the Patrologia Graeca, and the Internet Archive (in total, about 185.1M tokens). They fine-tuned PhilBERTa and PhilTa on not only AG but also Latin and English data. The latter come from the Corpus Corporum project (167.5M tokens) and a collection of English texts from different sources (212.8M tokens), whose topics are similar to the ones found in AG and Latin sources (for example, English translations of AG and Latin texts), for a total of 565.4M tokens.

GreBERTa and PhilBERTa are encoder-only transformers providing token embeddings for prediction of word-related targets (i.e., UPOS, XPOS, UFeats, AllTags, HEAD, and DEPREL). Since not the original scripts but only the pretrained models are made available online (see also Section 8), it was not possible to test the former with the AGDT dataset (see Section 6): in the present experiment, therefore, the pretrained token embeddings were just used as inputs to dense layers outputting the final probability scores for each token. However, the parameters of the pretrained models were left trainable. GreTa and PhilTa are encoder-decoder transformers for character prediction, and we fine-tuned them for lemmatization.[15]

[13]See https://github.com/nlp-uoregon/trankit/issues/48.

[14]We use the names GreBERTa, PhilBERTa, GreTa, and PhilTa to also name the models obtained by our fine-tuning: context is sufficient to clarify what these names exactly refer to.

[15]We are grateful to Frederick Riemenschneider, who provided us with a script for lemma prediction similar to the one used for his paper.

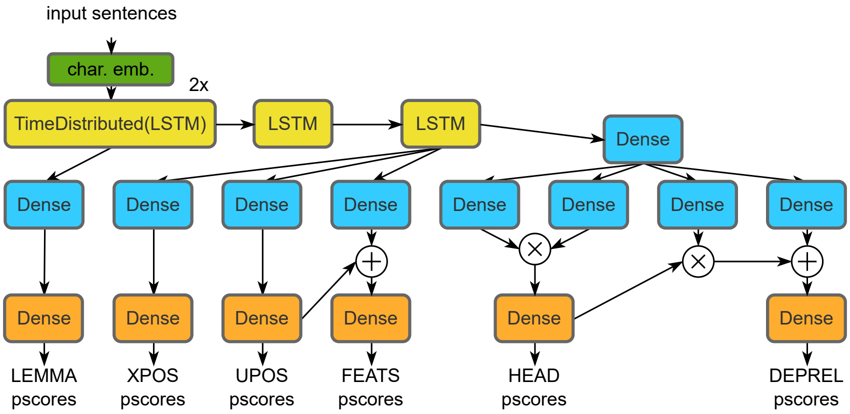


Figure 1: Main layers of Dithrax, the baseline model architecture. Blue color stands for `tanh(linear(x))`, while orange for `softmax(linear(x))` (with $\times$ meaning dot product and $+$ concatenation).

5 Results

Table 2 shows the mean F1 scores and related standard deviations[16] for the models trained with 5-fold cross-validation, with each split being used twice with different random seeds (in total, 10 models for each architecture). The mean scores are based on the F1 scores returned by the evaluation script of the CoNLL 2018 Shared Task applied to the results outputted by each model when tested on the test set. The models created by the runs with the best scores (see Table 3) are made available online.[17]

Table 2 also displays Riemenschneider and Frank's (2023) results for the models trained on UD Perseus data, i.e., a small subset of the dataset used for the present study, which were evaluated using the same CoNLL 2018 Shared Task script.[18] Even if the UD annotation scheme and the AGDT one are similar, there are differences that are likely to impact parsing results. For example, Keersmaekers (2021) argues that UD annotation style of coordination allows one to achieve higher scores for UAS and LAS. Moreover, UD data, unlike the AGDT data used for the present study, do not contain elliptical nodes. This means that comparison of F1 scores between UD models and the ones of the present study can only be loose, especially with reference to UAS and LAS.

The mean scores for PhilBERTa shown in Table 2 are the lowest ones and their related standard deviations are remarkably high (>20) because the model performed very poorly in one of the runs. However, even if that run were not considered, the mean scores would still be lower and the standard deviations would remain rather high in comparison to the values of the other models: POS: 92.87 (3.63); XPOS: 87.42 (4.37); Feats: 91.92 (3.11); AllTags: 86.44 (4.94); UAS: 67.44 (6.9); LAS: 60.94 (7.11).

Figures 2, 3, 4, 5, 6, 7, and 8 show the posterior distributions of the mean differences of F1 scores between all models pairwise returned by Benavoli et al.'s (2017) Bayesian correlated t-test.[19]

In each of the above-mentioned figures except Figure 8, the top-left, top-middle, top-right, bottom-left, bottom-middle, bottom-right plots show, respectively, the posteriors for the pairs Dithrax-Trankit, Dithrax-PhilBERTa, Dithrax-GreBERTa, Trankit-PhilBERTa, Trankit-GreBERTa, and GreBERTa-PhilBERTa. In Figure 8, which visualizes Lemmas scores, the left, middle, and right plots represent the posteriors for Dithrax-PhilTa, Dithrax-GreTa, and GreTa-PhilTa, respectively—as noted above, Trankit could not be trained for lemmatization because of an internal code error. Each above-mentioned Figure is coupled with a table (i.e., Tables 5, 6, 7, 8, 9, 10, and 11 in Appendix C), which reports the values of the areas under the curve.

Each single plot gives information about the probabilities that the mean differences of F1 scores between two models are practically negative, practically equivalent, and practically positive. For

[16]SDs have been calculated using `numpy.std` with `ddof=1`.

[17]See footnotes 2 and 3.

[18]Results for Trankit are taken from `https://trankit.readthedocs.io/en/latest/performance.html` (`Ancient_Greek-Perseus` treebank).

[19]The Python package documented at `https://github.com/janezd/baycomp` was used for the plots and calculations.

example, the bottom-left plot in Figure 4 and the corresponding Table 7 show:

- the posterior probability that the mean difference of F1 scores between PhilBERTa and Trankit is practically negative, i.e., the integral of the posterior over the interval $(-\infty, -1)$, equal to ≈ 0.80. This is the probability that Trankit is practically **better** than PhilBERTa.

- the posterior probability that the mean difference of F1 scores between PhilBERTa and Trankit is practically equivalent, i.e., the integral of the posterior over the ROPE interval $[-1, 1]$, equal to ≈ 0.06. This is the probability that PhilBERTa and Trankit are practically **equivalent**.

- the posterior probability that the mean difference of F1 scores between PhilBERTa and Trankit is practically positive, i.e., the integral of the posterior over the interval $(1, +\infty)$, equal to ≈ 0.14. This is the probability that PhilBERTa is practically **better** than Trankit.

6 Discussion

Table 2 seems to suggest that Trankit is the best model in each morphosyntactic task. This is only *partly* confirmed by the Bayesian statistical analysis.

Even if Trankit's results for POS, XPOS, and Feats are the highest in absolute terms, its performance can be considered to be practically equivalent to that of the baseline model Dithrax with reference to these metrics. Indeed, the corresponding Tables 5, 6, and 7 show that the area under the curve within the ROPE is ≈ 1 for POS and Feats, and ≈ 0.88 for XPOS.[20]

On the other hand, the models PhilBERTa and GreBERTa perform practically worse than both Dithrax and Trankit with respect to these same metrics: there is at least an ≈ 0.79 probability (see Dithrax-PhilBERTa in Table 5)[21] that Dithrax or Trankit performs practically better.

This is an interesting result because, unlike Trankit, PhilBERTa, and GreBERTa, Dithrax does not rely on pretrained (but randomly initialized) character embeddings and its architecture has a lower overall number of parameters (see Table 4):[22] this suggests that classification tasks such as POS, XPOS, and Feats can be successfully addressed without use of more expensive model architectures—however, as shown in Table 4, Dithrax has a longer training time. The AllTags F1 score is a metric for POS+XPOS+Feats. Trankit turns out to perform practically better than any other model (Table 8), including Dithrax.

Syntactic prediction is notoriously more complex, and this is shown in the lower results reported in Table 2 for UAS and LAS. Trankit's performance is clearly superior to that of any other model, even if its scores are much lower than the POS and XPOS ones.

Syntactic analysis is a much more challenging task because HEAD and DEPREL values heavily depend on contextual information. Even if a pretrained transformer such as GreBERTa or PhilBERTa outputs context-aware token embeddings, it turns out to predict syntax poorly without a further modeling strategy.

In the GreBERTa and PhilBERTa models, the pretrained token embeddings were used as input to dense layers outputting probabilities for morphology and syntax in a multi-output model; however, while results for morphology are comparable to those of the other models, those for syntax clearly are not (see also Section 8): as Table 2 shows, UAS and LAS scores for GreBERTa and PhilBERTa are remarkably lower, and there is an ≈ 0.93 or higher probability that Dithrax or Trankit performs practically better than them (Tables 9 and 10).

This can be explained by the fact that, contrary to GreBERTa and PhilBERTa, Dithrax and Trankit employ a modeling strategy on top of embeddings: Dithrax models sentence syntax through adjacency matrices (Rybak and Wróblewska, 2018), while Trankit implements Dozat and Manning's (2017) biaffine attention mechanism, both of which aim to capture the complex relationship between heads and dependents within a sentence.[23]

Lemmatization is performed best by GreTa. While Dithrax simply employs LSTM layers over character embeddings, GreTa and PhilTa are seq2seq models: Table 11 shows that, while the seq2seq models perform practically better than Dithrax (≈ 1.00), there is an ≈ 0.75 probability that

[20] A threshold of 0.80 can be chosen when comparing the models.

[21] 0.79 is actually lower than the threshold of 0.80, but the difference is minimal.

[22] However, Trankit has fewer trainable parameters than Dithrax.

[23] To filter syntactic cycles, the Chu-Liu-Edmonds algorithm is applied to each parser's output.

GreTa performs practically better than PhilTa and an ≈ 0.25 probability that their performance is practically equivalent.

If we compare Trankit's results on the AGDT dataset with those on the UD dataset (see Table 2), scores for POS, XPOS, Feats, and AllTags are considerably higher in absolute terms on the AGDT dataset, with differences of ≈ 2.21, ≈ 4.3, ≈ 2.95, and ≈ 4.33, respectively; UAS and LAS scores, however, are higher on the UD dataset, with differences of ≈ 1.2 and ≈ 1.89, respectively. Interestingly, UAS and LAS scores do not seem to be impacted by the much larger size of the AGDT dataset; however, the model trained on the AGDT data can be expected to generalize much better than that trained on the UD data due to the much larger variety of texts used during training.

7 Conclusions

A comparison of six model architectures (Dithrax, Trankit, PhilBERTa, GreBERTa, PhilTa, and GreTa) was documented to select state-of-the-art models for annotation of morphosyntax and lemmata of literary texts according to the AGDT annotation scheme. A Bayesian statistical analysis was adopted to interpret cross-validation scores, which suggests that Trankit annotates syntax better than the other models do, while GreTa's performance for lemmatization is the best. The study shows that the baseline model Dithrax can also achieve state-of-the-art performance for morphological annotation—it employs randomly initialized character embeddings and a lower overall number of parameters, but its training time is longer.

A noteworthy finding of the study is that, although pretrained embeddings, such as GreBERTa and PhilBERTa, rely on complex model architectures vectorizing tokens with embeddings calculated on a very large collection of AG texts, they do not perform well for syntactic prediction (i.e., UAS and LAS scores), unless a further modeling strategy aimed at capturing syntax information within a sentence is put in place, such as adjacency matrices or biaffine attention.

8 Limitations

The study aimed to document state-of-the-art models for morphosyntactic analysis and lemmatization of Ancient Greek. The dataset used for training contains manual annotations produced over many years by different (single) annotators (some were students, others scholars). Therefore, as is often the case with manual annotations, annotation consistency within the dataset cannot be guaranteed because of either annotation errors or different annotation styles, the first annotation guidelines[24] not being sufficiently specific regarding a number of morphosyntactic phenomena—it should also be noted that the morphosyntactic annotation of Ancient Greek literary texts is arguably much more complex than that of modern texts.

For this reason, the present study set aside the question of how annotation quality/consistency affects parsing results. Similarly, no experiment was conducted with respect to corpus composition, under the assumption that model architectures are powerful enough to capture distinctions between texts of different genres and/or composition dates. Moreover, as stated in Section 1, the focus of the study was to select a morphosyntactic parser and a lemmatizer that performed best overall based on well-known metrics and a statistical analysis: a model error analysis would be of interest, but lies beyond the scope of this study.

The reuse of models and model architectures for comparison was often limited: either they are not released or the provided code is partial. The latter case is that of PhilBERTa and GreBERTa: they achieved state-of-the-art UAS and LAS scores on the UD Perseus treebank, but the original scripts have not been released,[25] and therefore their original model architectures could not be used in the present study.

Acknowledgments

This work has been supported by the German Research Foundation (DFG project number 408121292).

References

David Bamman and Gregory Crane. 2011. The Ancient Greek and Latin dependency treebanks. In *Language Technology for Cultural Heritage*, pages 79–98, Berlin, Heidelberg. Springer Berlin Heidelberg.

[24] https://github.com/PerseusDL/treebank_data/blob/master/v1/greek/docs/guidelines.pdf; newer annotated texts should follow the much more specific annotation guidelines at https://github.com/PerseusDL/treebank_data/blob/master/AGDT2/guidelines/.

[25] https://github.com/Heidelberg-NLP/ancient-language-models/tree/main.

Alessio Benavoli, Giorgio Corani, Janez Demšar, and Marco Zaffalon. 2017. Time for a change: A tutorial for comparing multiple classifiers through Bayesian analysis. *Journal of Machine Learning Research*, 18(77):1–36.

Giuseppe G. A. Celano. 2019. The dependency treebanks for Ancient Greek and Latin. *Digital Classical Philology*, pages 279–298.

Giuseppe G. A. Celano. 2023. A neural network approach to ellipsis detection in Ancient Greek. In *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*, pages 151–158, Online. Association for Computational Linguistics.

Giuseppe G. A. Celano. 2024. Opera Graeca Adnotata: Building a 34m+ token multilayer corpus for Ancient Greek. *Preprint*, arXiv:2404.00739.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *International Conference on Learning Representations*.

Timothy Dozat, Peng Qi, and Christopher D. Manning. 2017. Stanford's graph-based neural dependency parser at the CoNLL 2017 shared task. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 20–30, Vancouver, Canada. Association for Computational Linguistics.

Vanessa B. Gorman. 2020. Dependency treebanks of Ancient Greek prose. *Journal of Open Humanities Data*, 6(1).

Alek Keersmaekers. 2021. The GLAUx corpus: methodological issues in designing a long-term, diverse, multi-layered corpus of Ancient Greek. In *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021*, pages 39–50, Online. Association for Computational Linguistics.

Alek Keersmaekers and Toon Van Hal. 2023. Creating a large-scale diachronic corpus resource: Automated parsing in the Greek papyri (and beyond). *Natural Language Engineering*, pages 1–30.

Jan Kostkan, Márton Kardos, Jacob Palle Bliddal Mortensen, and Kristoffer Laigaard Nielbo. 2023. OdyCy – A general-purpose NLP pipeline for Ancient Greek. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 128–134, Dubrovnik, Croatia. Association for Computational Linguistics.

Thomas Müller, Ryan Cotterell, Alexander Fraser, and Hinrich Schütze. 2015. Joint lemmatization and morphological tagging with lemming. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2268–2274, Lisbon, Portugal. Association for Computational Linguistics.

Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 80–90, Online. Association for Computational Linguistics.

Frederick Riemenschneider and Anette Frank. 2023. Exploring large language models for classical philology. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15181–15199, Toronto, Canada. Association for Computational Linguistics.

Piotr Rybak and Alina Wróblewska. 2018. Semi-supervised neural system for tagging, parsing and lemmatization. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 45–54, Brussels, Belgium. Association for Computational Linguistics.

Helmut Schmid and Florian Laws. 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 777–784, Manchester, UK. Coling 2008 Organizing Committee.

Pranaydeep Singh, Gorik Rutten, and Els Lefever. 2021. A pilot study for BERT language modelling and morphological analysis for Ancient and Medieval Greek. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 128–137, Punta Cana, Dominican Republic (online). Association for Computational Linguistics.

Ivan P. Yamshchikov, Alexey Tikhonov, Yorgos Pantis, Charlotte Schubert, and Jürgen Jost. 2022. BERT in Plutarch's shadows. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6071–6080, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.

A Model statistics

Model	APar	TPar	TTime
Dithrax	$58,906,077$	$58,906,077$	$\approx$14.6h
Trankit	$283,463,421$	$5,419,773$	$\approx$6.9h
GreBERTa	$127,860,506$	$127,860,506$	$\approx$2.6h
GreTa	$247,539,456$	$247,539,456$	$\approx$11.4h
PhilBERTa	$137,076,506$	$137,076,506$	$\approx$2.6h
PhilTa	$296,691,456$	$296,691,456$	$\approx$12.3h

Table 4: Model statistics consisting of number of all parameters (APar), trainable parameters (TPar), and approximate training time (TTime) calculated on an NVIDIA RTX4500 ADA 24GB GDDR6.

B Model hyperparameters

The present section reports the relevant hyperparameters for the training of the models. Dithrax (TensorFlow/Keras): batch size 28, epochs 100 with early stopping (patience 2, best model saved), and Adam optimizer with clipvalue 4.5, $\beta_1 = 0.9$, $\beta_2 = 0.9$, weight decay $1\mathrm{e}{-4}$, and learning rate using piecewise constant decay with boundaries $[15000, 27000]$ and values $[0.001, 0.0001, 0.00007]$.

PhilBERTa and GreBERTa (TensorFlow/Keras/Transformers): batch size 28, epochs 100 with early stopping (patience 2, best model saved), and Adam optimizer with clipvalue 4.5, $\beta_1 = 0.9$, $\beta_2 = 0.9$, weight decay $1\mathrm{e}{-4}$, and learning rate using piecewise constant decay with boundaries $[10000]$ and values $[0.001, 0.0001, 0.00007]$.

Trankit (PyTorch/Transformers): token embeddings `xlm-roberta-base`, batch size 16, epochs 100 (best model saved), and a linear scheduler with warmup steps 80, training steps 160, and AdamW optimizer with learning rate $1\mathrm{e}{-3}$ and weight decay $1\mathrm{e}{-4}$.

PhilTa and GreTa (PyTorch/Transformers): `Seq2SeqTrainingArguments` with batch size 128, epochs 10, learning rate $1\mathrm{e}{-4}$, weight decay $1\mathrm{e}{-3}$, gradient accumulation steps 1, generation max length 30, and generation number of beams 20.

C Scores from the Bayesian correlated t-tests

Model pair	Left	ROPE	Right
Dithrax-Trankit	$\approx$0.00	$\approx$1.00	$\approx$0.00
Dithrax-PhilBERTa	$\approx$0.79	$\approx$0.04	$\approx$0.17
Dithrax-GreBERTa	$\approx$0.98	$\approx$0.02	$\approx$0.00
Trankit-PhilBERTa	$\approx$0.80	$\approx$0.04	$\approx$0.16
Trankit-GreBERTa	$\approx$1.00	$\approx$0.00	$\approx$0.00
GreBERTa-PhilBERTa	$\approx$0.75	$\approx$0.05	$\approx$0.20

Table 5: Integrals on the intervals $(-\infty, -1)$, $[-1, 1]$, and $(1, +\infty)$ for plots in Figure 2 (POS).

Model pair	Left	ROPE	Right
Dithrax-Trankit	$\approx$0.00	$\approx$0.88	$\approx$0.12
Dithrax-PhilBERTa	$\approx$0.80	$\approx$0.04	$\approx$0.16
Dithrax-GreBERTa	$\approx$0.97	$\approx$0.03	$\approx$0.00
Trankit-PhilBERTa	$\approx$0.82	$\approx$0.04	$\approx$0.14
Trankit-GreBERTa	$\approx$1.00	$\approx$0.00	$\approx$0.00
GreBERTa-PhilBERTa	$\approx$0.76	$\approx$0.05	$\approx$0.19

Table 6: Integrals over the intervals $(-\infty, -1)$, $[-1, 1]$, and $(1, +\infty)$ for plots in Figure 3 (XPOS).

Model pair	Left	ROPE	Right
Dithrax-Trankit	$\approx$0.00	$\approx$1.00	$\approx$0.00
Dithrax-PhilBERTa	$\approx$0.80	$\approx$0.06	$\approx$0.14
Dithrax-GreBERTa	$\approx$0.86	$\approx$0.14	$\approx$0.00
Trankit-PhilBERTa	$\approx$0.80	$\approx$0.06	$\approx$0.14
Trankit-GreBERTa	$\approx$0.98	$\approx$0.02	$\approx$0.00
GreBERTa-PhilBERTa	$\approx$0.75	$\approx$0.07	$\approx$0.18

Table 7: Integrals over the intervals $(-\infty, -1)$, $[-1, 1]$, and $(1, +\infty)$ for plots in Figure 4 (Feats).

Model pair	Left	ROPE	Right
Dithrax-Trankit	$\approx$0.00	$\approx$0.00	$\approx$1.00
Dithrax-PhilBERTa	$\approx$0.80	$\approx$0.04	$\approx$0.17
Dithrax-GreBERTa	$\approx$0.95	$\approx$0.05	$\approx$0.00
Trankit-PhilBERTa	$\approx$0.82	$\approx$0.03	$\approx$0.14
Trankit-GreBERTa	$\approx$1.00	$\approx$0.00	$\approx$0.00
GreBERTa-PhilBERTa	$\approx$0.76	$\approx$0.04	$\approx$0.19

Table 8: Integrals over the intervals $(-\infty, -1)$, $[-1, 1]$, and $(1, +\infty)$ for plots in Figure 5 (AllTags).

Model pair	Left	ROPE	Right
Dithrax-Trankit	$\approx$0.00	$\approx$0.00	$\approx$1.00
Dithrax-PhilBERTa	$\approx$0.93	$\approx$0.02	$\approx$0.05
Dithrax-GreBERTa	$\approx$1.00	$\approx$0.00	$\approx$0.00
Trankit-PhilBERTa	$\approx$0.97	$\approx$0.01	$\approx$0.02
Trankit-GreBERTa	$\approx$1.00	$\approx$0.00	$\approx$0.00
GreBERTa-PhilBERTa	$\approx$0.36	$\approx$0.08	$\approx$0.56

Table 9: Integrals over the intervals $(-\infty, -1)$, $[-1, 1]$, and $(1, +\infty)$ for plots in Figure 6 (UAS).

Model pair	Left	ROPE	Right
Dithrax-Trankit	$\approx$0.00	$\approx$0.00	$\approx$1.00
Dithrax-PhilBERTa	$\approx$0.93	$\approx$0.02	$\approx$0.05
Dithrax-GreBERTa	$\approx$1.00	$\approx$0.00	$\approx$0.00
Trankit-PhilBERTa	$\approx$0.98	$\approx$0.01	$\approx$0.02
Trankit-GreBERTa	$\approx$1.00	$\approx$0.00	$\approx$0.00
GreBERTa-PhilBERTa	$\approx$0.39	$\approx$0.09	$\approx$0.52

Table 10: Integrals over the intervals $(-\infty, -1)$, $[-1, 1]$, and $(1, +\infty)$ for plots in Figure 7 (LAS).

Model pair	Left	ROPE	Right
Dithrax-PhilTa	$\approx$0.00	$\approx$0.00	$\approx$1.00
Dithrax-GreTa	$\approx$0.00	$\approx$0.00	$\approx$1.00
GreTa-PhilTa	$\approx$0.75	$\approx$0.25	$\approx$0.00

Table 11: Integrals over the intervals $(-\infty, -1)$, $[-1, 1]$, and $(1, +\infty)$ for plots in Figure 8 (Lemmas).

D Posteriors

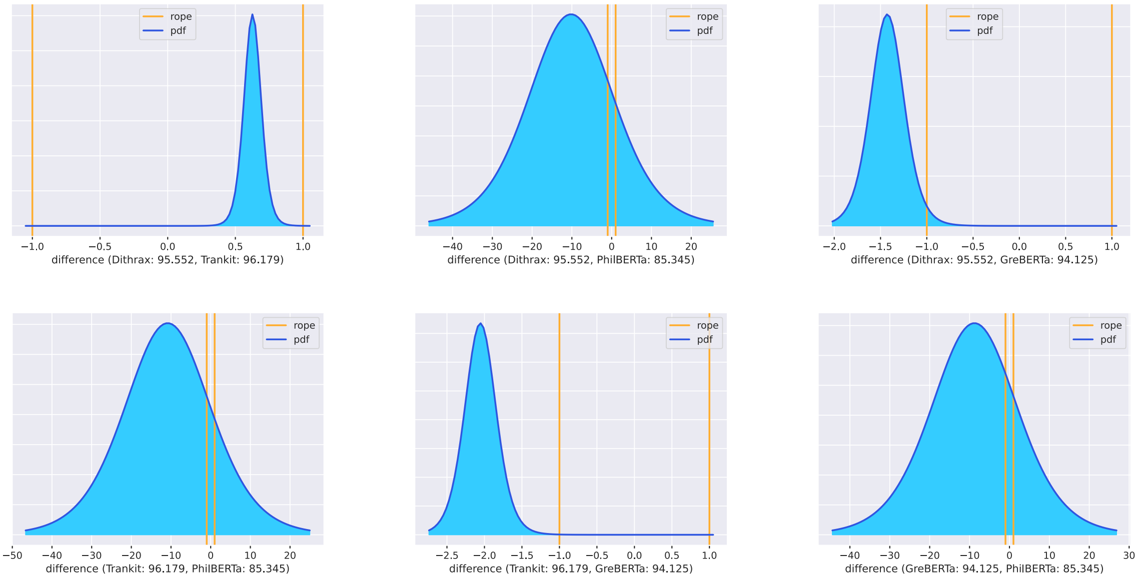


Figure 2: Posteriors of the Bayesian correlated t-test for all model pairs with reference to POS scores.

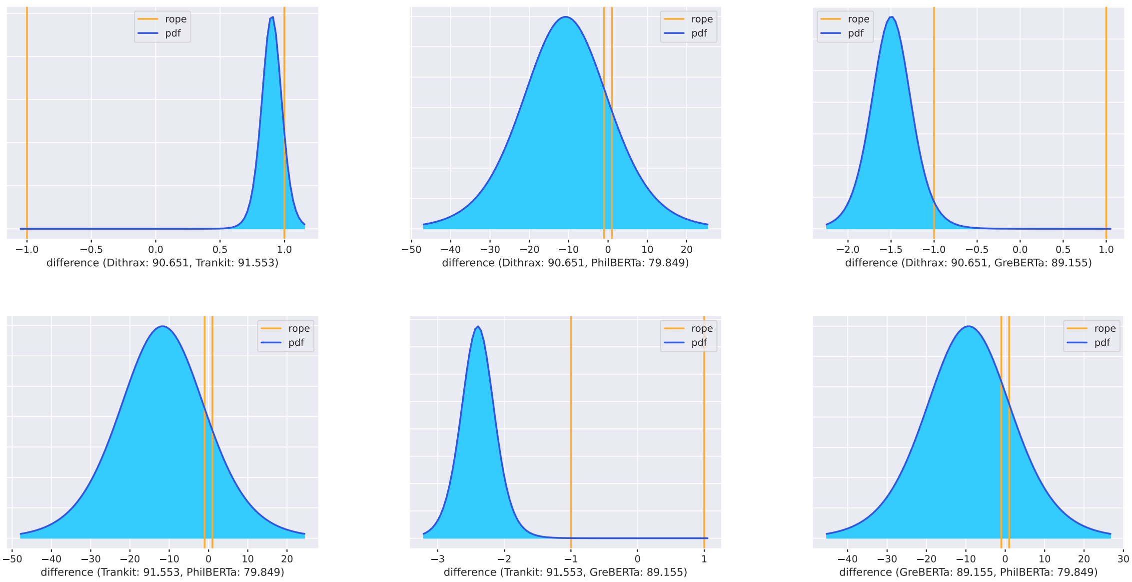


Figure 3: Posteriors of the Bayesian correlated t-test for all model pairs with reference to XPOS scores.

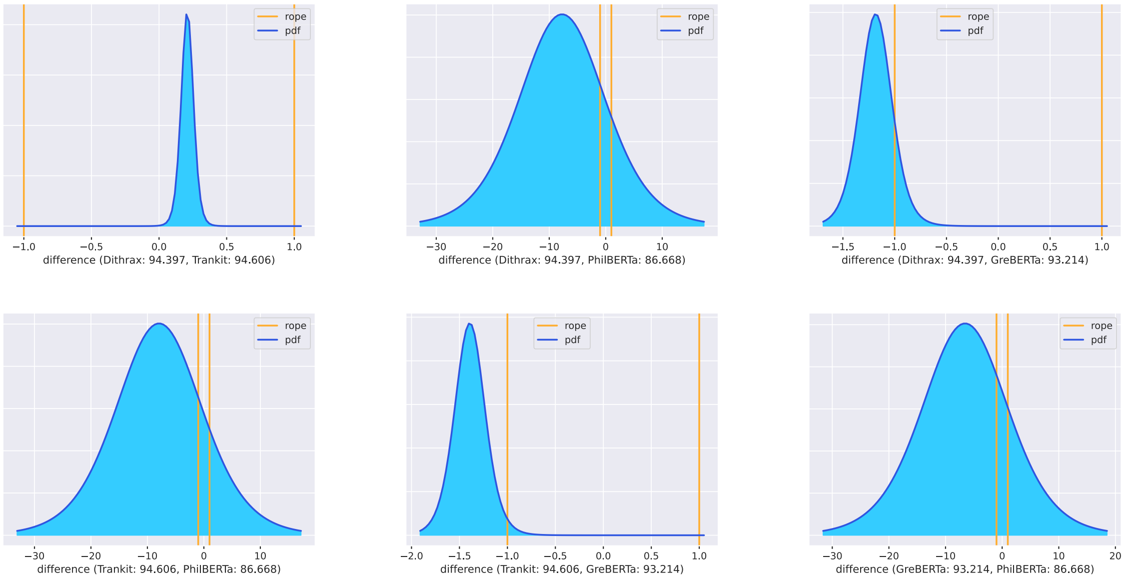


Figure 4: Posteriors of the Bayesian correlated t-test for all model pairs with reference to Feats scores.

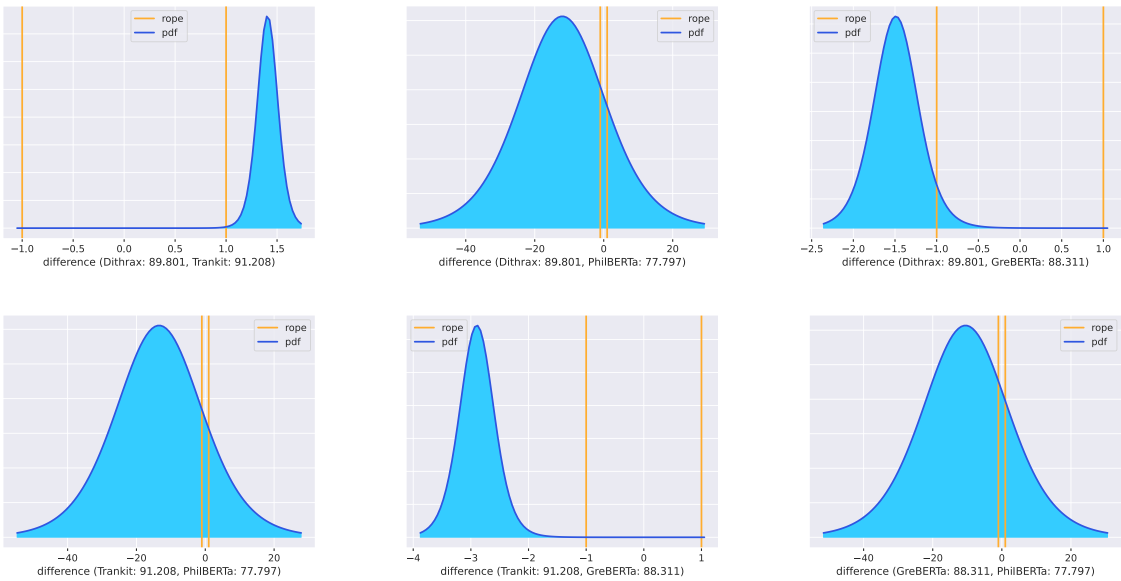


Figure 5: Posteriors of the Bayesian correlated t-test for all model pairs with reference to AllTags scores.

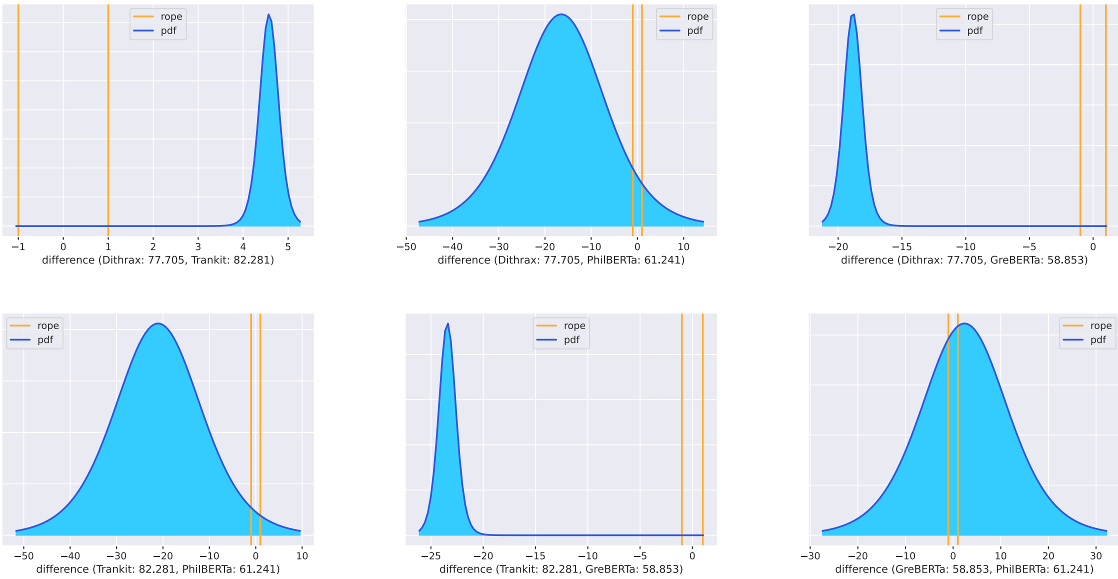


Figure 6: Posteriors of the Bayesian correlated t-test for all model pairs with reference to UAS scores.

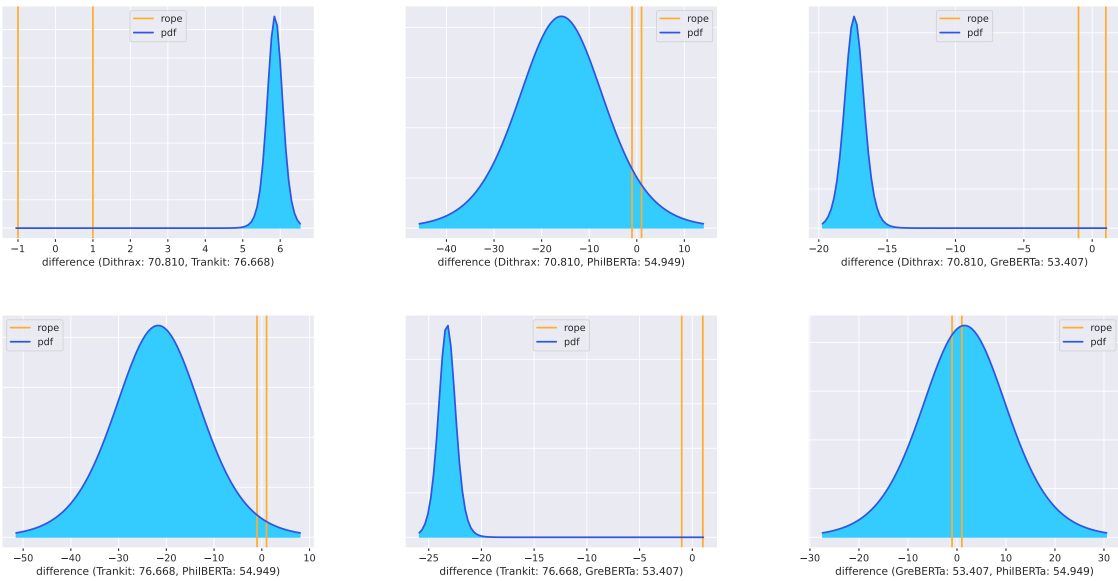


Figure 7: Posteriors of the Bayesian correlated t-test for all model pairs with reference to LAS scores.

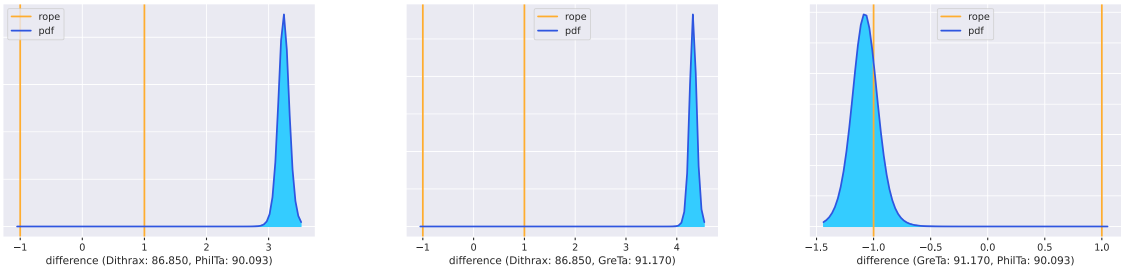


Figure 8: Posteriors of the Bayesian correlated t-test for all model pairs with reference to Lemmas scores.

E Texts

The following tables provide details of the texts used in the training, validation, and test sets (see also Table 3 for a more concise presentation). The authors, titles, and dates of each work were retrieved primarily from the file https://github.com/OperaGraecaAdnotata/OGA/tree/main/work_chronology/texts/chronology_greek_works.xml. This file contains work and title metadata derived from the canonical-greekLit[26] and First1KGreek[27] GitHub repositories, as well as from the Perseus Catalogue.[28] The dates of the works, expressed in ISO 8601 format, were manually annotated by a single annotator,[29] who used reference sources documented in the file mentioned above. All metadata should be regarded as work in progress.

[26] https://github.com/PerseusDL/canonical-greekLit.

[27] https://github.com/OpenGreekAndLatin/First1KGreek.

[28] https://catalog.perseus.org/.

[29] The annotator is an expert in AG literature and was paid fairly in accordance with German law.

CTS	Author	Title	Date	Tokens
tlg0003.tlg001	Thucydides	History of the Peloponnesian War	–0430-01/–0410-12	32, 344
tlg0005.tlgxxx	Theocritus	Fragments	–0299-01/–0259-12	304
tlg0006.tlg003	Euripides	Medea	–0430-01/–0430-12	9, 845
tlg0007.tlg004	Plutarch	Lycurgus	+0096-01/+0120-12	10, 709
tlg0007.tlg015		Alcibiades	+0096-01/+0120-12	11, 439
tlg0007.tlg086		On the Fortunes of the Romans	+0060-01/+0065-12	5, 232
tlg0007.tlg087		On the Fortune or the Virtue of Alexander I and II	+0096-01/+0120-12	9, 823
tlg0008.tlg001	Athenaeus of Naucratis	The Deipnosophists	+0175-01/+0200-12	45, 653
tlg0009.tlg001	Sappho	Fragments	–0699-01/–0599-12	4, 530
tlg0010.tlg002	Isocrates	Against Callimachus	–0401-01/–0401-12	4,109
tlg0010.tlg020		To Philip	–0345-01/–0345-12	466
tlg0011.tlg001	Sophocles	Trachiniae	–0449-01/–0449-12	9, 026
tlg0011.tlg002		Antigone	–0442-01/–0437-12	8, 990
tlg0011.tlg003		Ajax	–0438-01/–0435-12	9, 751
tlg0011.tlg004		Oedipus Tyrannus	–0418-01/–0415-12	11, 521
tlg0011.tlg005		Electra	–0417-01/–0406-12	10, 806
tlg0012.tlg001	Homer	Iliad	–0799-01/–0700-12	130, 479
tlg0012.tlg002		Odyssey	–0799-01/–0700-12	105, 612
tlg0013.tlg002	Homeric Hymns	Hymn 2 to Demeter	–0624-01/–0574-12	3, 968
tlg0014.tlg001	Demosthenes	First Olynthiac	–0348-01/–0348-12	2, 194
tlg0014.tlg004		First Philippic	–0350-01/–0350-12	3, 951
tlg0014.tlg007		On Halonnesus	–0342-01/–0341-12	2, 886
tlg0014.tlg017		On the Treaty with Alexander	–0330-01/–0330-12	2, 076
tlg0014.tlg018		On the Crown	–0329-01/–0329-12	26, 435
tlg0014.tlg027		Against Aphobus I	–0363-01/–0362-12	5, 346
tlg0014.tlg036		For Phormio	–0349-01/–0348-12	4, 649
tlg0014.tlg037		Against Pantaenetus	–0346-01/–0346-12	4, 528
tlg0014.tlg039		Against Boeotus I	–0347-01/–0346-12	3, 351
tlg0014.tlg041		Against Spudias	–0363-01/–0358-12	2, 333
tlg0014.tlg042		Against Phaenippus	–0329-01/–0329-12	2, 624
tlg0014.tlg045		Against Stephanus I	–0349-01/–0348-12	6, 839
tlg0014.tlg046		Against Stephanus II	–0349-01/–0348-12	2, 168
tlg0014.tlg047		Against Evergus and Mnesibulus	–0354-01/–0354-12	6, 235
tlg0014.tlg049		Apollodorus Against Timotheus	–0361-01/–0361-12	5, 005
tlg0014.tlg050		Apollodorus Against Polycles	–0359-01/–0359-12	5, 306
tlg0014.tlg051		On the Trierarchic Crown	–0359-01/–0357-12	1, 580
tlg0014.tlg052		Apollodorus Against Callippus	–0368-01/–0367-12	2, 490
tlg0014.tlg053		Apollodorus Against Nicostratus	–0367-01/–0366-12	2, 340

CTS	Author	Title	Date	Tokens
tlg0014.tlg054	Demosthenes	Against Conon	–0354-01/–0340-12	3,755
tlg0014.tlg057		Against Eubulides	–0345-01/–0344-12	5,498
tlg0014.tlg059		Theomnestus and Apollodorus Against Neaera	–0342-01/–0339-12	10,489
tlg0016.tlg001	Herodotus	Histories	–0429-01/–0424-12	33,150
tlg0017.tlg003	Isaeus	The Estate of Pyrrhus	–0388-01/–0388-12	4,959
tlg0019.tlg001	Aristophanes	Acharnians	–0424-01/–0424-12	8,984
tlg0019.tlg008		Thesmophoriazusae	–0410-01/–0410-12	9,073
tlg0020.tlg001	Hesiod	Theogony	–0899-01/–0700-12	8,234
tlg0020.tlg002		Works and Days	–0899-01/–0700-12	7,116
tlg0020.tlg003		Shield of Heracles	–0899-01/–0700-12	3,934
tlg0026.tlg001	Aeschines	Against Timarchus	–0345-01/–0344-12	15,971
tlg0027.tlg001	Andocides	On the Mysteries	–0399-01/–0398-12	5,964
tlg0028.tlg001	Antiphon	Against the Stepmother for Poisoning	–0419-01/–0410-12	2,046
tlg0028.tlg002		First Tetralogy	–0479-01/–0410-12	2,915
tlg0028.tlg005		On the Murder of Herodes	–0417-01/–0417-12	7,458
tlg0028.tlg006		On the Choreutes	–0418-01/–0418-12	4,014
tlg0032.tlg001	Xenophon	Hellenica	–0361-01/–0353-12	27,401
tlg0032.tlg002		Memorabilia	–0409-01/–0353-12	27,840
tlg0032.tlg004		Symposium	–0369-01/–0360-12	7,291
tlg0032.tlg006		Anabasis	–0379-01/–0359-12	18,737
tlg0032.tlg007		Cyropaedia	–0368-01/–0365-12	50,690
tlg0032.tlg008		Hiero	–0356-01/–0356-12	6,953
tlg0032.tlg015		Constitution of the Athenians	–0442-01/–0405-12	3,723
tlg0041.tlg001	Chion	Epistulae	+0001-01/+0100-12	5,577
tlg0058.tlg001	Aeneas Tacticus	Poliorcetica	–0374-01/–0349-12	7,207
tlg0059.tlg001	Plato	Euthyphro	–0398-01/–0346-12	6,349
tlg0059.tlg002		Apology	–0398-01/–0389-12	10,457
tlg0059.tlg003		Crito	–0398-01/–0389-12	5,093
tlg0059.tlg029		Cleiphon	–0398-01/–0346-12	1,875
tlg0060.tlg001	Diodorus of Sicily	Historical Library	–0059-01/–0029-12	25,692
tlg0061.tlg001	Lucian of Samosata	Asinus	+0125-01/+0180-12	11,054
tlg0081.tlg001	Dionysius of Halicarnassus	Antiquitates Romanae	–0007-01/–0006-12	30,312
tlg0085.tlg001	Aeschylus	Supplices	–0465-01/–0458-12	6,071
tlg0085.tlg002		Persians	–0471-01/–0471-12	6,381
tlg0085.tlg003		Prometheus Bound	–0459-01/–0455-12	7,222
tlg0085.tlg004		Seven against Thebes	–0466-01/–0466-12	6,372
tlg0085.tlg005		Agamemnon	–0457-01/–0457-12	10,037
tlg0085.tlg006		Libation Bearers	–0457-01/–0457-12	5,846
tlg0085.tlg007		Eumenides	–0457-01/–0457-12	6,518

CTS	Author	Title	Date	Tokens
tlg0086.tlg035	Aristotle	Politics	–0399-01/–0299-12	19,867
tlg0093.tlg009	Theophrastus	Characters	–0316-01/–0316-12	8,265
tlg0096.tlg002	Aesop	Aesop's Fables	–0599-01/–0500-12	5,221
tlg0255.tlg001	Mimnermus of Colophon	Fragmenta	–0699-01/–0599-12	213
tlg0260.tlg001	Semonides of Amorgos	Fragmenta	–0699-01/–0599-12	767
tlg0343.tlg001	Ezechiel	Exagoge	–0199-01/–0099-12	1,939
tlg0429.tlg001	Cephisodorus Comicus	Fragmenta	–0401-01/–0401-12	29,490
tlg0526.tlg004	Josephus Flavius	The Jewish War	+0075-01/+0075-12	24,987
tlg0527.tlg001	Septuaginta	Genesis	–0299-01/–0200-12	19,235
tlg0537.tlg012	Epicurus	Epistula ad Menoeceum	–0310-01/–0270-12	1,523
tlg0540.tlg001	Lysias	On the Murder of Eratosthenes	–0402-01/–0401-12	2,834
tlg0540.tlg012		Against Eratosthenes	–0402-01/–0402-12	5,638
tlg0540.tlg013		Against Agoratus	–0399-01/–0397-12	5,641
tlg0540.tlg014		Against Alcibiades 1	–0394-01/–0394-12	2,801
tlg0540.tlg015		Against Alcibiades 2	–0394-01/–0394-12	688
tlg0540.tlg019		On the Property of Aristophanes	–0386-01/–0386-12	3,624
tlg0540.tlg023		Against Pancleon	–0399-01/–0398-12	896
tlg0540.tlg024		On the Refusal of a Pension	–0402-01/–0402-12	1,665
tlg0541.tlg007	Menander of Athens	Dyscolus	–0315-01/–0315-12	8,069
tlg0543.tlg001	Polybius	Histories	–0167-01/–0117-12	105,693
tlg0544.tlg002	Sextus Empiricus	Adversus Mathematicos	+0201-01/+0300-12	16,218
tlg0548.tlg001	Apollodorus	Library	+0101-01/+0200-12	1,265
tlg0551.tlg017	Appianus of Alexandria	Civil Wars	+0101-01/+0200-12	25,665
tlg0554.tlg001	Chariton	De Chaerea et Callirhoe	+0075-01/+0125-12	6,265
tlg0557.tlg001	Epictetus	Discourses	+0108-01/+0108-12	7,204
tlg0559.tlg002	Hero of Alexandria	De Automatis	+0062-01/+0085-12	10,321
tlg0561.tlg001	Longus	Daphnis and Chloe	+0101-01/+0300-12	672
tlg0585.tlg001	Phlegon of Tralles	Book of Marvels	+0100-01/+0200-12	5,642
tlg1220.tlg001	Batrachomyomachia	Batrachomyomachia Homerica	–0099-01/–0029-12	2,212
tlg2003.tlg001	Julian	Panegyric in Honor of the Emperor Constantinus	+0355-01/+0355-12	1,405
tlgxxxx.tlgxxx	Paeanius	Brevarium	+0337-01/+0379-12	6,184

It takes a village to grammaticalize

Joseph Larson and **Patrícia Amaral**
Department of Spanish and Portuguese
Indiana University
joelarso@iu.edu
pamaral@iu.edu

Abstract

This paper investigates the grammaticalization of the noun *caleta* 'cove, village' to an intensifier, as part of the system of degree words in Chilean Spanish. We use word embeddings trained on a corpus of tweets to show the ongoing syntactic and semantic change of *caleta*, while also revealing how high degree is expressed in colloquial Chilean Spanish.

1 Introduction

Studies of language change using distributional methods have shown the potential of word embeddings (both static and contextualized) to trace syntactic and semantic change over time (Hamilton et al., 2016; Kutuzov et al., 2018; Periti et al., 2024, a.o.).[1] However, such research tends to focus on predicting changes that affect sets of lexical items shifting from one semantic domain to another, which typically reflects cultural and societal changes. Fewer studies have explored both semantic and morphosyntactic change (but see Fonteyn et al. 2022). In this paper, we focus on the semantic and syntactic shift from lexical to grammatical, known as grammaticalization (Meillet, 1912; Hopper and Traugott, 2003), and the stages of this process. Specifically, we study the creation of degree expressions like English *very, a lot*.

Traditionally, degree expressions have been associated with adjectives, considered the prototypical gradable category. However, degree modification is also compatible with nouns and verbs, which shows that gradability cuts across syntactic categories (Bolinger, 1972; Neeleman et al., 2004; Doetjes, 2008). As a word becomes a degree expression over time, it typically expands its distribution along different categories: e.g. it first combines with nouns before co-occurring with verbs and adjectives. Hence, the grammaticalization of degree expressions provides insight into the semantics of degree and patterns in the distribution of degree words (Amaral, 2016; Luo et al., 2019). This paper examines an understudied variety, Chilean Spanish, and uses word embeddings to investigate the emerging system of degree words to which one grammaticalized word shifts. We investigate the grammaticalization of *caleta* in Chilean Spanish, from a noun denoting 'cove, hiding place (where merchandise can be stored)','village', as in ex. (1), to a quantifier and degree adverb 'much, a lot', as in (2), where *caleta* modifies the verb and denotes high degree.

(1) Esta experiencia la realizamos
this experience CL.FEM.SG.ACC do.PST.1PL
en Zapallar, en la caleta de pescadores
in Zapallar in the caleta of fishermen
'We did this experience in Zapallar, in the fishermen's cove'

(2) me gustó caleta
CL.1SG.DAT like.PST.3SG caleta
'I liked it a lot.'

We use word embeddings to examine to what extent the grammaticalization of *caleta* has developed while also shedding light on the system of degree modifiers in Chilean Spanish. We ask, (i) how far along has *caleta* grammaticalized in Chilean Spanish, and (ii) what types of evidence do word embeddings provide of different stages of grammaticalization of degree words?

2 Previous Work

Linguists have provided analyses of the gradual process by which lexical items acquire grammatical functions: for example, in this diachronic change, nouns lose their categorial properties like occurring after a determiner or being pluralized.

[1] For a recent state-of-the-art survey comparing different approaches to semantic change using large language models, see Periti and Montanelli 2024.

Proceedings of the First Workshop NLP and Language Models for Digital Humanities associated with RANLP 2025, pages 66–74, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-106-6-006

The grammaticalization of nouns into degree adverbs (e.g. the development from *lot* 'a set of objects' to *a lot* 'much') is well attested cross-linguistically: other examples are French adverb *beaucoup* from *un beau coup* 'a good strike' and English *a bit* from 'a bite, a portion that fits in the mouth' (Abeillé et al., 2004; Marchello-Nizia, 2006; Verveckken, 2012; Traugott, 2008; Amaral, 2020).

This research has shown that a typical structure in which nouns occur - modification by a prepositional phrase, as in *a lot* [$_{PP}$ *of chairs], a mountain* [$_{PP}$ *of books]* - provides a starting point for quantity and degree interpretations. This structure undergoes subsequent syntactic reanalysis, where the head noun (e.g. *lot*) loses nominal properties and *a lot of* becomes an adverb modifying the second noun. The development of so-called binominal structures Det N_1 of N_2, which may or may not further evolve to a fully adverbial category, plays a crucial role in the grammaticalization of degree words. In our study, we also include the structure *(Det) caleta of N*, hence we investigate the distribution of *caleta de*.

As argued by Doetjes, 2008, degree words across languages show a systematic behavior in terms of the words they can modify. These well-attested patterns correspond to types along a continuum of syntactic-semantic word classes, where a degree expression can modify all word classes (like French *trop*, type C) or just a subset of classes, gradable adjectives (like English *very*, type A), see Figure 1. As words develop into one type, they are predicted to modify words in the order along the continuum; for instance, if a word co-occurs with words of category V, it is expected to co-occur with words of category IV before it appears with words of category III. [2] As we investigate whether *caleta* has grammaticalized into a degree word, we will examine its stage of development with respect to Doetjes' continuum.

While some computational studies of grammaticalization have adopted case-driven approaches similar to ours (Fonteyn and Manjavacas, 2021; Amaral et al., 2023; Nagata et al., 2024), we also investigate how a distributional analysis of *caleta* can provide insight on the set of degree expressions currently used in colloquial Chilean Spanish. In other words, we aim to examine not just the grammaticalization of *caleta* but also how this word fits in the system of degree words in Chilean Spanish and types of degree expressions across languages.

[2](Doetjes, 2008) differentiates between 'gradable' and 'eventive' adjectives and verbs by whether or not the modifier is targeting the degree or is quantifying over events. The example she gives is from Dutch: *Jan is veel ziek* 'Jan is sick a lot' vs. *Jan is erg ziek* 'Jan is very sick.' In the former, *veel* as a quantifier targets eventive adjectives, thus it can only modify the quantity of sick events. In the latter, *erg* expresses the degree of sickness, i.e. the severity of his illness.

<table>
<tr><th>Category</th><th>Word Class</th><th></th><th></th><th></th><th></th><th></th></tr>
<tr><td>I</td><td>gradable adjectives</td><td>Type A
very^E</td><td rowspan="3">Type B
erg^D
očen'^{R}</td><td rowspan="6">Type C
trop^F
muito^P
molto^I</td><td rowspan="6"></td><td rowspan="7"></td></tr>
<tr><td>IIa</td><td>gradable nominal predicates</td><td rowspan="7">Type D
beaucoup^F
a lot^E</td></tr>
<tr><td>IIb</td><td>gradable verbs</td></tr>
<tr><td rowspan="3">III</td><td>eventive verbs</td><td rowspan="5">Type E
veel^D
mnogo^R</td></tr>
<tr><td>eventive adjectives</td></tr>
<tr><td>comparatives</td></tr>
<tr><td>IV</td><td>mass nouns</td><td rowspan="2">Type F
a mountain^E</td></tr>
<tr><td>V</td><td>plural nouns</td><td>Type G
many^E</td></tr>
</table>

Figure 1: Typology of degree expressions according to their distribution along a continuum of word classes. Table adapted with modifications from (Doetjes, 2008, 138). Superscripts indicate language: R for Russian, D for Dutch, F for French, E for English, P for Portuguese, and I for Italian.

3 Methodology

3.1 Corpus Creation

To ensure we had a good representation of colloquial Chilean Spanish, we created a subcorpus from an already existing corpus of online data (Ortiz-Fuentes, 2023). The already existing corpus contained roughly 19GB of data, from diverse sources, including news, tweets, online reviews and other miscellaneous web content. We chose to create a subcorpus just from tweets to reduce the computational load for our later experiments and since we only wanted informal instances of language; *caleta* typically only occurs in less formal registers. The resulting subcorpus of $27,306,582$ tweets consisted of exactly $342,979,307$ tokens. The time span of these tweets is from 2010 to 2020.

3.2 Preprocessing

We first normalized the text in the corpus: we removed case, punctuation, diacritics, URLs, hashtags, and any repeated letters. For this last step, we only allowed double letters where they occur within normative Spanish orthography (i.e. $< r >$, $< c >$, $< l >$), elsewhere only single letters were allowed. Then we input the corpus into a plain text file separated by newlines. The resulting file was then lemmatized using SpaCy's Spanish lemmatizer (Honnibal et al., 2020).

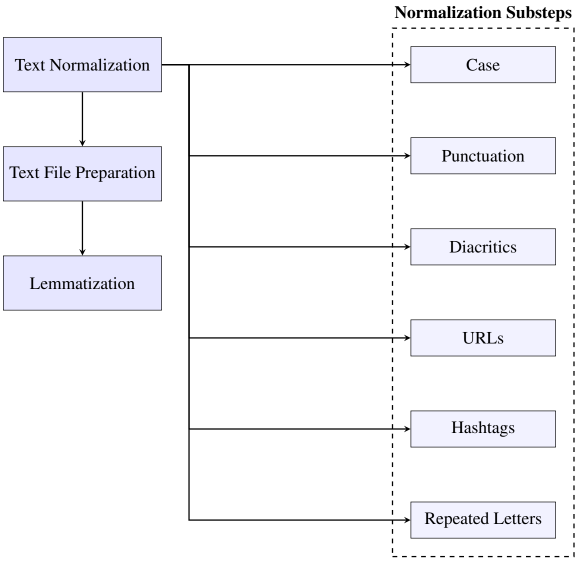


Figure 2: Preprocessing steps.

3.3 Model Selection

To represent the distributional patterns of words in our corpus, we decided to use static word embeddings over contextualized word embeddings. Non-contextualized embeddings allow us to compare our target word with other words in Chilean Spanish to examine the current stage of grammaticalization of *caleta* as determined by its closeness to different subsystems in the language.

The algorithm we use is Skip-Gram with Negative Sampling (SGNS) implemented in word2vec (Mikolov et al., 2013) to extract embeddings, based on previous research that showed good results for studies of semantic change (Hu et al., 2022, a.o.). For this reason, we do not consider it necessary to use a more computationally expensive operation (e.g. dynamic word embeddings). We trained each model for five epochs, a minimum token count of 10 and the skip-gram algorithm. Initially, we experimented with several hyperparameters: the window size, the minimal word count and the vector size. The only hyperparameter that proved to be significant was the window size (see next section for more details). The resulting model used a vector length of 100 and a minimal word count of 10. To verify the validity of the model, we used analogy tests targetting gender-based morphological and semantic relations (see Table 1 for specifics). We found the analogy used always 100% accurate.

Relationship	Word Pair 1		Word Pair 2	
	Word A	Word B	Word A	Word B
Age-based	*Hombre* 'Man'	*Mujer* 'Woman'	*Niño* 'Boy'	*Niña* 'Girl'
Familial	*Padre* 'Father'	*Madre* 'Mother'	*Hijo* 'Son'	*Hija* 'Daughter'
Feline	*Niño* 'Boy'	*Gato* 'Cat (male)'	*Niña* 'Girl'	*Gata* 'Cat (female)'
Canine	*Niño* 'Boy'	*Perro* 'Dog (male)'	*Niña* 'Girl'	*Perra* 'Dog (female)'

Table 1: The four analogy tests used to validate Word2Vec model. The equation used was $WB_2 = WA_1 - WA_2 + WB_1$.

3.4 Window Size

As mentioned in the previous section, the only hyperparameter we adjusted for the model was the window size. We extracted models for $w = [1, 10]$. Our hypothesis was that lower window sizes would be more adequate for showing grammaticalization, since the scope of grammatical words like quantifiers lies within its immediate neighbors, whereas higher window sizes show neighbors within the same semantic field (therefore its lexical use). However, since we use a corpus of tweets, window size is fairly limited by the genre itself (a possible limitation we address later).

4 Results

4.1 *Caleta*

Here we display only the results of the experiments with a small ($w = 1$) and a large ($w = 10$) window size. This allows us to compare the information obtained by manipulating this parameter. In Figure 3, the word embeddings show both neighbors of the lexical noun and neighbors of the degree word. Nearest neighbors of the noun are toponyms (i.e. names of villages) and other nouns with related meanings (e.g. *playa* 'beach' and *balneario* 'bathhouse'). As for the neighbors of the degree word, we find degree expressions, both adverbs and quantifiers like *mucho* and *ene*, both meaning 'a lot'. *Caleta de* also appears among the neighbors (please see subsequent section for these results).

The co-occurrence of neighbors of both meanings shows that *caleta* has partially grammaticalized; it still retains its lexical use as a noun. These findings provide evidence for a situation of layering (Hopper, 1991), i.e. the synchronic coexistence of older and more recent functions of a form in a language.

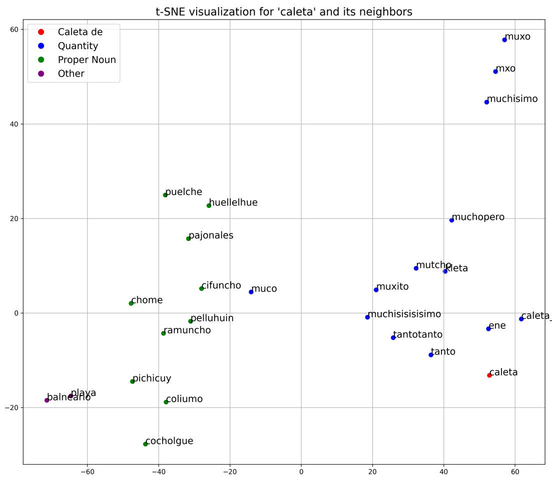


Figure 3: TSNE representation of *caleta* and its top 25 neighbors. Embeddings were created with a window size of 1. Blue corresponds to words that are quantifiers, green corresponds to toponyms (i.e. names of villages), and purple corresponds to semantically related nouns.

If we now use a larger window size, the results are different, with more neighbors associated with the lexical item. In Figure 4 we find the plural noun (*caletas*); as mentioned in historical analyses, the ability to be pluralized is a syntactic property of nouns. This attests to the persistence of some nominal categorial properties of *caleta*. We also find the noun *pescadores* 'fishermen', as the noun *caleta* typically refers to a village of fishermen and hence the nouns often co-occur (in *caleta de pescadores*), and related nouns like *muelle* 'pier' and *poza* 'puddle'.

4.2 *Caleta de*

We analyzed the results of *caleta de* separately from those of *caleta* since the former is the vestige of a binominal quantifier preceding the grammaticalization of the latter. Figure 5 and Figure 6 show the TSNE representations of the nearest neighbors of *caleta de*. For the smaller window size, we see other quantifiers like *ene* (more in the next section), *caleta*, etc. The majority of neighbors here are

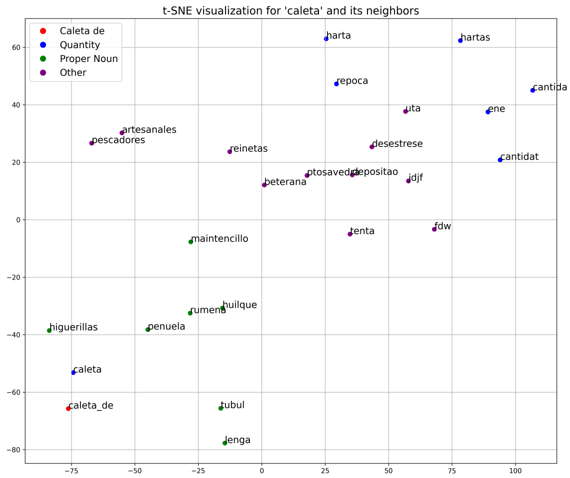


Figure 4: TSNE representation of *caleta* and its top 25 neighbors. Embeddings were created with a window size of 10. Blue corresponds to words that are quantifiers, green corresponds to toponyms (i.e. examples of *caletas*), and purple corresponds to semantically related nouns.

quantifiers in their orthographical variants found in tweets (e.g. *mucho*, *mxo*, *nucho*, etc). Two other words that form part of binominal quantifiers are also present, *monton* and *montones*, both meaning 'pile' and 'piles', but which have grammaticalized in the same fashion as *caleta* to denote a large quantity (*un montón de N* 'a lot of N'). In this window size, only one proper noun is present, *Chorromil*, the name of a village. Lastly, we find other quantifiers, like *cualquiers* and *cualesquiers*, both orthographical variations of *cualquier*, 'whichever', and *puras*, a determiner in Chilean Spanish.

In the larger window size, we see *caleta* as its nearest neighbor. Other quantifiers like *mucho*, *ene*, *harto*, etc. are present, but they are much further away than semantically related nouns like *pescadores* 'fishermen', *artesanales* 'craftsmen', *reinetas*, a plural noun denoting a variety of white fish, as well as toponyms that are names of *caletas*. These results show once more how important the hyperparameter of window size is in capturing the grammatical meaning of relatively newly grammaticalized words in a language.

4.3 *Ene*

We decided to display the top 10 neighbors for the word *ene*, since *ene* always appeared as a top neighbor for *caleta* and *caleta de*. *Ene* comes from the Spanish pronunciation of the grapheme $< n >$ and is used in Mathematics to denote an unspeci-

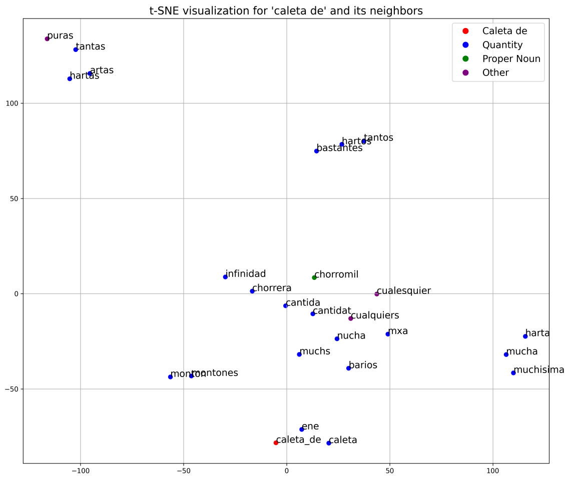


Figure 5: TSNE representation of *caleta de* and its top 25 neighbors. Embeddings were created with a window size of 1. Blue corresponds to words related to quantity, green corresponds to toponyms (i.e. examples of *caletas*), and purple corresponds to syntactically and semantically-related words.

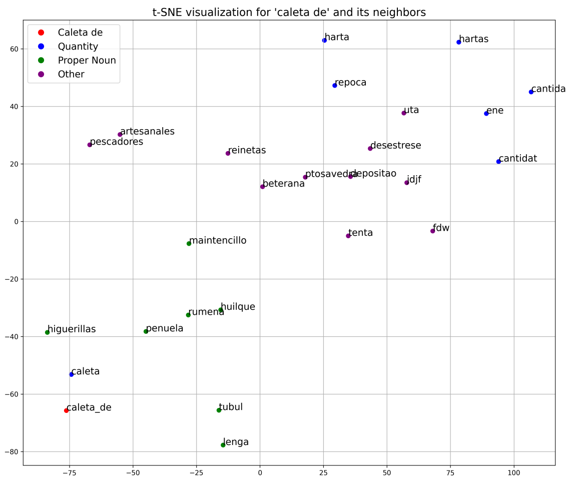


Figure 6: TSNE representation of *caleta de* and its top 25 neighbors. Embeddings were created with a window size of 10. Blue corresponds to words related to quantity, green corresponds to toponyms (i.e. examples of *caletas*), and purple corresponds to syntactically and semantically-related words.

fied integer. Over time, in this variety of Spanish *ene* has grammaticalized like *caleta* to denote a large quantity and high degree. Our results show that *ene* is another example of a grammaticalized degree word, albeit in a different stage of grammaticalization. To the best of our knowledge, this has not been observed or studied. Example (3) shows a lexical use of *ene*, taken from the Dictionary of the Spanish Real Academy (Real Academia Española, 2025), since no such example could be found in our corpus. Example (4) shows the degree adverb (here, modifying a verb), i.e. the grammaticalized item. Lastly, example (5) shows *ene* in combination with *ctm*, a commonly used abbreviation of the phrase *concha (de) tu madre* (literally 'your mother's pussy'), which is used as a vulgar intensifier similar to *fucking* in English.

(3) El fenómeno se repite
The phenomenon CL.REFL repeat.PRS.3SG
ene veces.
n times
'The phenomenon is repeated *n* times.'

(4) me gustó ene
CL.1SG.DAT like.PST.3SG ene
'I liked it a lot.'

(5) me gustó ene ctm
CL.1SG.DAT like.PST.3SG ene ctm
'I fucking liked it a lot.'

Table 2 and 3 show the closest neighbors for *ene* in our corpus. For both window sizes, none of the neighbors are semantically related to Mathematics, which would be expected if *ene* still retained some of its original lexical meaning. For the smaller window size, all of the neighbors are degree words meaning 'much' (including the noun *cantidad* which can appear in a binominal structure *cantidad de N* 'a large quantity of N'). For the larger window size, half of the neighbors are quantifiers. We also see the expressive *puxis* (an orthographical variation of *pucha*, meaning 'darn'), spellings of laughter and the vulgar term *autodelicioso*. This is evidence for what has been described previously in the literature that degree modifiers, as highly volatile units of language, are subject to change easily and become expressives (Ito and Tagliamonte, 2003).

Rank	Word	Score
1	***caleta de*** 'a lot of'	0.78
2	***cantitat*** (*cantidad*, orthographical variation, 'quantity')	0.67
3	***harto*** 'a lot'	0.66
4	***caleta*** 'a lot' or 'village'	0.66
5	***kleta*** 'caleta' (orthographical variation)	0.65
6	***arto*** 'harto' (orthographical variation)	0.64
7	***mucho*** 'a lot'	0.64
8	***tanto*** 'so much'	0.63
9	***mxo*** 'mucho' (orthographical variation)	0.62
10	***muchopero*** (*mucho pero* as one word, 'a lot but...')	0.61

Table 2: Ranked words with their scores (cosine) for *ene* for $w = 1$

Rank	Word	Score
1	***kleta*** (orthographical variation of *caleta*)	0.71
2	***caleta de*** ‘a lot of’	0.68
3	***cantitat*** (*cantidad*, orthographical variation, ‘quantity’)	0.67
4	*graziash* (*gracias*, orthographical variation, ‘thanks’)	0.66
5	*jsjsjd* ‘laughter’	0.66
6	***harto*** ‘a lot’	0.66
7	*puxis* (orthographic variation of *pucha*, ‘darn’)	0.66
8	*autodelicioso* (lit. ‘self-delicious’, term used for masturbation)	0.64
10	***muchosaño*** (*muchos años* as one word, ‘many years’)	0.63

Table 3: Ranked words with their scores (cosine) for *ene* for $w = 10$. Bold words correspond to quantifiers.

4.4 Other quantifiers

Lastly, we show word embeddings of other degree words, in this case ‘stable’ quantifiers in Chilean Spanish: *harto* ‘a lot’, *mucho* ‘a lot’, *tanto* ‘so many.’ It is worth mentioning that unlike *caleta*, *caleta de* and *ene* (which syntactically can be considered degree adverbs), these quantifiers inflect for gender and number when modifying a noun. The purpose of using the lemmatizer was to control for this, but as the results show, some inflected tokens of these quantifiers were not properly lemmatized.

Tables 4, 5, 6, 7, 8 and 9 show the nearest neighbors for *harto*, *mucho* and *tanto* at the two window sizes. For *harto*, we see that the majority of its neighbors are other quantifiers for both window sizes, as well as orthographical variations (e.g. *harrto*, *arto*) and inflected versions of the lexeme, like the feminine form *harta*. Likewise, *tanto* as its neighbors for the smaller window size shows mostly orthographical variations (e.g. *tsnto*, *tabto*), while for the larger window size we can see similar results to *ene*, where nouns like ‘laughter’ are amongst the neighbors. For *mucho*, we can see mostly orthographical variants for the smaller window size (e.g. *muxo*, *muxho*) and for the larger window size we see less orthographical variations and more of other quantifiers, even its antonym *poco*, which also occurs with intensifying affixes: *re-poco* and *poc-azo* ‘very little’.

5 Discussion

Our word embedding results for *caleta* show that nowadays the word is used to express high degree. In addition, in our results both the lexical noun and the degree modifier are present. The choice of hyperparameters, specifically window size, has important consequences: a small window size yields nearest neighbors for both forms, while a larger window size results in more neighbors of

Rank	Word (Gloss)	Score
1	***arto*** ‘harto’ (orthographical variation)	0.94
2	***mucho*** ‘a lot’	0.84
3	***bastante*** ‘quite’	0.78
4	***harrto*** ‘harto’ (orthographical variation)	0.74
5	***mxo*** ‘mucho’ (orthographical variation)	0.72
6	***muchisimo*** ‘mucho’ (superlative)	0.71
7	***muxo*** ‘mucho’ (orthographical variation)	0.69
8	***mutcho*** ‘mucho’ (orthographical variation)	0.68
9	***mucjo*** ‘mucho’ (orthographical variation)	0.67
10	***nucho*** ‘mucho’ (orthographical variation)	0.66

Table 4: Ranked words with their scores (cosine) for *harto* for $w = 1$. Bold words correspond to quantifiers.

Rank	Word (Gloss)	Score
1	***arto*** ‘harto’ (orthographical variation)	0.81
2	***mucho*** ‘a lot’	0.72
3	*sosi* (*eso sí*, abbreviation, ‘though’)	0.69
4	***bastante*** ‘quite’	0.68
5	***harta*** ‘a lot’	0.68
6	***ene*** ‘a lot’	0.66
7	*pucha* ‘darn’	0.63
8	***haarto*** ‘harto’ (orthographical variation)	0.63
9	***repoco*** ‘poco’ (intensifier)	0.63
10	***pocazo*** ‘poco’ (augmentative)	0.61

Table 5: Ranked words with their scores (cosine) for *harto* for $w = 10$. Bold words correspond to quantifiers.

Rank	Word (Gloss)	Score
1	***tsnto***‘tanto’ (orthographical variation)	0.76
2	***demasia*** (*demasiado*, phonetic variation, ‘too much’	0.70
3	***tantotanto*** ‘tanto’ (repeated)	0.69
4	***mucho*** ‘a lot’	0.69
5	***tantoy*** (*tanto y* as one word, ‘so much and’)	0.69
6	***tabto*** ‘tanto’ (orthographical variation)	0.68
7	***tantisimo*** ‘tanto’ (superlative)	0.67
8	***tnto*** ‘tanto’ (orthographical variation)	0.64
9	***tanro*** ‘tanto’ (orthographical variation)	0.64
10	***mutcho*** ‘mucho’ (orthographical variation)	0.64

Table 6: Ranked words with their scores (cosine) for *tanto* for $w = 1$. Bold words correspond to quantifiers.

Rank	Word (Gloss)	Score
1	***mucho*** ‘a lot’	0.71
2	***tsnto*** ‘tanto’ (orthographical variation)	0.65
3	***tantotanto*** ‘tanto’ (repeated)	0.63
4	**tantisimo** ‘tanto’ (superlative)	0.60
5	***simuchas*** (*sí muchas* as one word, ’yes a lot’)	0.60
6	*jskdkd* ‘laughter’	0.60
7	*jajajajajajaun* ‘laughter’	0.60
8	***muchogracias*** (muchas gracias as one word, ‘thanks a lot’)	0.59
9	*tisin* (*tí sin* as one word, ‘you (prepositional), without)	0.58
10	*pueso* (portmanteau of *pues eso*, ‘exactly’)	0.58

Table 7: Ranked words with their scores (cosine) for *tanto* for $w = 10$. Bold words correspond to quantifiers.

the lexical noun. We hypothesize that this is due to the fact that as a degree word, *caleta* is a modi-

Rank	Word (Gloss)	Score
1	***muchisimo*** 'mucho' (superlative)	0.91
2	***mxo*** 'mucho' (orthographical variation)	0.88
3	***harto*** 'a lot'	0.82
4	***muxo*** 'mucho' (orthographical variation)	0.81
5	***mucjo*** 'mucho' (orthographical variation)	0.80
6	***muchi*** 'mucho' (diminutive)	0.77
7	***muho*** 'mucho' (orthographical variation)	0.77
8	***muxho*** 'mucho' (orthographical variation)	0.77
9	***arto*** 'harto' (orthographical variation)	0.76
10	***nucho*** 'mucho' (orthographical variation)	0.75

Table 8: Ranked words with their scores (cosine) for *mucho* for $w = 1$. Bold words correspond to quantifiers.

Rank	Word (Gloss)	Score
1	***muchisimo*** 'mucho' (superlative)	0.79
2	***harto*** 'a lot'	0.74
3	***tanto*** 'so much'	0.71
4	***poco*** 'a little'	0.67
5	***muchoy*** (*mucho y* as one word, 'a lot and'	0.65
6	***muccho***'mucho' (orthographical variation)	0.65
7	***bastante*** 'quite'	0.65
8	***muchopero*** (*mucho pero* as one word, 'a lot but')	0.64
9	*aunpero* (*aún pero* as one word, 'still but')	0.63
10	***muchisisisimo*** 'mucho' (repeated superlative)	0.61

Table 9: Ranked words with their scores (cosine) for *mucho* for $w = 10$. Bold words correspond to quantifiers.

fier, and occurs in close adjacency to the modified word. Hence, a small window captures this distribution. On the other hand, as a lexical noun *caleta* is less syntactically constrained, with more positional freedom and semantic content.

While cosine similarity scores give us insight into a changing word's distribution, they alone do not tell us about its syntactic properties in detail. To better understand *caleta*'s current status as a degree modifier, we performed a *post-hoc* analysis of the top 20 collocates of *caleta* and *caleta de*. We looked specifically at the top tokens that immediately precede and proceed the two strings in our unlemmatized corpus. We were interested in the kinds of words that *caleta* and *caleta de* have come to modify, in accordance to Doetjes's typology of degree modifiers (see Section 2).

Our analysis shows that *caleta* has evolved extensively beyond its original lexical usage, wherein it was only compatible with count nouns that were semantically related e.g. *pescadores* 'fishermen' *camarones* 'shrimp (plural)', headed by the preposition *de*. The structure *caleta de* is now compatible with count nouns beyond the semantic domain of a fishing village: *años* 'years', *veces* 'times/instances' (see (6)), as well as mass nouns e.g. *plata* 'money (informal), *tiempo* 'time' (see (7)). It can also modify comparatives e.g. *mejor* 'better', *peor* 'worse' (see (9)); eventive verbs e.g. *dormir* 'to sleep', *reír* 'to laugh' (see (8)); gradable verbs *gustar* 'to like', *querer* 'to want' (see (2); and finally gradable nominal predicates[3] e.g. *hambre* 'hunger', *pena*, 'sorrow', as in (10).

(6) Hace caleta de años
make.PRS.3SG caleta of years
'Many years ago'

(7) es caleta de plata
be.PRS.3SG caleta of money
'it's a lot of money.'

(8) Yo igual reí caleta.
1SG.NOM same laugh.PST.1SG caleta
'I laughed a lot, anyway.'

(9) hay que cuidarse
be.existential.PRS.3SG that care.INF.REF
caleta mejor...
caleta better
'one has to take care of themselves much better.'

(10) Hace caleta de frío.
make.PRS.3SG caleta of coldness
'It's really cold.'

There were no cases of *caleta* modifying either eventive adjectives or gradable adjectives within our corpus. This, according to Doetjes's classification, indicates that *caleta* has evolved into a type D degree modifier. Figure 7 shows *caleta*'s position in this typology, in comparison to the other degree expressions in Chilean Spanish that we have discussed in this paper. Our results align with claims in the literature that Type C and D are the most common in the Romance languages (Doetjes, 2008). Lastly, within our results, *caleta* has no nearest neighbors with Type A modifiers (e.g. *muy* 'very'), which combine exclusively with gradable adjectives. This is not surprising since Type A modifiers have no overlap in word classes with Type D modifiers; their distributions are disjoint. This highlights how embeddings capture syntactic properties of words, as opposed to just similarity of meaning.

Our study has two main findings, which answer the research questions above. First, we have shown that *caleta* is undergoing grammaticalization: both

[3]Gradable nominal predicates, in Doetjes's definition, are nouns that are generally the objects of light verb expressions. The examples she gives are from French e.g. *Elle a très soif* 'She is very thirsty.' In Spanish, such light verb constructions also exist, so we consider cases like *tener sed* 'to be thirsty (lit. to have thirst)' to also be examples of nominal predicates.

<table>
<tr><th>Category</th><th>Word Class</th><th></th><th></th><th></th><th></th><th></th></tr>
<tr><td>I</td><td>gradable adjectives</td><td>Type A</td><td rowspan="3">Type B</td><td rowspan="6">Type C
harto
bastante
demasiado</td><td rowspan="6">Type F
un montón
cantidad
montones</td><td rowspan="5"></td></tr>
<tr><td>IIa</td><td>gradable nominal predicates</td><td rowspan="5">Type D
caleta
ene
mucho
tanto</td></tr>
<tr><td>IIb</td><td>gradable verbs</td></tr>
<tr><td>III</td><td>eventive verbs
eventive adjectives
comparatives</td><td rowspan="3">Type E</td></tr>
<tr><td>IV</td><td>mass nouns</td></tr>
<tr><td>V</td><td>plural nouns</td><td>Type G
vario</td></tr>
</table>

Figure 7: Degree words found in our results and their corresponding types according to Doetjes' model; modified table from (Doetjes, 2008, 138)

the older and the new meaning are captured by the word embeddings. Importantly, we see a difference in the results depending on the window size, when compared to other degree words which are grammatical items and not undergoing change, like *mucho* and *harto*. In the latter case, window size does not significantly impact the neighbors. Additionally, our *post-hoc* analysis provided insight on the properties of *caleta* as a degree word.

Second, our word embeddings have allowed us to reveal the inventory of degree words in colloquial Chilean Spanish, including a word that to date had never been investigated, *ene*. These words denote high degree (intensifiers), words that are known to change rapidly due to social and expressive pressure (Ito and Tagliamonte, 2003). Since *caleta* and *ene* are not normative forms, they are left out of traditional studies. This entails that we may miss instances of change possibly of interest to current linguistic theory. Hence, word embeddings can be a tool to study lesser-known subsystems of a language and capture ongoing changes in synchrony.

6 Conclusion

Our study contributes to studies of language change by analyzing intensifiers in colloquial Chilean Spanish (an understudied variety) from the past twenty years. We reveal an ongoing change that had not been previously studied. Using spontaneous speech from tweets, we gained access to informal speech where speakers communicate in an unedited way, which has allowed us to study the use of older and more recent degree expressions. Hence, our study shows how Digital Humanities as an interdisciplinary field can expand our knowledge of low-resource language varieties. In our specific case, the examination of the data through language processing revealed instances of grammaticalization that to the best of our knowledge had not been analyzed before.

We have shown that static word embeddings provide evidence for this change and can reveal meaning relations not previously studied. Moreover, we show that different choices of hyperparameters have an effect on which meaning (the lexical vs. the grammatical) of the word undergoing change, *caleta*, is represented.

Some limitations of our study are due to the genre itself. One such limitation is the difficulty with lemmatization: as we have mentioned, these are tweets, so we find strings that do not conform to normative orthography (for example, typos, abbreviations etc), therefore the lemmatizer has difficulty with detecting words of the same lexeme. In addition, Tweeter users tend to adopt orthographical forms that reflect pronunciation and sometimes are intended to be expressive, like repeating vowels in a word to express a very high degree. Furthermore, using a corpus of tweets means that the character limit has an impact on the possible window sizes. To obviate this problem, further studies on *caleta* could use longer texts that have the same register as tweets, e.g. blog posts.

References

Anne Abeillé, Olivier Bonami, Danièle Godard, and Jesse Tseng. 2004. The Syntax of French de-N' Phrases. *Proceedings of the International Conference on Head-Driven Phrase Structure Grammar*, pages 6–26.

Patrícia Amaral. 2016. When Something Becomes a Bit. *Diachronica*, 33:151–186.

Patrícia Amaral. 2020. Bocado: Scalar Semantics and Polarity Sensitivity. *Zeitschrift für romanische Philologie*, 136(4):1114–1136.

Patrícia Amaral, Hai Hu, and Sandra Kübler. 2023. Tracing semantic change with distributional methods: The contexts of algo. *Diachronica*, 40(2):153–194.

Dwight Bolinger. 1972. *Degree Words*. De Gruyter Mouton, Berlin, Boston.

Jenny Doetjes. 2008. Adjectives and Degree Modification. In *Adjectives and Adverbs: Syntax, Semantics,* *and Discourse*, pages 123–155. Oxford University Press.

Lauren Fonteyn and Enrique Manjavacas. 2021. Adjusting scope: a computational approach to case-driven research on semantic change. In *Proceedings of the Workshop on Computational Humanities Research (CHR 2021)*, volume 2898 of *CEUR Workshop Proceedings*, pages 280–298.

Lauren Fonteyn, Enrique Manjavacas, and Sara Budts. 2022. Exploring morphosyntactic varation and change with distributional semantic models. *Journal of Historical Syntax*, 6:1–41.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2116–2121, Austin, Texas. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adrienne Boyd. 2020. spacy: Industrial-strength natural language processing in python. *The Journal of Open Source Software*, 5(53):2914.

Paul Hopper. 1991. On some principles of grammaticization. In *Approaches to Grammaticalization*, pages 17–35. Benjamins.

Paul J. Hopper and Elizabeth Closs Traugott. 2003. *Grammaticalization*, 2 edition. Cambridge Textbooks in Linguistics. Cambridge University Press.

Hai Hu, Patrícia Amaral, and Sandra Kübler. 2022. Word embeddings and semantic shifts in historical spanish: Methodological considerations. *Digital Scholarship in the Humanities*, 37(2):441–461.

Rika Ito and Sali Tagliamonte. 2003. Well weird, right dodgy, very strange, really cool: Layering and recycling in english intensifiers. *Language in Society*, 32(2):257–279.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yiwei Luo, Dan Jurafsky, and Beth Levin. 2019. From insanely jealous to insanely delicious: Computational models for the semantic bleaching of English intensifiers. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 1–13, Florence, Italy. Association for Computational Linguistics.

Christiane Marchello-Nizia. 2006. *Grammaticalisation et changement linguistique*. De Boeck.

Antoine Meillet. 1912. L' évolution des formes grammaticales. *Scientia*, 12:130–148.

Tomas Mikolov, Kai Chen, G.s Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *Proceedings of Workshop at ICLR*, 2013:1–12.

Ryo Nagata, Yoshifumi Kawasaki, Naoki Otani, and Hiroya Takamura. 2024. A Computational Approach to Quantifying Grammaticization of English Deverbal Prepositions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 211–220, Torino, Italia. ELRA and ICCL.

Ad Neeleman, Hans Van de Koot, and Jenny Doetjes. 2004. Degree expressions. *The Linguistic Review*, 21(1):1–66.

Jorge Ortiz-Fuentes. 2023. Chilean Spanish Corpus.

Francesco Periti, Pierluigi Cassotti, Haim Dubossarsky, and Nina Tahmasebi. 2024. Analyzing semantic change through lexical replacements. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4495–4510, Bangkok, Thailand. Association for Computational Linguistics.

Francesco Periti and Stefano Montanelli. 2024. Lexical semantic change through large language models: a survey. *ACM Comput. Surv.*, 56(11).

Real Academia Española. 2025. Diccionario de la lengua española.

Elizabeth Traugott. 2008. Grammaticalization, Constructions and the Incremental Development of Language: Suggestions from the Development of Degree Modifiers in English. *Variation, Selection, Development: Probing the Evolutionary Model of Language Change*, pages 219–250.

Katrien Verveckken. 2012. Towards a Constructional Account of High and Low Frequency binominal Quantifiers in Spanish. *Cognitive Linguistics*, 23(2).

Evaluating LLMs for Historical Document OCR: A Methodological Framework for Digital Humanities

Maria Levchenko
Italian Institute of Germanic Studies (IISG), Rome, Italy
University of Bologna, Bologna, Italy
marylevchenko@gmail.com

Abstract

Digital humanities scholars increasingly use Large Language Models for historical document digitization, yet lack appropriate evaluation frameworks for LLM-based OCR. Traditional metrics fail to capture temporal biases and period-specific errors crucial for historical corpus creation. We present an evaluation methodology for LLM-based historical OCR, addressing contamination risks and systematic biases in diplomatic transcription. Using 18th-century Russian Civil font texts, we introduce novel metrics including Historical Character Preservation Rate (HCPR) and Archaic Insertion Rate (AIR), alongside protocols for contamination control and stability testing. We evaluate 12 multimodal LLMs, finding that Gemini and Qwen models outperform traditional OCR while exhibiting ”over-historicization”—inserting archaic characters from incorrect historical periods. Post-OCR correction degrades rather than improves performance. Our methodology provides digital humanities practitioners with guidelines for model selection and quality assessment in historical corpus digitization.

1 Introduction

The evolution of large language models (LLMs) into powerful optical character recognition (OCR) tools has created new opportunities in digital humanities, especially for processing historical documents where traditional OCR systems struggle with non-standard typography, evolving orthographic conventions. However, evaluating LLM-based OCR requires fundamentally different methodological approaches than those developed for traditional machine learning systems. Unlike conventional OCR models, where researchers can control training data, modify architectures, and perform fine-tuning, LLMs present unique evaluation challenges: we cannot access their training data, modify their parameters, or retrain them for specific historical corpora. This constraint necessitates new evaluation frameworks that assess and optimize LLM performance through external factors such as prompt engineering, processing modes, and systematic bias detection.

Current OCR evaluation practices prove inadequate for LLM-based historical document processing. Standard metrics like Character Error Rate (CER) and Word Error Rate (WER) fail to capture LLM-specific behaviors such as temporal conflation, where models incorrectly apply orthographic features from different historical periods, or systematic insertion of anachronistic elements. Moreover, the risk of training data contamination, where evaluation texts may have been included in LLM pretraining corpora, undermines traditional benchmarking approaches that assume clean train-test separation.

We address these methodological gaps through a comprehensive evaluation framework, demonstrated via the challenging case study of 18th-century Russian texts printed in Civil font. This domain exemplifies evaluation challenges facing digital humanities: texts feature distinctive orthographic elements (і, ѣ, ъ at word endings), archaic grammatical forms, and syntactic structures unfamiliar to modern readers, while being underrepresented in digital corpora and thus effectively low-resource for LLMs. These linguistic elements are seldom preserved online; even the Russian National Corpus often presents 18th-century texts in post-1918 orthography (Savchuk, 2009). The diplomatic transcription requirement—preserving exact textual features including line breaks, hyphens, and original typographical errors—further demands precise character-level fidelity that tests LLM capabilities beyond normalized text processing.

Building on recent evidence that LLMs can outperform specialized OCR systems through holistic page processing and prompt engineering (Humphries et al., 2024; Kim et al., 2025; Sohail et al., 2024), our framework introduces key innovations: (1)

Proceedings of the First Workshop NLP and Language Models for Digital Humanities associated with RANLP 2025, pages 75–85, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-106-6-007

contamination-aware dataset creation protocols ensuring evaluation integrity, (2) novel metrics designed to capture LLM behaviors in historical contexts, including Historical Character Preservation Rate (HCPR) and Archaic Insertion Rate (AIR), (3) systematic analysis of processing modes and prompt engineering strategies, (4) comprehensive stability testing accounting for LLM output variability, and (5) feature sensitivity analysis identifying document characteristics that affect performance.

Using this framework, we evaluate 12 leading commercial and open-source multimodal LLMs on a novel dataset of 1,029 pages from 428 unique 18th-century Russian books, revealing systematic patterns in LLM behavior previously undocumented in historical OCR literature. Our analysis uncovers ”over-historicization”—a phenomenon where LLMs systematically insert archaic characters eliminated from the target historical period—demonstrating how LLMs exhibit unexpected temporal biases that standard evaluation approaches cannot detect.

2 Literature review

Early research found LLM-based OCR often outperforms state-of-the-art pipelines. Multimodal LLMs often transcribe unseen manuscripts zero-shot for printed and even handwritten documents in English, Finnish, Italian and Japanese. For instance, Humphries et al. (2024) report GPT-4–class models achieved CER around 5–7% on 18th–19th century English manuscripts—a 14% relative improvement over Transkribus—and further reduced CER to 1.8% with LLM-based post-correction (Humphries et al., 2024). Similarly, Kim et al. (2025) found general-purpose LLMs outperforming tools like Tesseract and TrOCR on historical tables, and early benchmarks highlight the importance of prompt design (e.g., two-shot prompting, line-by-line input) (Kim et al., 2025).

However, recent studies underscore limitations. Crosilla et al. (2025) benchmarked LLMs against Transkribus on multilingual historical datasets and found no consistent overall winner (Crosilla et al., 2025). Proprietary models excelled in English, while open-source LLMs and non-English scripts showed weaker performance, reflecting pretraining data biases. Unpredictable generative outputs and hallucinations remain challenges (Thomas et al., 2024; Boros et al., 2024). While instruction-tuning can aid post-OCR correction, zero-shot self-correction abilities are still limited.

In summary, while LLMs have advanced OCR for some languages, new risks arise: contamination from training data, unpredictable outputs, and the need for task-specific prompt engineering. Notably, little work has evaluated LLMs on Russian historical texts, motivating our focus.

3 Methodology and Data Integrity Controls

Preventing Training Data Contamination. Evaluation of LLMs on OCR tasks is complicated by the risk of test set contamination, as standard benchmarks are often present in LLM pretraining corpora. Prior studies have used n-gram overlap, membership inference attacks (MIAs), and surprisal-based probes, but these methods are limited, especially for historical material (Chang et al., 2023; Ravichander et al., 2025). To ensure robust evaluation, we created a novel dataset of 18th-century Russian texts, digitized from sources never previously recognized or published, and kept strictly offline during all known LLM pretraining periods.

Dataset. Our corpus consists of 1,029 scanned pages from 428 unique 18th-century books printed in Russian Civil font, sourced from the National Library of Russia’s limited-access collection ”Русская книга гражданской печати XVIII в. в библиотеках РФ” (Russian Civil Print Books of the 18th Century in Russian Federation Libraries). We stratified the data by publication period (1750–1800), text density, decorative elements, and subject (fiction, science, religion, etc.) to ensure diversity (see Appendix B for details). Images below 150ppi were excluded, following evidence of poor LLM OCR performance at low resolution (Inoue, 2025).

The ground truth (GT) for this corpus was prepared through a multi-stage process:

Layout Analysis: a YOLOv8 model (Varghese and M., 2024), fine-tuned on a 495-page subset of this corpus, performed region detection; line detection within regions utilized a pre-trained Riksarkivet model.

Initial OCR: A TrOCR model, also fine-tuned on the same 495-page subset (13,456 lines), generated initial transcriptions for the entire corpus. On a held-out portion of the tool-training data, the TrOCR model achieved a CER of 1.83%, WER of 7.82%, and line Exact Match Rate (EMR) of 99.84%.

Manual Correction: 100% manual review using the eScriptorium interface. Our transcription adheres to diplomatic principles, preserving period-

OCR System	CER (%)	WER (%)
Surya (BT5)	45.96	78.33
Tesseract OCR 4.0	21.55	126.10
Transkribus PyLaia	26.93	29.07
Fine-tuned TrOCR*	1.83	7.82

Table 1: OCR results for Old Russian orthography
* Fine-tuned on our dataset; represents an upper bound, not indicative of typical generalization.

specific orthography, hyphenation, original errors, typos, and spacing conventions to accurately reflect the source documents. The resulting GT for the entire 1,029-page corpus (which serves as the evaluation set for the LLMs) comprises 28,657 lines and 146,690 words. It will be released upon publication.

GT was produced by a single expert annotator using a two-pass protocol. We audited a stratified sample of 500 lines with a second verifier under the same guidelines; line-level exact-match was 98.6%, and character-level accuracy was 99.93%.

Baseline: Traditional OCR for Historical Texts. Traditional OCR systems struggle with 18th-century Russian texts in Civil font due to a confluence of challenges, including visually confusable character pairs (e.g., і/ї, т/ш), divergent historical orthographic conventions, typographic inconsistencies from printer-specific variations, and complex page layouts with decorative elements.

To quantify these difficulties, we tested both a general-purpose OCR system (Tesseract), the multilingual BT5 model from the Surya OCR framework (Paruchuri and Team, 2025), and a specialized "Russian print XVIII cent PyLaia" model trained on similar material via Transkribus (reporting 2.40% CER on its own data). On a 100-page sample from our dataset, as shown in Table 1, Surya struggled with the Old Russian orthography (45.96% CER), PyLaia showed a markedly poorer CER of 26.93%, and Tesseract performed significantly worse. For reference, a TrOCR model fine-tuned on our data achieved 1.83% CER—an upper bound, not typical of generic OCR models.

This significant performance drop, even for specialized models not fine-tuned on our specific corpus, underscores the generalization limits of traditional OCR and the impracticality of achieving usable results without extensive, resource-intensive retraining for specific collections. Such limitations motivate our investigation into Large Language Models (LLMs) as a more adaptable alternative.

Our study addresses three research questions: **RQ1:** How do input parameters (processing mode, text density, decorative elements, subject) affect LLM OCR performance for 18th-century Russian Civil font? **RQ2:** What is the impact of prompt engineering on period-specific orthographic fidelity? **RQ3:** What are the characteristic error patterns of LLM-based OCR on historical Russian?

4 Experiment Setup

We evaluated 12 leading LLMs (see Appendix A), including commercial models (Claude, GPT, Gemini) and open-source models (Llama, Qwen). Models were accessed either via their official APIs or, for open-source models, through the TogetherAI service.

Model Stability Evaluation Protocol. To assess performance consistency, we re-evaluated a subset of our models on a fixed sample of 20 documents daily for seven consecutive days. Stability was measured by the Coefficient of Variation (CV) of daily Word Accuracy scores.

Recognition Modes. *Single Line Processing*: Each text line is processed independently, mirroring traditional OCR. This mode provides minimal context and is efficient but may miss cross-line dependencies.

Full Page Processing: The entire page image is provided as a single input, maximizing contextual information. While this may resolve ambiguities, it risks hallucinations or detail loss on dense or complex layouts.

Sliding Window Processing: Fixed-size windows (e.g., 3 lines at a time, transcribing the center) provide more context than single-line but may be more robust to local errors than full-page mode.

Prompt Engineering Experiments. We conducted systematic prompt variation experiments (see Appendix C), including 1) a baseline prompt with basic image information ("Extract the OCR text from this 18th-century Russian book line. Preserve the original Old Russian orthography."), 2) context-enhanced prompts in English (including book information and character list), 3) context-enhanced prompts in Russian.

Evaluation Metrics. We employed an evaluation framework with multiple metrics to assess OCR accuracy, historical fidelity, and case sensitivity:

Standard OCR Metrics. Character Error Rate (CER) and Word Error Rate (WER), using Levenshtein distance between prediction and ground truth.

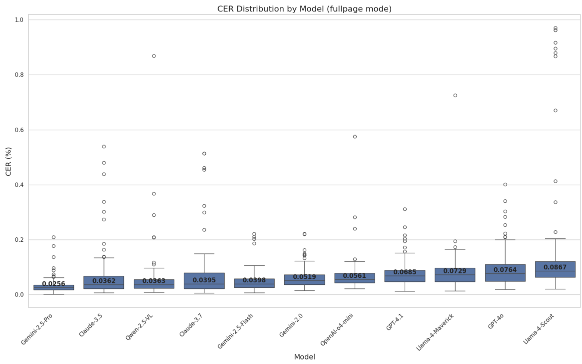


Figure 1: CER distribution by models (full page mode)

Case-Insensitive Metrics. CER and WER after lowercasing, to isolate character recognition from case errors.

Historical Fidelity Metrics. Historical Character Preservation Rate (HCPR) for period-specific characters (i/ï, ѣ, ъ); Archaic Insertion Rate (AIR) for insertion of obsolete, pre-Petrine characters.

Case Preservation Accuracy. Case Error Rate (CaseER) to specifically assess case assignment errors, with particular focus on visually distinctive characters such as ѣ.

5 Experiments and Results

We evaluated all models across the three recognition modes using standard metrics (CER, WER, CI-CER, CI-WER, historical character fidelity, and case accuracy). Table 2 summarizes model performance for each mode; lower values indicate better performance.

Recognition Mode Effectiveness. Table 3 reports performance in full-page mode (the best mode for most models). For each metric, we provide mean values and the observed range (in parentheses) across all documents. Gemini-2.5-Pro achieved the lowest error rates overall. All models showed higher error rates for historical character preservation than for general character recognition, indicating persistent difficulty with period-specific features.

Stability Testing. We assessed performance consistency by processing 20 documents daily with each model for seven consecutive days. Table 4 ranks models by the coefficient of variation (CV) of daily word accuracy; lower CV indicates greater stability. Gemini-2.5-Pro showed both the highest stability and the highest mean word accuracy, while Claude-3.5 exhibited the highest variability. The distribution of CERs by model (Figure 1) further illustrates these differences in stability, with boxplots indicating both

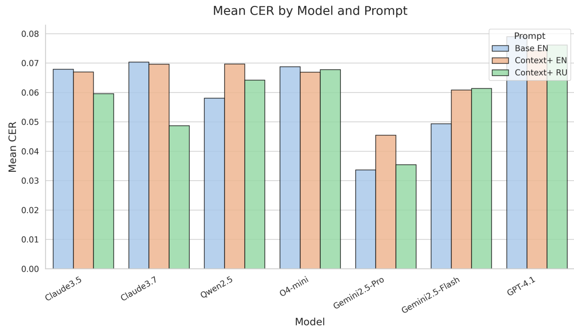


Figure 2: Mean character error rate (CER) by model and prompt strategy (simple English, context-enhanced English, context-enhanced Russian). Lower values indicate better performance.

the central tendency and the frequency of outlier cases for each model.

No model's daily performance deviated by more than one standard deviation from the previous day, suggesting overall day-to-day consistency.

Prompt Engineering Impact. For the top-performing models in full-page mode, we tested three prompting strategies: a simple English prompt, a context-enhanced English prompt, and a context-enhanced Russian prompt. Context-enhanced Russian prompts led to statistically significant CER and WER reductions for several models (e.g., Claude-3.7, Claude-3.5, Gemini-2.5-Flash), with mean CER reductions of up to 0.02 and WER reductions of up to 0.03 ($p < 0.05$, paired t-test). Models such as GPT-4.1 and o4-mini were less affected by prompt type, suggesting greater robustness. In some cases, context-enhanced English prompts increased error rates, underscoring the impact of prompt language and structure on performance. Figure 2 illustrates the mean CER for each model and prompt type, demonstrating the relative gains (or lack thereof) from prompt engineering across systems.

Parameter Impact Visualization. We quantified each model's sensitivity to 17 document and image features by computing the absolute correlation of each feature with CER and WER (see Figure 3). The most robust models (Gemini 2.5-Pro, Gemini 2.5-Flash, Qwen 2.5) achieved the lowest overall error rates and demonstrated the lowest sensitivity to layout complexity and line count—features that most strongly predict increased error for weaker models (e.g., Claude 3.5, Llama4-Mav). Notably, while most models were only moderately affected by old-character content, layout complexity (r up to 0.39) and line count (r up to 0.55) sharply increased

Model	Full page		Single Line		Sliding Window	
	CER (%)	WER (%)	CER (%)	WER (%)	CER (%)	WER (%)
Gemini-2.5-Pro	**3.36**	**4.69**	9.35	15.99	7.83	11.77
Gemini-2.5-Flash	4.94	6.70	18.79	26.21	25.63	30.77
Qwen-2.5-VL	5.81	7.48	7.70	11.29	8.87	12.72
Gemini-2.0	6.14	10.33	10.04	16.43	14.90	19.50
Claude-3.5	6.79	8.46	**5.73**	9.61	**7.17**	11.07
OpenAI-o4-mini	6.87	9.07	9.35	13.89	8.17	11.67
Claude-3.7	7.32	9.47	**5.63**	**9.13**	7.35	**10.03**
GPT-4.1	7.90	9.76	7.55	11.89	9.59	13.35
Llama-4-Maverick	8.29	11.87	8.98	16.62	11.57	16.81
GPT-4o	9.23	13.66	23.75	28.30	11.93	17.13
Llama-4-Scout	15.94	20.51	8.98	15.41	14.95	20.78

Table 2: Model Performance Comparison: Character Error Rate (CER) and Word Error Rate (WER) across three recognition modes for each model. Best scores in each column are bolded.

Model	CER (%)	WER (%)	CI-CER (%)	Hist. Char. Error (%)
Gemini-2.5-Pro	3.36 (0.14–20.95)	4.69 (0.08–31.43)	3.19	9.83
Gemini-2.5-Flash	4.94 (0.75–22.11)	6.70 (0.41–22.82)	4.81	12.86
Qwen-2.5-VL	5.81 (0.81–86.86)	7.48 (0.99–90.14)	5.54	16.40
Gemini-2.0	6.14 (1.51–22.16)	10.33 (1.58–30.55)	5.66	32.00
Claude-3.5	6.79 (0.70–53.96)	8.46 (0.00–51.09)	5.75	15.24
OpenAI-o4-mini	6.87 (2.18–57.54)	9.07 (2.37–58.85)	6.76	18.38
Claude-3.7	7.32 (0.61–51.40)	9.47 (0.21–53.93)	6.21	15.29
GPT-4.1	7.90 (1.20–31.14)	9.76 (1.96–31.85)	7.80	16.94
Llama-4-Maverick	8.29 (1.34–72.57)	11.87 (1.68–69.81)	7.77	22.33
GPT-4o	9.23 (1.89–40.08)	13.66 (1.30–48.87)	9.07	20.70
Llama-4-Scout	15.94 (2.09–97.00)	20.51 (1.70–99.18)	14.95	42.23

Table 3: Full page mode results. For CER and WER, ranges in parentheses show minimum and maximum values across all documents.

error rates for several models. Regression analysis confirmed that text features explain the majority of variance in error (R^2 up to 0.83), with image features only adding modest predictive power.

Document features such as line count and layout complexity are the most predictive of model errors, and only the top-performing models demonstrate resilience to these challenges.

Error analysis. A striking and unexpected finding is that LLMs consistently "over-historicize" 18th-century Russian texts by inserting archaic Slavonic characters that had already been eliminated by Peter the Great's reforms. Instead of modernizing texts (the expected error direction), models frequently introduced obsolete characters, suggesting a systematic bias.

Table 5 summarizes both the top archaic character insertions and the most frequent error types for each model. While OpenAI and Gemini models are prone to introducing pre-Petrine archaic letters, all models struggle most with the preservation of 'ї' and accurate handling of the hard sign 'ъ'.

Over-historicization appears most prominently in OpenAI models, with GPT-4o inserting archaic characters in 59% of files. These insertions are not random, but follow recognizable patterns:

Medieval Slavonic characters: 'ѧ' (little yus), 'ѡ' (omega), 'ȣ' (monograph uk), and 'ѿ' (ot) were standard in medieval manuscripts but had been eliminated from Civil font by the mid-18th century.

Context-sensitive insertions: Models insert archaic characters in predictable linguistic contexts—'ѧ' typically replaces 'я' in reflexive verb endings and after palatalized consonants, 'ѡ' appears in prepositions and prefixes, and 'ȣ' substitutes for 'у' in specific word positions.

Over-Complication with Diacritics. Models often insert diacritical marks and combining characters

Rank	Model	CV	Mean Word Accuracy	StdDev
1	Gemini-2.5-Pro	0.037	0.9620	0.036
2	Gemini-2.0-Flash-Lite	0.051	0.9300	0.048
3	Gemini-2.5-Flash	0.081	0.9430	0.077
4	GPT-4.1	0.113	0.9044	0.102
5	o4-mini	0.118	0.9070	0.107
6	GPT-4o	0.227	0.8620	0.195
7	Claude-3.7	0.271	0.8486	0.230
8	Claude-3.5	0.307	0.8340	0.256

Table 4: Models ranked by output stability over seven days (lower CV = higher stability).

Model	Top Archaic Insertions	Most Common Errors
GPT-4o	ѡ, ѧ, ꙋ, ѿ, ꙗ	ї → і, ъ → ь, т → ш
GPT-4.1	ѡ, ѧ, ꙋ, ꙗ	ї → і, ъ → ь, т → ш
o4-mini	ѧ, ѭ, ѥ	ї → і, ъ → ь, ѣ → е
Gemini-2.5-Flash	ѧ, ѥ	ї → і, т → ш, ъ → ь
Claude-3.7	Minimal archaic insertions	ъ → Ъ, ї → і, ь → ъ
Qwen2.5	Minimal archaic insertions	ї → і, т → п, ъ → ь

Table 5: Archaic character insertions and most common OCR errors by model.

that are not present in 18th-century Civil font, further complicating the transcription and introducing anachronistic features. This tendency may be exacerbated by visual noise and typographic ambiguity in the source material. For example, faded ink, paper discoloration, or ink bleed-through can produce artifacts that models misinterpret as diacritics or additional marks. Similarly, nonstandard or worn-out typefaces might blur the distinction between basic characters and diacritical elements, especially for visually similar Cyrillic forms.

Character Preservation and Confusion. Distinct error patterns are evident in the handling of period-specific characters:

'ї' vs. 'і': Although 'ї' is legitimate in 18th-century Civil font, models frequently replace it with 'і': the most common substitution error across all models.

'ѣ' *(yat) preservation*: Rates vary widely, from 77.30% (Claude-3.5) to 89.03% (Gemini-2.0).

Hard/soft sign confusion: All models have trouble with the terminal hard sign 'ъ'—commonly omitted (Claude), replaced with 'ь' (Gemini), or incorrectly capitalized 'ъ→Ъ' (Claude-3.5/3.7).

Visual similarity errors: Certain character pairs are frequently confused due to visual similarity—'т→ш' (Gemini), 'т→п' (Qwen, o4-mini). This confusion is exacerbated not only by scan degradation or low resolution, but also by the nature of 18th-century typography. Figure 4 illustrates how the Civil font renders "т" and "ш" in ways that may appear nearly identical.

These systematic error patterns offer key insights into LLM behavior on historical text: **1) Temporal conflation**: Models conflate orthographic features from different periods (pre-Petrine Church Slavonic, 18th-century Civil font, modern Russian), struggling to maintain strict period boundaries; **2) Contextual over-fitting**: There is a correlation between text subject and error types; models seem to apply different orthographic standards by genre, likely reflecting biases in their training data; **3) Model family signatures**: Error profiles differ by provider (OpenAI, Anthropic, Google), suggesting differences in training data and strategies regarding historical texts.

6 Discussion

LLM Behavior: Over-Historicization and Error Patterns. A surprising result is that LLMs systematically "over-historicize" 18th-century Russian texts, introducing archaic Slavonic characters that had been eliminated by the era in question. Rather than modernizing spelling, models often default to pre-reform or even medieval forms.

This likely reflects how LLMs, lacking explicit period awareness, generalize from a noisy mixture of training data: rare or visually distinctive archaic forms become signals for "historical text" regardless of actual period accuracy. Multimodal and text-only corpora contain heterogeneous historical Russian

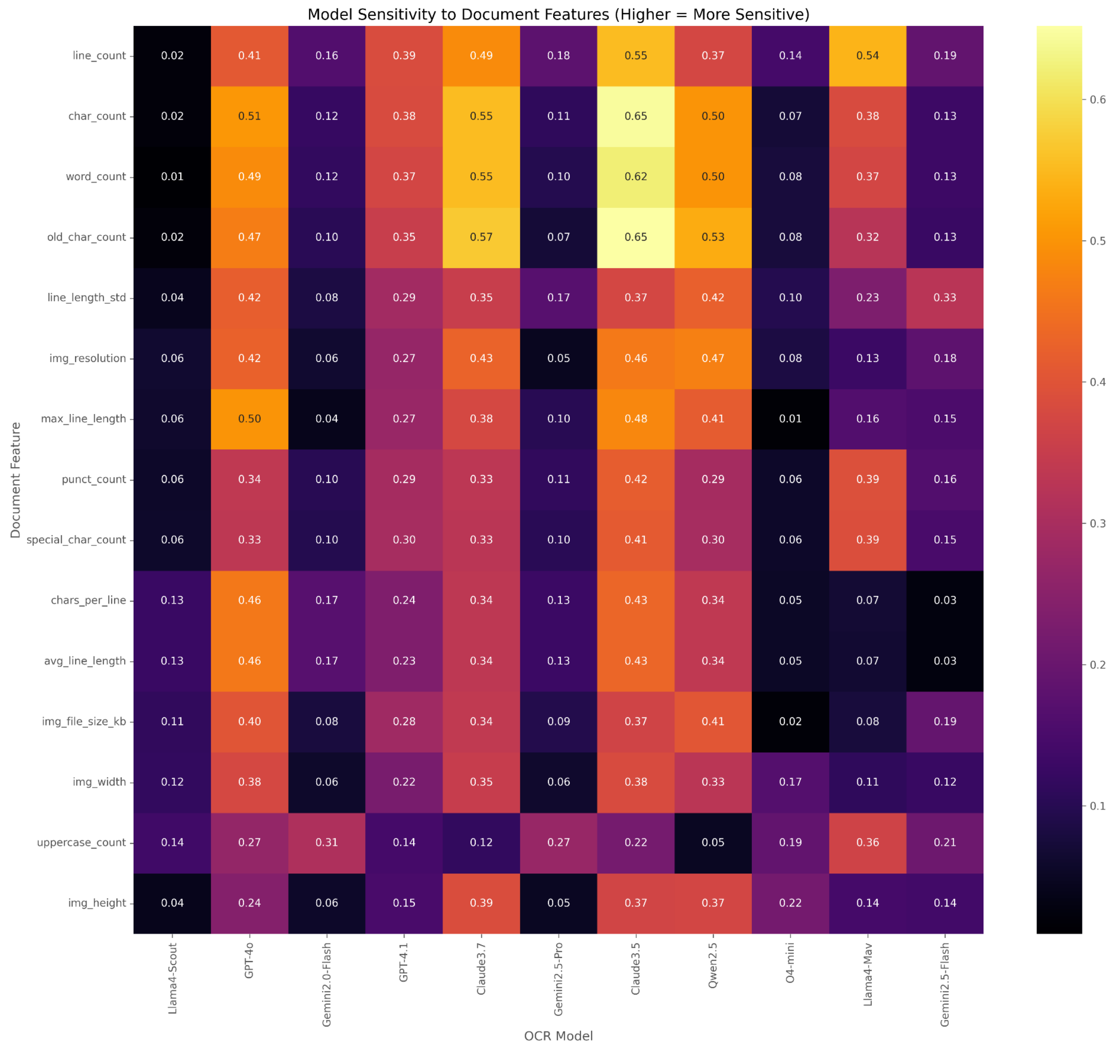


Figure 3: Model sensitivity to document features. Each cell shows the absolute correlation between a given feature (rows) and OCR error rates (CER/WER, averaged) for each model (columns; names shortened for readability). Higher values indicate greater sensitivity—that is, a model's performance degrades more as that document feature increases. The most robust models (e.g., Gemini-2.5-Pro, o4-mini) exhibit consistently low sensitivity, while others (e.g., Claude3.5, Llama4-Mav) show heightened sensitivity to line count, old-character content, and layout complexity.

(Church Slavonic, pre-reform, post-1918), but models lack explicit period tags, so "historic" cues (yus letters, omega, diacritics) become generic signals for "old text". Visual ambiguity in the typography and degraded print quality may further reinforce these mistakes, with models erring on the side of complexity and inserting diacritics or combining marks absent from authentic 18th-century Civil font.

Optimal Model and Prompting Strategies for Historical OCR. Our results indicate that Gemini and Qwen models are the most robust and accurate models across diverse document types, especially where high line counts or layout complexity would otherwise increase error rates. Prompt engineering can enhance performance, particularly when prompts specify period features or are given in Russian, but the best models are less dependent on prompt tweaks. Full-page mode generally yields the best accuracy, but for models highly sensitive to document length, line-by-line mode can be preferable.

Post-OCR Correction Analysis. Our experiments reveal counterintuitive findings about LLM post-correction effectiveness: when providing both image and OCR text to higher-performing models, performance does not exceed the correcting model's direct OCR capabilities—models essentially re-perform OCR rather than correct the provided text. We suggest that models rarely apply constrained edits; instead they re-decode from the image, using the text as weak context. Text-only correction (without

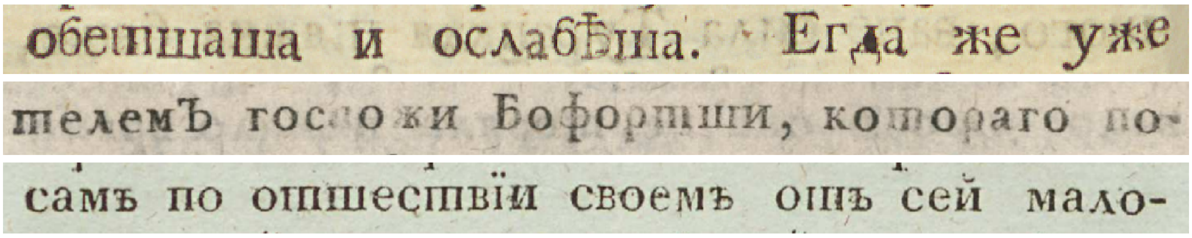


Figure 4: Excerpt from an 18th-century Russian book printed in Civil font. The letters "ш" (as in обетшаша, ослабѣша, Бофортши, отшествїи) display notable typographic variability, occasionally resembling the "т" glyph. Such variability, inherent to period printing, contributes to frequent "т→ш" substitution errors.

source images) consistently degrades performance, with models introducing errors that corrupt the original transcription. Two mechanisms likely apply: (i) attention dilution/position effects—LLMs are known to unevenly use long contexts; (ii) editor vs. generator mismatch—chat-tuned models prioritize fluent regeneration over minimal edits unless decoding is constrained. These findings suggest practitioners should focus on selecting optimal models for direct OCR rather than post-correction pipelines, as correction attempts either provide no benefit or actively harm accuracy.

7 Conclusion

This paper introduces a comprehensive methodological framework for evaluating large language models (LLMs) on historical OCR tasks, exemplified by the case study of 18th-century Russian prints in Civil font. Our results demonstrate that LLM-based approaches substantially outperform traditional OCR systems for these challenging materials, and our work sets out best practices for reliable evaluation and practical implementation. Our stratified coverage across printers, decades, genres, and layouts supports transfer to other historical prints with period-specific orthography; applying the same protocol with a collection-specific grapheme inventory for HCPR/AIR typically requires only a brief 10–20-page pilot.

Looking ahead, we note that LLMs are rapidly improving; having a clearly defined evaluation protocol, public metrics, and detailed error analysis will allow ongoing, transparent tracking of model progress. However, the publication of ground-truth datasets for evaluation is a double-edged sword: once released, they risk being incorporated into future model training, compromising their utility for truly unseen evaluation. Even a single benchmarking release may affect evaluation integrity if outputs are shared or scraped. The trade-off between transparency, reproducibility, and long-term benchmark validity remains an open question for the community.

8 Limitations

Our dataset is specific to Russian Civil font print from the second half of the 18th century, and our manual ground truth verification process, while rigorous, may still be subject to rare annotation errors, especially for visually ambiguous or degraded source material. All evaluated LLMs were accessed via their respective APIs; however, we excluded OpenAI o3 due to prohibitive usage costs. For consistency, we requested structured (JSON) outputs when supported (e.g., OpenAI models) and programmatically extracted lines from unstructured outputs otherwise. Alternative output formats, such as Markdown or raw text, may yield different recognition results and could be further investigated in future work. Additionally, our model stability experiments revealed that LLM outputs can vary between runs for the same document and model, though this variance was relatively minor within our observation window. Nonetheless, this inherent non-determinism may affect reproducibility and should be considered when interpreting comparative results.

During evaluation, if a model's response consisted of a clear API error message (e.g., "Unable to process image" or an explicit failure code), we resubmitted the OCR request to ensure that temporary API or service issues did not affect the results. However, if a model returned a plausible but off-target output (such as an explanation, commentary, or unrelated generative text instead of a transcription), we recorded this as the model's result without resubmission, in line with our goal of measuring real-world output quality rather than optimizing for best-case scenarios.

References

Emanuela Boros, Maud Ehrmann, Matteo Romanello, Sven Najem-Meyer, and Frédéric Kaplan. 2024. Post-correction of historical text transcripts with large language models: An exploratory study. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 133–159, St. Julians, Malta. Association for Computational Linguistics.

Kent Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023. Speak, memory: An archaeology of books known to ChatGPT/GPT-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7312–7327, Singapore. Association for Computational Linguistics.

Giorgia Crosilla, Lukas Klic, and Giovanni Colavizza. 2025. Benchmarking large language models for handwritten text recognition.

Mark Humphries, Lianne C. Leddy, Quinn Downton, Meredith Legace, John McConnell, Isabella Murray, and Elizabeth Spence. 2024. Unlocking the archives: Using large language models to transcribe handwritten historical documents.

Kotaro Inoue. 2025. Context-independent ocr with multimodal llms: Effects of image resolution and visual complexity. *arXiv preprint arXiv:2503.23667*.

Seorin Kim, Julien Baudru, Wouter Ryckbosch, Hugues Bersini, and Vincent Ginis. 2025. Early evidence of how llms outperform traditional systems on ocr/htr tasks for historical records.

Vikas Paruchuri and Datalab Team. 2025. Surya: A lightweight document ocr and analysis toolkit. `https://github.com/VikParuchuri/surya`. GitHub repository.

Abhilasha Ravichander, Jillian Fisher, Taylor Sorensen, Ximing Lu, Maria Antoniak, Bill Yuchen Lin, Niloofar Mireshghallah, Chandra Bhagavatula, and Yejin Choi. 2025. Information-guided identification of training data imprint in (proprietary) 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 1962–1978, Albuquerque, New Mexico. Association for Computational Linguistics.

Svetlana Savchuk. 2009. The corpus of eighteenth-century texts: Towards the construction of the diachronic corpus of russian. In *Proceedings of the Corpus Linguistics Conference 2009*.

Muhammad Abdullah Sohail, Salaar Masood, and Hamza Iqbal. 2024. Deciphering the underserved: Benchmarking llm ocr for low-resource scripts.

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

Rejin Varghese and Sambath M. 2024. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, pages 1–6.

A Evaluated Models

Provider	Model Name
Anthropic	Claude 3.7 Sonnet
Anthropic	Claude 3.5 Sonnet 20241022
OpenAI	GPT-4o-2024-08-06
OpenAI	GPT-4.1-2025-04-14
OpenAI	o4-mini-2025-04-16
Google	Gemini 2.0 Flash
Google	Gemini 2.5 Pro (05-06)
Google	Gemini 2.5 Flash (04-17)
Google	Gemini 2.0 Flash-Lite
Qwen AI	Qwen2.5-VL-72B-Instruct
Meta	Llama-4 Maverick 17B 128E Instruct FP8
Meta	Llama-4 Scout 17B 16E Instruct

B Dataset Description

The evaluation dataset comprises 1,029 page images sampled from 428 unique Russian books published between 1752 and 1801, with the majority printed in the 1780s and 1790s. The collection covers a broad range of genres, with the largest shares contributed by fiction (22.7%), religion (15.7%), history (15.0%), and science (12.9%). This diversity helps ensure that both typographical and linguistic variation in Russian print is well-represented for OCR evaluation. All texts are printed in the Civil font, introduced by Peter the Great's typographic reform.

The resulting corpus contains 28,657 lines and 146,690 words. The 100-page sample from the corpus is published online alongside the LLM-based OCR results (github repository contains ground truth transcriptions and model outputs). The year and subject distributions are shown in Figures 5 and 6. The dataset is dominated by fiction, religion, history, and science, but maintains coverage across a variety of genres, supporting generalizable evaluation of historical OCR models

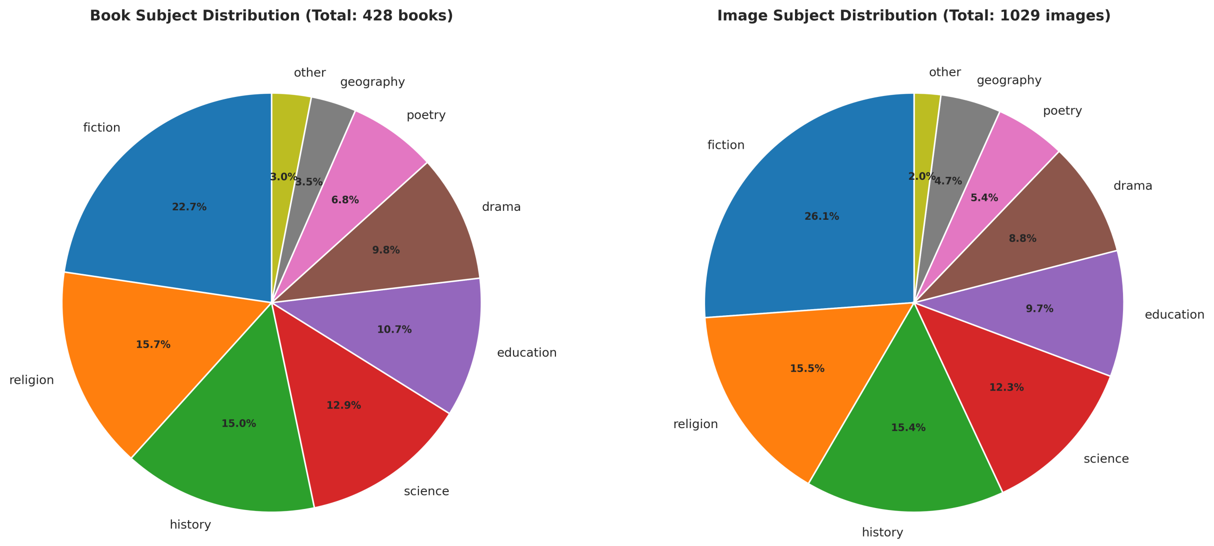


Figure 5: Subject distribution in the evaluation dataset. **Left:** Distribution by unique books (N=428). **Right:** Distribution by sampled page images (N=1029). The dataset is dominated by fiction, religion, history, and science, but maintains coverage across a variety of genres, supporting generalizable evaluation of historical OCR models.

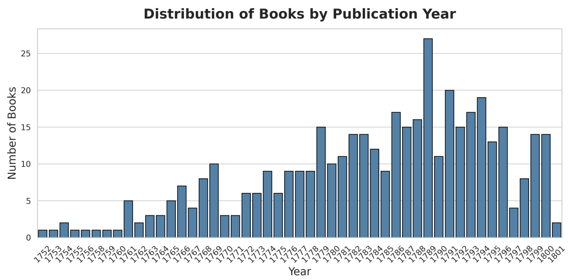


Figure 6: Year distribution in the evaluation dataset. The corpus reflects the rapid growth of Russian print in the late 18th century.

C Prompts

Single Line Mode Prompt. Extract the OCR text from this 18th-century Russian book line. Preserve the original Old Russian orthography. Respond with ONLY a JSON object containing the extracted text in the `'line'` field.

Sliding Window Mode Prompt. Extract the text from these consecutive lines of an 18th-century Russian book. Focus on the middle line while using surrounding lines as context. Preserve the original Old Russian orthography. Respond with ONLY a JSON object containing the extracted text of the middle line in the `'line'` field. Do not include any additional explanations.

Full Page Mode Prompt. Extract the OCR text from this full page of an 18th-century Russian book. Preserve the original Old Russian orthography. Process each line independently. Respond with ONLY a JSON array where each object has a `'line'` field containing the transcribed text. Do not include any additional explanations.

Full Page Context-Enhanced Prompt (English). You are an expert OCR system specialized in processing 18th-century Russian texts. Your task is to accurately transcribe text from an image of a page from a {book_year} Russian book titled "{book_title}" published in {publication_info}.

Instructions:

Analyze the entire image thoroughly before beginning transcription.

Process the text line by line, maintaining the exact layout of the original page.

Preserve all original Old Russian orthography, including:

– special characters: ѣ, Ѳ, ѳ, ѵ, і, ї, ъ

– Original punctuation

– Capitalization as it appears in the original text.

Respond with ONLY a JSON array where each object has a `'line'` field containing the extracted text. Do not include any explanations or additional formatting in your response.

Full Page Context-Enhanced Prompt (Russian). Вы являетесь экспертной OCR-системой, специализирующейся на обработке русских текстов XVIII века, напечатанных гражданским шрифтом после реформы Петра I (1708–1710 гг.), но до реформы орфографии 1918 года. Ваша задача — точно транскрибировать текст с изображения страницы из русской книги {book_year} года под названием “{book_title}”, опубликованной в {publication_info}. Особенности орфографии этого периода включают:

Наличие специфических букв: ѣ (ять), і (и десятеричное) или ї, ѳ (фита), ѵ (ижица), ъ (твёрдый знак на конце слов)

Отсутствие букв церковнославянского алфавита (ѡ, ѧ, ѫ, ѯ, ѱ, etc.)

Использование гражданского шрифта вместо устава или полуустава

Инструкции:

Тщательно проанализируйте всё изображение перед началом транскрипции.

Обрабатывайте текст построчно, сохраняя точное расположение оригинальной страницы.

Сохраняйте всю оригинальную старорусскую орфографию, включая:

– специальные символы: ѣ, Ѳ, ѳ, ѵ, і, ї и ъ,

– оригинальную пунктуацию,

– заглавные буквы так, как они представлены в оригинальном тексте.

Отвечайте ТОЛЬКО JSON-массивом, где каждый объект имеет поле `’line’`, содержащее каждую извлеченную строку текста. Не включайте никаких пояснений или дополнительного форматирования в ваш ответ.

Finding the Plea: Evaluating the Ability of LLMs to Identify Rhetorical Structure in Swedish and English Historical Petitions

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

Abstract

Large language models (LLMs) have shown impressive capabilities across many NLP tasks, but their effectiveness on fine-grained content annotation, especially for historical texts, remains underexplored. This study investigates how well GPT-4, Gemini, Mixtral, Mistral, and LLaMA can identify rhetorical sections (*Salutatio*, *Petitio*, and *Conclusio*) in 100 English and 100 Swedish petitions using few-shot prompting with varying levels of detail. Most models perform very well, achieving F1 scores in the high 90s for *Salutatio*, though *Petitio* and *Conclusio* prove more challenging, particularly for smaller models and Swedish data. Cross-lingual prompting yields mixed results, and models generally underestimate document difficulty. These findings demonstrate the strong potential of LLMs for assisting with nuanced historical annotation while highlighting areas for further investigation.

1 Introduction

In recent years, large language models (LLMs) have demonstrated remarkable capabilities across many NLP tasks, including translation, summarisation, and question answering. However, their performance on fine-grained, content-aware text annotation tasks, particularly those involving historical texts and moderately resourced languages, remains a relatively unexamined area.

In this paper, we investigate to what extent LLMs can be used to analyse and annotate a specific type of historical document, the petition. In pre-modern and pre-democratic societies, petitions allowed ordinary people to seek redress or support from those in positions of authority — such as courts, parliaments, landlords, or monarchs (Houston, 2014). Despite their potential to shed light on the everyday lives, concerns, and ways of navigating authority in the past, petitions have been relatively neglected in both historical and computational research.

The interdisciplinary project *Speaking to One's Superiors: Petitions as Cultural Heritage and Sources of Knowledge*, led by Uppsala University's *Gender and Work* (GaW) research project and funded by the Swedish Research Council, investigates 18th-century Swedish petitions.[1] Thousands of documents have been digitised, annotated, and made publicly accessible to shed light on how women and men in Early Modern Sweden made a living and asserted their rights.

Petitions in early modern Europe often followed a classical rhetorical structure, typically divided into five sections: *Salutatio*, *Exordium*, *Narratio* (including *Argumentatio*), *Petitio*, and *Conclusio* (Dodd, 2011; Israelsson, 2016). This study explores the use of LLMs to automatically identify such sections, with the aim of supporting information extraction for historians and other scholars working with petitions. We focus in particular on three key components: the greeting (*Salutatio*), the request (*Petitio*), and the ending (*Conclusio*).

To support this work, we have created a dataset of 100 historical petitions in Swedish and 100 in English, each annotated to mark the locations of the targeted rhetorical sections. Each document is also assigned a difficulty score, reflecting the level of annotation difficulty, allowing us to evaluate the relationship between model confidence and human-perceived difficulty, for a more nuanced assessment of LLM performance on complex historical texts.

With our experiments, we evaluate how effectively LLMs can find and annotate rhetorical components in historical petitions using few-shot prompting. We have four research goals:

1. Comparing model performance across a range of commercial and open-source LLMs.
2. Studying how prompt complexity (less vs. more detailed instructions) and one vs three

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

Proceedings of the First Workshop NLP and Language Models for Digital Humanities associated with RANLP 2025, pages 86–101, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-106-6-008

output examples affect annotation accuracy.

3. Assessing cross-linguistic generalisation by testing both English and Swedish, and varying the prompt language used on Swedish data.

4. Comparing human and model uncertainty by analysing how well the performance of the models and their self-assigned difficulty scores align with human difficulty ratings.

The models under investigation include a diverse mix of architectures and scales: GPT-4 (OpenAI), Gemini 1.5 Pro (Google DeepMind), Mistral 7B and Mixtral 8x22B (Mistral AI), and LLaMA 3 (Meta AI) in both 8B and 70B configurations. We evaluate the model outputs and also analyse how closely model confidence aligns with human difficulty judgments. By combining comparative evaluation, prompt design variation, and confidence modeling, this work aims to illuminate both the capabilities and the limitations of LLMs in performing nuanced annotation tasks on historical texts across diverse settings.

2 Related Work

Applying LLMs to structured tasks like rhetorical analysis depends critically on methods used to guide the model's output, a field broadly known as prompt engineering. Sahoo et al. (2024) emphasise that obtaining accurate and structured information from LLMs is a non-trivial challenge that requires carefully designed interaction strategies. This can be argued to be particularly true for annotation and extraction tasks, where the desired output is not free-form text but a structured representation. Cheng et al. (2024) address this for Named Entity Recognition (NER) by proposing a standardised prompting method. They demonstrate that a combination of a clear task definition, illustrative few-shot examples, and a strict output format specification is essential for improving the reliability of structured data extraction in a few-shot context. In the context of rhetorical analysis, Maekawa et al. (2024) tackle discourse parsing by translating traditional parsing strategies into effective prompts for a decoder-only LLM. By combining this with parameter-efficient fine-tuning (QLoRA), they achieve state-of-the-art results with strong generalisation, demonstrating that LLMs can model complex rhetorical hierarchies.

The potential of using LLMs to process historical data is gaining attention, offering informative new tools for fields from behavioral science to the digital humanities (Varnum et al., 2024). This development has led to the application of LLMs across all phases of historical research. At the most foundational level, researchers are using LLMs to overcome long-standing barriers, such as transcribing handwritten historical documents to unlock previously inaccessible archives (Humphries et al., 2025). Moving beyond data preparation to analysis, Cohen et al. (2025) investigate the potential of BERT and GPT-4o models to detect irony in 19th-century Latin American newspaper texts, demonstrating how LLMs can be used in context-dependent tasks given historical linguistic changes. Overall, this body of research indicates that LLMs could be highly effective for automated rhetorical annotation of historical texts, a task that to the best of our knowledge has not been explored previously.

Test Set	Period	# Docs	# Toks	Avg Toks/Doc
Swedish	1709–1800	100	24,904	249 ± 116
English	1692–1799	100	28,831	288 ± 172

Table 1: Overview of the Swedish and English test sets: time period, document count, total token count, average and standard deviation of tokens per document.

3 The Petition Data Sets

An overview of the test set statistics is presented in Table 1. Below, we describe the composition and annotation process for each dataset in more detail.

3.1 The Swedish Data Set

The Swedish petition data set consists of 100 petitions from the 18th century, transcribed by a historian. These petitions were originally written between 1719 and 1800 and submitted to the regional administration in Örebro, Sweden. We also make use of an additional 10 petitions as a development set, used when developing the code and prompts to our experiments.

3.2 The English Data Set

The English dataset is drawn from the *London Lives 1690–1800* archive[2] (Hitchcock et al., 2012), a large digital collection of legal and social records focusing on everyday Londoners. We use a digitised subset of these materials curated for the London Lives Petitions Project (Howard, 2016).[3] The

[2]https://www.londonlives.org/

[3]https://github.com/sharonhoward/llpp?tab=readme-ov-file

whole digitised collection includes around 10,000 petitioning documents submitted to magistrates and courts, from which we select petitions addressed to the courts of the Old Bailey and Middlesex Sessions and City of London. The petitions were originally transcribed using a double rekeying process, where two (non-academic) typists transcribe text, the two versions are compared and only discrepancies are manually checked. We randomly select 100 petitions from this collection in a stratified manner based on court for our English test set and 10 petitions for a development set.

3.3 Rhetorical Structure of Petitions

In many parts of premodern Europe, the structure of petitions followed a classical rhetorical framework, typically comprising five or six sections (Hansson, 1988; Sokoll, 2006; Israelsson, 2016):

1. *Salutatio*: Formal salutation to the addressee.
2. *Exordium*: Brief opening phrase appealing to the recipient's greatness or capacity to help.
3. *Narratio*: Narration of the circumstances leading to the petition, often mixed with arguments (4. *Argumentatio*).
5. *Petitio*: Specific request or plea being made.
6. *Conclusio*: Final phrase(s) of courtesy and/or inferiority, often including a signature.

In this study, we focus both manual annotation and model evaluation on the sections *Salutatio*, *Petitio*, and *Conclusio*. The *Exordium* is excluded, as it is typically a brief phrase, may be absent from some texts, and its identification is often more subjective. The sections *Narratio* and *Argumentatio* are likewise omitted, as they are frequently intertwined and difficult to distinguish reliably. As a result, these parts remain unannotated, and the majority of unmarked content in the corpus should correspond to one or both of these rhetorical functions. Our experiments thus primarily test the ability of language models to identify the three selected sections.

3.4 Annotated Gold Data Sets

To identify the targeted rhetorical sections in both the Swedish and English datasets, we manually inserted start and end tags for each section. Three annotators carried out the work, with each petition annotated by two of them. The data was divided into batches, and after each round, the specific disagreements were resolved and general principles agreed upon to support consistency in later batches.

Swedish Dataset				
Diff	Section	%Exact	κ	TokDist
1.60	Overall	48.0	0.82	5.76
	Salutatio	100.0	1.00	0.00
	Petitio	48.0	0.48	16.70
	Conclusio	98.0	0.92	0.60
English Dataset				
Diff	Section	%Exact	κ	TokDist
1.49	Overall	68.0	0.88	2.15
	Salutatio	99.0	0.99	0.02
	Petitio	69.0	0.68	5.84
	Conclusio	95.0	0.95	0.59

Table 2: Inter-annotator agreement for Swedish and English datasets. Diff = average difficulty score, %Exact = percent exact matches, κ = Cohen's kappa, TokDist = mean token distance.

During the annotation process, each document was also assigned a difficulty score ranging from 0 to 2. Scoring was based on annotator agreement to reflect the level of annotation difficulty. Documents that received a score of 1 often exhibit mild ambiguities, such as blended rhetorical sections or unusual phrasing, leading to minor disagreements, which were typically resolved quickly. A score of 2 was assigned to cases that required extended discussion to resolve disagreement. These documents often present interpretive challenges due to older/non-standard orthography, incomplete phrases, or heavy use of abbreviations, which complicates clear identification of rhetorical boundaries. In particular, separating *Petitio* from *Narratio/Argumentatio* was frequently experienced as challenging in these cases. Examples of annotation agreements and the provided difficulty scores can be found in the Appendix. By including difficulty annotations, we can assess if and how model confidence aligns with human-perceived complexity thereby enriching the evaluation of LLMs on historically and linguistically complex texts.

Table 2 presents the average difficulty scores and inter-annotator agreement scores for each dataset. The exact match score (%Exact) measures the percentage of petitions or sections where the two annotations are token identical, while the κ score is Cohen's kappa. The token distance measure (TokDist) is the average number of tokens that differ between the two annotations. It is worth noting that *Petitio* emerges as the most challenging petition segment to annotate, as indicated by its lowest percentage of identical tags and Cohen's kappa scores, as well as the highest mean token distance scores. This is particularly evident in the Swedish dataset.

4 Method

With our experiments, we aim to evaluate how well LLMs can annotate rhetorical sections of historical petitions using few-shot prompting. The key components of our method are presented below.

Assess model performance on the annotation task across several leading LLMs, including both commercial and open-source systems.

Investigate the role of prompt design by comparing less vs. more detailed instructions and by providing either one or three output examples to understand how prompt complexity influences tagging accuracy.

Evaluate cross-linguistic generalisation by comparing results on English (a high-resource language) and Swedish (a moderately resourced language in the LLM training ecosystem). To further explore cross-lingual effects, we prompt the Swedish data set using both Swedish and English instructions in the prompts (apart from the given output examples), assessing how the prompt language influences model tagging performance.

Compare human and model uncertainty by assessing how well models can self-estimate uncertainty in comparison to human judgments of difficulty. Each text in the dataset has not only been annotated for rhetorical sections but has also been assigned a difficulty score by the human annotators, on a scale from 0 (easy) to 2 (difficult) (see more details in Section 3.4). To compare with human difficulty judgments, we instruct the language models to return a difficulty score alongside each predicted annotation.

4.1 Prompting Settings and Variations

To test whether and how model performance is affected by prompt design, we developed three prompt variations:

Prompt 1: short 1-shot This prompt includes a less detailed description of the task, a list of the tags to be used, the required output format, and one example output showing an annotated petition.

Prompt 2: long 1-shot Similar to Prompt 1 but with a more detailed and dataset-specific description, providing additional context and clarification about the task.

Prompt 3: long 3-shot Extends Prompt 2 by including three example outputs of annotated petitions, giving the model more extensive demonstrations of the expected tagging and formatting.

These variations were designed to evaluate how the level of detail and the number of examples influence the ability of the models to perform the tagging task accurately. For the Swedish dataset, we also examine whether the language of the prompt influences model performance by testing each prompt in both Swedish and English. Examples of prompts for both datasets can be found in the Appendix.

4.2 Models

We evaluate six contemporary LLMs with varying architectures and sizes: GPT-4 (Achiam et al., 2023) from OpenAI, Gemini Pro 1.5 (Team et al., 2024) from Google DeepMind, two LLaMA open-weight transformers from Meta AI in both 8B and 70B configurations (Touvron et al., 2023) and the open-source models Mixtral 8x22B (Jiang et al., 2024) and Mistral 7B (Jiang et al., 2023) from Mistral AI. All models are accessed via APIs (e.g., OpenAI, Google, Mistral), and we ensure consistent prompt formatting and settings across runs for comparability. To promote deterministic generation and reproducibility, all prompts are run with a temperature setting of 0.

4.3 Evaluation Procedure

To evaluate the performance of the LLMs on the rhetorical annotation task, we make use of a unified evaluation framework applicable across all models and languages. Evaluation is conducted separately on English and Swedish data sets to compare model performance on a high-resource versus a moderately resourced language. For the Swedish data set, we evaluate both the use of Swedish instructions and English instructions. We perform an evaluation per prompt type, to analyse the impact of prompt complexity by comparing shorter versus more detailed instructions.

Annotation The models are instructed to annotate texts using predefined rhetorical tags, as described in Section 3.4. We evaluate model predic-

Results for the English Dataset

Prompt	Model	salutatio P	salutatio R	salutatio F1	salutatio TD	petitio P	petitio R	petitio F1	petitio TD	conclusio P	conclusio R	conclusio F1	conclusio TD
short 1-shot prompt	GPT-4	68.4	**100.0**	81.2	0.46	94.9	97.3	96.1	0.11	97.2	97.6	**97.4**	0.03
	Gemini	95.5	**100.0**	97.7	0.06	94.7	97.2	96.0	0.12	98.2	91.0	94.5	0.06
	Mixtral	98.5	**100.0**	99.2	0.03	91.2	96.7	93.9	0.17	89.3	91.6	90.5	0.05
	Mistral	98.5	99.9	99.2	0.03	74.3	86.0	79.8	0.40	87.7	72.6	79.4	0.23
	LLaMA 70B	99.7	**100.0**	99.8	0.02	95.0	96.7	95.8	0.13	89.3	91.9	90.6	0.09
	LLaMA 8B	98.8	99.9	99.4	0.02	90.6	90.5	90.5	0.18	64.4	21.2	31.8	0.68
long 1-shot prompt	GPT-4	77.9	97.7	86.7	0.36	96.1	97.1	96.6	0.09	91.6	**98.7**	95.0	0.03
	Gemini	98.6	**100.0**	99.3	0.02	96.3	96.1	96.2	0.15	98.1	89.4	93.6	0.04
	Mixtral	98.8	**100.0**	99.4	0.03	89.7	96.5	93.0	0.25	82.2	91.9	86.8	0.07
	Mistral	98.5	99.9	99.2	0.03	68.3	92.3	78.5	0.43	89.4	84.8	87.1	0.11
	LLaMA 70B	99.4	**100.0**	99.7	0.03	96.4	97.1	96.7	0.18	97.1	94.3	95.7	0.06
	LLaMA 8B	98.9	**100.0**	99.5	0.03	88.7	93.1	90.9	0.19	74.3	49.7	59.6	0.48
long 3-shot prompt	GPT-4	89.8	**100.0**	94.6	0.15	95.6	**97.8**	96.7	0.08	89.9	97.5	93.5	0.04
	Gemini	99.6	**100.0**	99.8	0.01	97.2	94.4	95.8	0.16	**100.0**	83.6	91.0	0.06
	Mixtral	98.1	**100.0**	99.0	0.03	90.6	96.2	93.3	0.29	87.0	84.8	85.9	0.06
	Mistral	98.5	99.9	99.2	0.03	59.4	91.1	71.9	0.53	83.9	63.7	72.4	0.21
	LLaMA 70B	**99.8**	**100.0**	**99.9**	0.02	**97.3**	97.5	**97.4**	0.15	81.7	95.1	87.9	0.09
	LLaMA 8B	98.3	99.9	99.1	0.02	86.4	94.5	90.3	0.25	88.2	67.6	76.6	0.25

Table 3: Results for English data across three prompt types and six models. Scores includes token-level precision (P), recall (R), and F1 in percentage, and mean token-level edit distance for each predicted span.

tions against gold annotations using two metrics. First, we compute token-level precision, recall, and F1-score by collecting all tokens that occur inside predicted spans and comparing them to all tokens inside the corresponding gold spans for each rhetorical tag. Second, we calculate the mean token-level edit distance: for each predicted span, we compute the minimum normalised edit distance (Levenshtein distance over whitespace-tokenised words) to any gold span of the same tag. Distances are averaged across all predicted spans, including perfect matches (where distance = 0). A tag that is missing in both the model prediction and in the gold annotation is scored as a perfect match.

Difficulty Estimation To evaluate the alignment between model-assigned and human-assigned difficulty ratings, we calculate the mean error (ME) as the average difference between model-predicted and human-assigned difficulty scores:

$$\text{Error} = \text{Model} - \text{Human} \tag{1}$$

A positive mean error indicates that the model systematically rates documents as more difficult than human annotators, whereas a negative mean error indicates that the model rates documents as easier than humans do. We also use Spearman's rank correlation coefficient (ρ) (Spearman, 1904) to measure how well the order of difficulties assigned by the model agrees with the order assigned by humans.

5 Results and Discussion

The results for the English dataset are presented in Table 3, those for the Swedish dataset with Swedish prompts in Table 4, and for the Swedish dataset with English prompts in Table 5.

5.1 Results on the Annotation Task

Although results vary between specific models, prompts, and datasets, an overall view suggests that the models generally perform the task of annotating the petitions very well, with many F1 scores reaching into the high 90s. In particular, *Salutatio* shows very strong results, aligning well with the high inter-annotator agreement among human annotators.

Results for the English data For English petitions, *Salutatio* results were consistently strong. Surprisingly, although *Petitio* posed the greatest challenge for human annotators, models more often had difficulty with *Conclusio*, particularly smaller models such as LLaMA 8B and Mistral 7B, and to some extent Mixtral.

LLaMA 70B and Gemini performed consistently well across all parts. Interestingly, GPT-4, while very strong on *Petitio* and *Conclusio*, showed weaker performance on *Salutatio* than other models, often including paragraphs presenting the petitioner (which usually followed the *Salutatio* in the English dataset). By contrast, the smaller models,

Results for the Swedish Dataset - using Swedish Instructions

Prompt	Model	salutatio				petitio				conclusio			
		P	R	F1	TD	P	R	F1	TD	P	R	F1	TD
short 1-shot prompt	GPT-4	98.7	99.9	99.3	0.01	82.5	79.2	80.8	0.29	92.3	92.5	**92.4**	0.10
	Gemini	94.0	**100.0**	96.9	0.02	80.8	80.7	80.7	0.35	**100.0**	62.1	76.6	0.34
	Mixtral	93.8	99.6	96.6	0.02	64.5	66.1	65.3	0.55	79.8	64.0	71.0	0.34
	Mistral	85.3	99.8	92.0	0.04	46.4	83.9	59.7	0.67	60.5	62.1	61.3	0.39
	LLaMA 70B	98.7	99.9	99.3	0.01	82.1	83.6	82.8	0.38	83.0	88.2	85.5	0.19
	LLaMA 8B	98.8	96.1	97.5	0.14	73.8	44.9	55.8	0.58	82.6	53.1	64.6	0.49
long 1-shot prompt	GPT-4	98.7	**100.0**	99.3	0.01	85.3	80.8	83.0	0.27	85.6	93.4	89.3	0.10
	Gemini	94.0	**100.0**	96.9	0.02	**86.5**	80.1	83.2	0.31	**100.0**	68.5	81.3	0.32
	Mixtral	95.7	99.7	97.7	0.02	81.6	62.1	70.5	0.47	91.0	58.2	71.0	0.35
	Mistral	80.1	99.4	88.7	0.05	53.2	**89.3**	66.7	0.55	95.3	59.4	73.2	0.31
	LLaMA 70B	98.7	99.9	99.3	0.01	84.5	83.0	83.7	0.36	81.2	82.9	82.1	0.21
	LLaMA 8B	98.8	96.7	97.7	0.12	75.3	42.1	54.0	0.56	87.2	54.6	67.1	0.47
long 3-shot prompt	GPT-4	98.7	**100.0**	99.3	0.01	85.7	82.8	**84.2**	0.22	86.2	**95.0**	90.4	0.08
	Gemini	**100.0**	**100.0**	**100.0**	0.00	**86.5**	81.2	83.8	0.30	**100.0**	71.6	83.4	0.28
	Mixtral	94.1	99.8	96.9	0.02	79.6	57.9	67.0	0.53	86.7	51.6	64.7	0.35
	Mistral	**100.0**	99.1	99.5	0.04	47.7	82.6	60.5	0.62	88.7	38.4	53.6	0.36
	LLaMA 70B	97.6	99.8	98.7	0.01	86.2	77.8	81.8	0.36	92.0	75.7	83.0	0.25
	LLaMA 8B	98.1	97.7	97.9	0.10	61.3	52.5	56.6	0.57	72.3	71.7	72.0	0.35

Table 4: Results for Swedish data across three Swedish prompt types and six models. Scores includes token-level precision (P), recall (R), and F1 in percentage, and mean token-level edit distance for each predicted span.

though competitive on *Salutatio*, mostly underperformed on *Petitio* and *Conclusio*.

Manual inspection, focusing on *Conclusio* errors from Mixtral, LLaMA 70B, and LLaMA 8B, highlighted different sources of difficulty. Beyond minor punctuation mismatches, some models omitted tags entirely or hallucinated content, such as fabricating a full *Conclusio* where none existed in the gold annotation, or adding phrases not present in the text. LLaMA 8B's particularly low scores for *Conclusio* were further explained by malformed outputs, where tags were placed after the relevant span instead of correctly wrapping it as specified in the prompt.

Results for the Swedish data For the Swedish petitions, as with the English data, strong results were observed for *Salutatio*. Unlike in the English dataset, GPT-4 did not struggle with annotating *Salutatio* in Swedish, achieving one of the highest F1 scores among the models. However, compared to the English data, the models generally found both *Petitio* and *Conclusio* more challenging to annotate in the Swedish Petitions, though performance varied across models.

When comparing models, larger models generally outperformed smaller ones on the Swedish petitions, with GPT-4 being the top-performing model for most petition parts and prompt types, followed by Llama 70B. Gemini and Mixtral also produced several high results, whereas the smaller models, Llama 8B and Mistral, received lower scores, especially for *Petitio* and *Conclusio*.

5.2 The Effect of Prompt Complexity

Across all models and datasets, there is no consistent pattern indicating that prompt length or the number of examples (short 1-shot vs. long 1-shot vs. long 3shot) systematically influences performance. While minor differences appear for specific models or rhetorical sections, these variations do not suggest a general advantage of more detailed or example-rich prompts for this annotation task.

5.3 Cross-Lingual Prompting

Comparing the results for Swedish petitions using Swedish versus English prompts reveals barely any consistent patterns. LLaMA 70B generally performed better with Swedish prompts, suggesting some advantage for this model, while other models showed similar results regardless of prompt language. There was a slight advantage for English prompts on *Salutatio*, whereas *Conclusio* saw marginally better performance with Swedish prompts. However, these differences were small, and the lack of a clear preference overall suggests that prompt language had little systematic effect.

5.4 Difficulty Ratings

Results for model performance and difficulty ratings in comparison to human ratings are presented in Table 6. Looking at the mean error (ME) scores,

Results for the Swedish Dataset - using English Instructions

Prompt	Model	salutatio				petitio				conclusio			
		P	R	F1	TD	P	R	F1	TD	P	R	F1	TD
short 1-shot prompt	GPT-4	98.7	**100.0**	99.3	0.01	84.1	78.9	81.4	0.29	87.3	**92.3**	89.7	0.11
	Gemini	98.7	**100.0**	99.3	0.01	80.1	82.2	81.1	0.34	**100.0**	59.0	74.2	0.36
	Mixtral	96.0	99.9	97.9	0.02	74.3	58.4	65.4	0.55	89.4	58.7	70.9	0.37
	Mistral	95.9	98.4	97.1	0.04	51.6	77.2	61.8	0.65	85.1	54.0	66.1	0.40
	LLaMA 70B	98.7	99.9	99.3	0.01	82.3	74.5	78.2	0.40	87.5	77.9	82.4	0.24
	LLaMA 8B	99.2	97.2	98.2	0.11	72.9	32.4	44.8	0.53	85.6	56.7	68.2	0.46
long 1-shot prompt	GPT-4	98.7	**100.0**	99.3	0.01	**87.1**	**84.3**	**85.6**	0.24	96.7	91.1	**93.9**	0.08
	Gemini	**100.0**	**100.0**	**100.0**	0.00	86.4	82.5	84.4	0.28	**100.0**	63.7	77.9	0.34
	Mixtral	95.8	99.9	97.8	0.01	84.8	68.1	75.5	0.40	86.4	55.5	67.6	0.34
	Mistral	95.7	98.4	97.1	0.04	55.5	72.5	62.8	0.64	85.8	48.0	61.6	0.41
	LLaMA 70B	82.9	60.1	69.7	0.55	78.1	38.0	51.2	0.77	73.7	37.8	50.0	0.74
	LLaMA 8B	99.2	97.4	98.3	0.10	77.1	38.8	51.6	0.48	84.4	53.2	65.3	0.44
long 3-shot prompt	GPT-4	97.7	96.4	97.1	0.06	84.8	80.0	82.3	0.28	85.8	88.5	87.1	0.16
	Gemini	**100.0**	**100.0**	**100.0**	0.00	86.4	81.9	84.1	0.30	99.9	64.3	78.2	0.33
	Mixtral	94.1	99.8	96.9	0.02	80.4	57.5	67.0	0.49	85.1	51.6	64.3	0.31
	Mistral	93.9	98.6	96.2	0.05	55.3	53.9	54.6	0.70	84.3	43.3	57.2	0.39
	LLaMA 70B	83.1	58.6	68.7	0.57	77.4	30.4	43.6	0.81	77.7	37.2	50.3	0.77
	LLaMA 8B	98.8	98.3	98.5	0.09	75.8	50.5	60.6	0.51	83.8	56.4	67.4	0.40

Table 5: Results for Swedish data across three English prompt types and six models. Scores includes token-level precision (P), recall (R), and F1 in percentage, and mean token-level edit distance for each predicted span.

the overwhelming majority of negative values indicates that models, with few exceptions, rate documents as less difficult than humans do. An interesting observation is that model difficulty ratings align most closely with human ratings when models are given detailed prompts with several examples (*long 3-shot*), as reflected by generally lower ME scores in this condition.

The Spearman's correlation coefficients further illustrate the relationship between model and human difficulty assessments. Across prompts and models, correlations ranged from -0.45 to +0.27, with most values being negative. This suggests that passages rated as more difficult by humans tended to be rated as easier by the models. Even the few positive correlations were weak, indicating minimal alignment in the ranking of document difficulty between models and humans.

When comparing model performance to human difficulty ratings, clearer trends are harder to discern. For both languages, model performance tends to be lowest on documents that humans rated as most difficult to annotate (Difficulty 2), but there is considerable variation as indicated by high standard deviations, and there is no clear difference between levels 0 and 1.

6 Conclusion and Future Work

This study has demonstrated that LLMs can perform remarkably well in annotating rhetorical sections within historical petitions, with many models achieving high F1 scores, particularly for *Salutatio*. The results highlight both the capabilities and limitations of current LLMs: while models generally perform strongly across datasets and prompt types, performance varies by section, with *Petitio* and *Conclusio* proving more challenging — particularly for the Swedish data and generally for smaller models. Additionally, although model performance and difficulty ratings correlate to some extent with human ratings, models tend to underestimate document difficulty, suggesting that while they can produce relative difficulty assessments, their ratings may not fully align with human judgments of annotation complexity.

Looking ahead, several avenues for future research emerge from these findings. Firstly, although few-shot prompting yields strong results, training or fine-tuning models specifically on rhetorical annotation tasks may further enhance performance, particularly for more challenging sections such as *Petitio* and *Conclusio*. Fine-tuning on domain-specific data could also improve model calibration, reducing the gap between model and human difficulty ratings. Secondly, future work should explore how segmentation and rhetorical annotation affect downstream tasks such as information retrieval, entity extraction, or social network reconstruction from historical petitions. Given that petitions often embed requests, narrations describ-

Model	Language	Prompt	Mean Err	Spearman ρ	p-value (ρ)	Difficulty 0	Difficulty 1	Difficulty 2
GPT-4	English	short 1-shot	-0.26	-0.05	0.64	91.6 ± 7.1	92.2 ± 6.4	89.4 ± 9.0
		long 1-shot	-0.27	-0.15	0.14	94.5 ± 5.1	94.7 ± 3.8	89.5 ± 8.8
		long 3-shot	**-0.22**	-0.12	0.25	97.2 ± 4.6	98.1 ± 2.9	90.9 ± 12.1
	Swedish	sv short 1-shot	-0.55	-0.40	< 0.01	92.5 ± 12.9	90.4 ± 16.2	81.0 ± 16.6
		sv long 1-shot	-0.37	-0.44	< 0.01	94.1 ± 11.1	92.1 ± 12.3	81.0 ± 16.5
		sv long 3-shot	**-0.03**	-0.36	< 0.01	94.0 ± 11.2	92.0 ± 15.8	86.1 ± 13.9
		eng short 1-shot	-0.56	-0.43	< 0.01	93.4 ± 12.2	90.8 ± 13.9	79.9 ± 16.9
		eng long 1-shot	-0.50	-0.43	< 0.01	94.2 ± 11.8	94.2 ± 10.4	82.6 ± 16.7
		eng long 3-shot	**-0.18**	-0.32	< 0.01	90.6 ± 16.9	86.9 ± 17.6	83.1 ± 19.8
Gemini 1.5 Pro	English	short 1-shot	**0.03**	-0.25	0.01	98.1 ± 3.1	98.2 ± 1.9	92.0 ± 11.2
		long 1-shot	0.08	-0.45	< 0.01	98.3 ± 3.8	97.4 ± 3.3	90.8 ± 11.0
		long 3-shot	0.26	-0.42	< 0.01	98.6 ± 2.6	96.9 ± 6.2	89.7 ± 12.0
	Swedish	sv short 1-shot	**-0.09**	-0.40	< 0.01	91.0 ± 10.5	87.5 ± 12.5	79.1 ± 14.8
		sv long 1-shot	-0.37	-0.20	0.04	89.8 ± 8.9	89.4 ± 10.4	82.7 ± 13.8
		sv long 3-shot	0.14	-0.07	0.49	87.3 ± 14.4	90.9 ± 8.4	86.6 ± 12.2
		eng short 1-shot	**-0.06**	-0.27	0.01	88.3 ± 14.4	88.0 ± 14.0	82.5 ± 12.0
		eng long 1-shot	-0.19	-0.30	< 0.01	90.6 ± 9.1	89.1 ± 10.5	82.4 ± 13.1
		eng long 3-shot	0.15	-0.12	0.22	88.9 ± 10.6	88.4 ± 10.5	86.0 ± 11.6
Mixtral 8x22B	English	short 1-shot	-0.36	-0.38	< 0.01	97.6 ± 4.5	96.2 ± 3.7	89.4 ± 11.6
		long 1-shot	-0.32	-0.39	< 0.01	97.0 ± 6.0	94.9 ± 4.5	92.9 ± 6.6
		long 3-shot	**-0.18**	-0.38	< 0.01	97.7 ± 3.8	95.1 ± 9.0	88.4 ± 12.4
	Swedish	sv short 1-shot	-0.33	-0.21	0.04	80.1 ± 20.7	75.1 ± 18.0	74.6 ± 14.6
		sv long 1-shot	-0.56	-0.33	< 0.01	83.9 ± 16.9	82.8 ± 16.1	71.2 ± 17.1
		sv long 3-shot	**-0.07**	-0.22	0.03	79.5 ± 18.1	72.3 ± 18.5	70.3 ± 21.1
		eng short 1-shot	-0.67	-0.27	0.01	80.9 ± 17.4	75.0 ± 21.9	69.0 ± 19.4
		eng long 1-shot	-0.66	-0.27	0.01	83.6 ± 16.7	87.9 ± 9.5	71.6 ± 19.7
		eng long 3-shot	**-0.13**	-0.23	0.02	79.8 ± 18.0	79.5 ± 18.0	68.4 ± 20.8
Mistral 7B	English	short 1-shot	-0.35	-0.13	0.21	89.7 ± 13.2	93.7 ± 9.0	79.4 ± 18.4
		long 1-shot	-0.31	-0.22	0.03	88.3 ± 13.2	90.3 ± 9.4	79.5 ± 13.5
		long 3-shot	**0.25**	0.04	0.71	82.7 ± 13.8	88.8 ± 8.0	79.6 ± 16.6
	Swedish	sv short 1-shot	-0.25	0.27	0.17	67.5 ± 17.6	78.1 ± 16.9	77.2 ± 15.1
		sv long 1-shot	-0.39	-0.21	0.28	81.8 ± 13.9	79.0 ± 15.9	75.8 ± 13.6
		sv long 3-shot	**-0.05**	-0.08	0.43	70.7 ± 24.5	64.9 ± 17.5	68.8 ± 20.1
		eng short 1-shot	-0.79	-0.03	0.74	69.8 ± 20.9	73.5 ± 18.0	66.9 ± 21.4
		eng long 1-shot	-0.78	-0.07	0.48	71.0 ± 20.0	77.9 ± 15.2	64.8 ± 23.4
		eng long 3-shot	**-0.10**	-0.37	< 0.01	72.8 ± 20.3	69.9 ± 19.9	50.5 ± 21.6
LLaMA-3 70B	English	short 1-shot	-0.34	-0.44	< 0.01	97.6 ± 5.8	98.4 ± 1.8	88.5 ± 13.5
		long 1-shot	-0.34	-0.45	< 0.01	98.7 ± 3.8	98.4 ± 2.1	92.5 ± 9.4
		long 3-shot	**-0.33**	-0.41	< 0.01	98.6 ± 4.4	98.9 ± 1.5	95.8 ± 4.3
	Swedish	sv short 1-shot	-0.55	-0.26	0.01	91.1 ± 11.6	87.5 ± 15.3	86.7 ± 11.6
		sv long 1-shot	-0.41	-0.22	0.03	89.8 ± 13.7	91.6 ± 7.5	87.1 ± 11.1
		sv long 3-shot	**0.01**	-0.23	0.02	89.8 ± 11.7	87.6 ± 11.9	84.8 ± 12.9
		eng short 1-shot	-0.67	-0.19	0.06	88.7 ± 14.0	82.3 ± 19.4	84.5 ± 17.7
		eng long 1-shot	-0.59	-0.02	0.86	54.0 ± 28.2	47.5 ± 23.4	54.7 ± 28.7
		eng long 3-shot	**-0.33**	-0.01	0.96	49.5 ± 25.6	46.4 ± 23.1	50.9 ± 27.5
LLaMA-3 8B	English	short 1-shot	**-0.06**	-0.18	0.07	93.2 ± 8.0	94.8 ± 5.2	83.9 ± 14.1
		long 1-shot	-0.20	-0.25	0.01	94.1 ± 7.7	95.9 ± 3.1	83.7 ± 14.4
		long 3-shot	0.37	-0.37	< 0.01	94.9 ± 7.7	94.7 ± 5.1	84.4 ± 11.5
	Swedish	sv short 1-shot	**0.02**	-0.19	0.06	72.3 ± 20.4	70.7 ± 15.7	62.3 ± 22.2
		sv long 1-shot	0.07	-0.11	0.27	68.6 ± 20.3	71.9 ± 22.0	61.6 ± 22.8
		sv long 3-shot	0.40	-0.19	0.05	76.1 ± 23.3	72.3 ± 24.2	67.6 ± 18.1
		eng short 1-shot	-0.30	-0.18	0.07	67.2 ± 20.7	69.4 ± 21.3	56.7 ± 23.6
		eng long 1-shot	-0.37	-0.24	0.01	71.6 ± 20.0	73.9 ± 20.8	57.4 ± 22.1
		eng long 3-shot	**-0.03**	-0.19	0.06	75.9 ± 22.0	71.0 ± 22.6	67.5 ± 21.9

Table 6: Comparison of model difficulty rating vs. human ratings expressed in Mean Error (Mean Err) and Spearman's correlation coefficients, including p-values, for each prompt type, together with performance of models on different human difficulty ratings.

ing the everyday lives of people, and expressions of social positioning within specific rhetorical sections, improved segmentation may directly enhance the accuracy and interpretability of subsequent analyses.

Finally, while this study has tested different prompt designs varying in length and number of examples, these variations do not yield systematic differences in model performance. It is possible that the different prompt types we employed do not differ substantially enough to impact results, suggesting that LLM outputs for this rhetorical annotation task may be relatively robust to prompt complexity. Moreover, although this work focuses on English and Swedish petitions, expanding to additional languages, including those with even fewer NLP resources, could further illuminate the limitations of current models and the potential for cross-lingual transfer. Together, such investigations would support the development of robust computational workflows for historical document analysis, enabling fine-grained, content-aware annotation at scale to advance humanities research.

Acknowledgments

We warmly thank our colleagues in the Petitions and GaW projects for the inspiring and engaging discussions we shared: Jezzica Israelsson, Örjan

Kardell, Jonas Lindström, Sofia Ling, Maria Ågren, Linda Oja and Fredrik Wahlberg. We are also deeply grateful to Sharon Howard for providing a plain text version of the London Lives dataset and for generously assisting us with our questions. Finally, we thank the anonymous reviewers for their constructive and valuable feedback. The research reported in this paper was supported by a grant from the Swedish Research Council (grant number 2018-06159).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Qi Cheng, Liqiong Chen, Zhixing Hu, Juan Tang, Qiang Xu, and Binbin Ning. 2024. A novel prompting method for few-shot NER via LLMs. *Natural Language Processing Journal*, 8:100099.

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

Gwilym Dodd. 2011. Writing wrongs: the drafting of supplications to the crown in later fourteenth-century England. *Medium Aevum*, 80(2):217–246.

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

Tim Hitchcock, Robert Shoemaker, Sharon Howard, Jamie McLaughlin, et al. 2012. London Lives, 1690–1800. https://www.londonlives.org. Version 1.1, 24 April 2012.

Rab Houston. 2014. *Peasant petitions: social relations and economic life on landed estates, 1600-1850*. Springer.

Sharon Howard. 2016. The London Lives Petitions Project. https://www.londonlives.org. Version 2.0, 2016, based on data from London Lives.

Mark Humphries, Lianne C Leddy, Quinn Downton, Meredith Legace, John McConnell, Isabella Murray, and Elizabeth Spence. 2025. Unlocking the archives: Using large language models to transcribe handwritten historical documents. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, pages 1–19.

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

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B.

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

Aru Maekawa, Tsutomu Hirao, Hidetaka Kamigaito, and Manabu Okumura. 2024. Can we obtain significant success in RST discourse parsing by using large language models? *arXiv preprint arXiv:2403.05065*.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.

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

Charles Spearman. 1904. The proof and measurement of association between two things. *American Journal of Psychology*, 15:72–101.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

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

Michael EW Varnum, Nicolas Baumard, Mohammad Atari, and Kurt Gray. 2024. Large language models based on historical text could offer informative tools for behavioral science. *Proceedings of the National Academy of Sciences*, 121(42):e2407639121.

A Annotation Examples

We include two annotation examples to illustrate how difficulty scores were assigned during annotation. Text included only by Annotator 1 is shown in *purple italics*, while text where both annotators agreed is shown in **green bold**.

Example 1: LMSMPS505520040_1765, disagreement in Petitio, difficulty score 1

<salutatio>To the Worshipfull his Majesty's Justices of the Peace for the County of Middlesex in their General Sessions of the Peace Assembled</salutatio>

The Humble Petition and Appeal of the Churchwardens and Overseers of the Poor of the Parish of Saint Mary le bone in the said County of Middlesex

Sheweth That by Virtue of a Pass Warrant or Order under the Hands & Seals of George Wrighte and Thomas Edwards Esquires two of his Majesty's Justices of the Peace for the City and Liberty of Westminster [...] (one whereof being of the Quorum) bearing Date the 12th. Day of August 1765 Elizabeth Gibson Wife of Bignall Gibson gone [...] James their Child were removd from the Parish of Saint James within the Liberty of Westminster in the said County to the Parish of Saint Mary Le Bone in the said County as the Place of the Last Legal Settlement of the said Bignall Gibson Wife and Child *<petitio>Whereby Your Petitioners Think themselves aggrieved and Appeal to this Court against the same*

<petitio>And Therefore humbly pray this Court will Please to Appoint a Time in this Sessions for hearing and determining the said Appeal And that all Persons removed may then attend.</petitio>

<conclusio>And your Petitioners shall ever Pray Etc</conclusio>

Example 2: LMSMPS502350016_1726, disagreement in Conclusio, difficulty score 2

<salutatio>To the Honble Bench of Justices Novemr. att Hickes Hall</salutatio>

The Humble Petition of the

Churchwarden and Overseer of the poor and Other Anchant Inhabitants of the Hamblett of Mile and New Term in the parish of Stepney on the County of Middxss:

<petitio>Humbly Sheweth that your petitioners begs the favour of this Honble: Bench that they would not Discharge John Bloom now in Custody at the Keeper of Bridwell</petitio> for that he being a Loose Idle Disorderly person and Absenting himself from his familey whereby the Said Hamblet has bin att great Expence and Charge for the two [...] last past for the Supps of the child

<conclusio>[...] Duty on [...]

[...] April [...] } Overseer of the

<conclusio>Joe Mills John Turner [...] } [...] Stable [...]</conclusio>

B Prompt Examples for the English Dataset

```
You are an expert on analysing historical texts. Your task is to identify and label rhetorical sections in petitions from the 18th century using three specific tags.

### Tags to apply:
1. <salutatio>...</salutatio> - opening formal greeting to the recipient(s) of the petition
2. <petitio>...</petitio> - main request(s) being made
3. <conclusio>...</conclusio> - final phrase(s) of courtesy and/or inferiority, often including a signature

At the end, provide an overall score (0-2) for how difficult the text was to tag.
   * **0 (Easy to tag):** All sections are clear and easily identifiable.
   * **1 (Somewhat difficult):** Some sections may be a bit blended or phrasing may be unusual, requiring careful judgment.
   * **2 (Very difficult):** The text is irregular or difficult to interpret, making identification more speculative. The distinction between narrative and request (petitio) can be ambiguous.

### Output Format:
Return only the following two sections. Do not add any explanations, comments, or other text before, between, or after the sections. Use the exact following headings and formatting:

### TAGGED TEXT:
<salutatio>...</salutatio> [any untagged text goes here] <petitio>...</petitio> [any untagged text goes here] <conclusio>...</conclusio>

### DIFFICULTY SCORE:
X

### Example Output:
### TAGGED TEXT:
<salutatio>To the Worshipfull his Majestys Justices of the Peace for the County of Middlesex in their General [---] Sessions of the Peace Assembled</salutatio>

The Humble Petition and Appeal of the Churchwardens and Overseers of the Poor of the Parish of Enfield in the County of Middlesex

Sheweth That by Virtue of a Pass Warrant or order of Removal under the Hands and Seals of John of Hesse and Saunders Welch Esquires two of his Majestys Justices of the Peace for the County of Middlesex (One whereof being of the Quorum) bearing Date the Twenty Sixth Day of October 1774 Robert Pearpoint and Elizabeth his Wife were removed and Conveyed from and out of the Parish of Paddington in the said County to the said Parish of Enfield in the said County of Middlesex as the Place of their last Legal Settlement Whereby your Petitioners conceive themselves to be agrieved

<petitio>Therefore humbly pray your Worships to appoint a Short Day in this present Session to hear and determine their said Appeal</petitio> <conclusio>And your Petitioners shall ever pray Etc

I Smart and Son Attorneys for Appellrs.</conclusio>

### DIFFICULTY SCORE:
0

### Now tag the following petition:
```

Figure 1: Prompt 1 for the English dataset, with less detailed instructions and one given example output.

```
You are an expert on analysing historical texts. Your task is to identify and label rhetorical sections in
     petitions from the 18th century using three specific tags.

### Tags to apply:
1. <salutatio>...</salutatio> - opening formal greeting to the recipient(s) of the petition
2. <petitio>...</petitio> - main request(s) being made
3. <conclusio>...</conclusio> - final phrase(s) of courtesy and/or inferiority, often including a signature

### Core Instructions
1. **Preserve Original Text:** Do NOT add, remove, or change any words, spelling, or punctuation in the
     original text.
2. **Tag Application:** Only apply tags where the content matches one of the three categories in the schema.
3. **Handle Missing Sections:** Sometimes a tag may be missing, though this should be rare.
4. **Handle Multiple Sections:** Tags may appear more than once, especially <petitio>...</petitio>, though
     this should be rare.
5. **Identify Functional Boundary:** When tagging the text segments, the functional boundaries should be
     prioritised over grammatical and/or syntactical boundaries if in conflict. Especially for petitio, this
      means separating circumstances or arguments from the request itself, e.g. "That your Petitioner
     conceives himself to be aggrievd by the said Conviction and humbly <petitio>appeals against the same</
     petitio>".
6. **Difficulty Score:** At the end, provide an overall score (0-2) for how difficult the text was to tag.
   * **0 (Easy to tag):** All sections are clear and easily identifiable.
   * **1 (Somewhat difficult):** Some sections may be a bit blended or phrasing may be unusual, requiring
        careful judgment.
   * **2 (Very difficult):** The text is irregular or difficult to interpret, making identification more
        speculative. The distinction between narrative and request (petitio) can be ambiguous.

### Output Format:
Return only the following two sections. Do not add any explanations, comments, or other text before, between
     , or after the sections. Use the exact following headings and formatting:

### TAGGED TEXT:
<salutatio>...</salutatio> [any untagged text goes here] <petitio>...</petitio> [any untagged text goes here
     ] <conclusio>...</conclusio>

### DIFFICULTY SCORE:
X

### Example Output:
/.../

### Now tag the following petition:
```

Figure 2: Prompt 2 for the English dataset, with more detailed instructions and one given example output (though the example text is left out in this figure).

C Swedish Prompt Examples for the Swedish Dataset

```
Du är expert på att analysera historiska texter. Din uppgift är att identifiera och märka upp retoriska
     segment i svenska suppliker från 1700-talet med hjälp av tre specifika taggar.

### Taggar att använda:
1. <salutatio>...</salutatio> - inledande formell hälsning till mottagaren av suppliken
2. <petitio>...</petitio> - framställning av den huvudsakliga begäran
3. <conclusio>...</conclusio> - avslutande artighets- och/eller underdånighetsfras, ofta inkluderande en
     signatur

Avslutningsvis, ange en övergripande svårighetsgrad (0-2) för hur svår texten var att tagga.
   * **0 - Lätt att taggga**: Alla sektioner är tydliga och lätta att identifiera.
   * **1 - Något svår**: Vissa sektioner kan flyta ihop något eller vissa formuleringar kan vara ovanliga,
       vilket kräver noggrant omdöme.
   * **2 - Mycket svår**: Texten är oregelbunden eller bitvis svårtolkad, vilket gör identifieringen mer
       spekulativ. Distinktionen mellan berättelse och begäran (petitio) kan vara tvetydig.

### Outputformat:
Returnera enbart följande två sektioner. Lägg inte till några förklaringar, kommentarer eller annan text fö
     re, mellan eller efter sektionerna. Använd exakt följande rubriker och formatering:

### TAGGAD TEXT:
<salutatio>...</salutatio> [eventuell otaggad text här] <petitio>...</petitio> [eventuell otaggad text här]
     <conclusio>...</conclusio>

### SVÅRIGHETSGRAD:
X

### Exempel på output:
### TAGGAD TEXT:
<salutatio>Högwälborne H Baron och Landshöfdinge
Nådige Herre</salutatio>


Inför Eders Nåde ähr iag fattige änkia högst Nödsakat mig att beswära, och ödmiukeligast tillkiänna gifwa
     huru som iag långt för detta dehlat om besittningen af 1/8 dehl uti helgiärds hemmanet bregården i
     Carlskouga sochn, med min swåger Oluf Larsson därstädes hwilken mig der ifrån trängt, oacktat hwad rätt
      iag der till äger och hoos gående högl kongl Bergs Collegii bref af d 8 Julij A 1711, samt det höga
     Landshöfdinge Embetets Resolutioner af d 1 och 10 Julij A 1717, mig nåd rättwiseligen tillägga uppå hög
      bem Kongl Collegii bref och dhe i mine Suppliqwer anförde skiähl, sedan hafwer och denna saak wähl
     warit före uti den wähl låfl lagmans tings rätten d 22 Aprill nästl, Men efter den war Incamminerat så
     i högl Kongl Bergs Collegium som och wyd detta Canceliet, Ty ähr den ej till afgiörande företagen
     worden wydare än hoosgående Resolution förmår och utwysar. Wetandes iag ej hwad för Resolution Oluf
     Larsson kunnat sig utwärka i Augusti månad A 1717. Ty så wyda han hållit sig intill Sanningen med sine
     berättelser som Eders Nåde täcktes skåda af min hoosfougade Documenter äro grundade På, så har han
     sannerligen intet Någon annan lydande Resolution kunnat utfå än iag; <petitio>Bönfaller för denskull
     till Eders höga Nåde iag alldra ödmiukast, at, I anseende till min anförde rättmätiga skiähl till be
     min hemmans dehl blifwa restituerat,</petitio> <conclusio>hwar öfwer, en nådig resolution afwacktandes
     deremot iag förblifwer.
Eders Nåds
Alldra ödmiukaste
tienarinna
Margreta Andersdotter
i österwyk.</conclusio>

### SVÅRIGHETSGRAD:
0

### Tagga nu följande supplik:
```

Figure 3: Swedish Prompt 1 for the Swedish dataset, with less detailed instructions and one given example output.

```
Du är expert på att analysera historiska texter. Din uppgift är att identifiera och märka upp retoriska
     segment i svenska suppliker från 1700-talet med hjälp av tre specifika taggar.

### Taggar att använda:
1. <salutatio>...</salutatio> - inledande formell hälsning till mottagaren av suppliken
2. <petitio>...</petitio> - framställning av den huvudsakliga begäran
3. <conclusio>...</conclusio> - avslutande artighets- och/eller underdånighetsfras, ofta inkluderande en
     signatur

### Huvudinstruktioner
1. **Bevara originaltexten:** Lägg INTE till, ta bort eller ändra några ord, stavningar eller skiljetecken i
      originaltexten.
2. **Taggtillämpning:** Använd taggar endast där innehållet matchar en av de tre kategorierna.
3. **Hantera saknade segment:** Det kan förekomma texter där någon eller några av segmenten saknas. Detta bö
     r dock vara sällsynt.
4. **Hantera flera segment:** Taggar kan förekomma mer än en gång, särskilt <petitio>. Även detta bör vara s
     ällsynt.
5. **Semantik vs syntax** Din taggning ska ta stor hänsyn till semantik, inte enbart till grammatik. När du
     taggar <petitio>, inkludera inte omgivande satser eller fraser som endast utgör argument eller
     bakgrundsinformation. Undantag: Du ska inkludera korta bindeord eller fraser i form av kausala markörer,
      som "därför" och "av detta skäl", om de direkt inleder eller avslutar själva begäran.
6. **Ange svårighetsgrad:** Avslutningsvis, ange en övergripande svårighetsgrad (0-2) för hur svår texten
     var att tagga.
  * **0 - Lätt att taggga**: Alla sektioner är tydliga och lätta att identifiera.
  * **1 - Något svår**: Vissa sektioner kan flyta ihop något eller vissa formuleringar kan vara ovanliga,
       vilket kräver noggrant omdöme.
  * **2 - Mycket svår**: Texten är oregelbunden eller bitvis svårtolkad, vilket gör identifieringen mer
       spekulativ. Distinktionen mellan berättelse och begäran (petitio) kan vara tvetydig.

### Outputformat:
Returnera enbart följande två sektioner. Lägg inte till några förklaringar, kommentarer eller annan text fö
     re, mellan eller efter sektionerna. Använd exakt följande rubriker och formatering:

### TAGGAD TEXT:
<salutatio>...</salutatio> [eventuell otaggad text här] <petitio>...</petitio> [eventuell otaggad text här]
     <conclusio>...</conclusio>

### SVÅRIGHETSGRAD:
X

### Exempel på output:
/.../

### Tagga nu följande supplik:
```

Figure 4: Swedish Prompt 2 for the Swedish dataset, with more detailed instructions and one given example output (though the example text is left out in this figure).

D English Prompt Examples for the Swedish Dataset

```
You are an expert on analysing historical texts. Your task is to identify and label rhetorical sections in Swedish petitions from the 18th century using three specific tags.

### Tags to apply:
1. <salutatio>...</salutatio> - opening formal greeting to the recipient(s) of the petition
2. <petitio>...</petitio> - main request(s) being made
3. <conclusio>...</conclusio> - final phrase(s) of courtesy and/or inferiority, often including a signature

At the end, provide an overall score (0-2) for how difficult the text was to tag.
   * **0 (Easy to tag):** All sections are clear and easily identifiable.
   * **1 (Somewhat difficult):** Some sections may be a bit blended or phrasing may be unusual, requiring careful judgment.
   * **2 (Very difficult):** The text is irregular or difficult to interpret, making identification more speculative. The distinction between narrative and request (petitio) can be ambiguous.

### Output Format:
Return only the following two sections. Do not add any explanations, comments, or other text before, between, or after the sections. Use the exact following headings and formatting:

### TAGGED TEXT:
<salutatio>...</salutatio> [any untagged text goes here] <petitio>...</petitio> [any untagged text goes here] <conclusio>...</conclusio>

### DIFFICULTY SCORE:
X

### Example Output:
### TAGGED TEXT:
<salutatio>Högwälborne H Baron och Landshöfdinge
Nådige Herre</salutatio>


Inför Eders Nåde ähr iag fattige änkia högst Nödsakat mig att beswära, och ödmiukeligast tillkiänna gifwa huru som iag långt för detta dehlat om besittningen af 1/8 dehl uti helgiärds hemmanet bregården i Carlskouga sochn, med min swåger Oluf Larsson därstädes hwilken mig der ifrån trängt, oacktat hwad rätt iag der till äger och hoos gående högl kongl Bergs Collegii bref af d 8 Julij A 1711, samt det höga Landshöfdinge Embetets Resolutioner af d 1 och 10 Julij A 1717, mig nåd rättwiseligen tillägga uppå hög bem Kongl Collegii bref och dhe i mine Suppliqwer anförde skiähl, sedan hafwer och denna saak wähl warit före uti den wähl låfl lagmans tings rätten d 22 Aprill nästl, Men efter den war Incamminerat så i högl Kongl Bergs Collegium som och wyd detta Canceliet, Ty ähr den ej till afgiörande företagen worden wydare än hoosgående Resolution förmår och utwysar. Wetandes iag ej hwad för Resolution Oluf Larsson kunnat sig utwärka i Augusti månad A 1717. Ty så wyda han hållit sig intill Sanningen med sine berättelser som Eders Nåde täcktes skåda af min hoosfougade Documenter äro grundade På, så har han sannerligen intet Någon annan lydande Resolution kunnat utfå än iag; <petitio>Bönfaller för denskull till Eders höga Nåde iag alldra ödmiukast, at, I anseende till min anförde rättmätiga skiähl till be min hemmans dehl blifwa restituerat,</petitio> <conclusio>hwar öfwer, en nådig resolution afwacktandes deremot iag förblifwer.
Eders Nåds
Alldra ödmiukaste
tienarinna
Margreta Andersdotter
i österwyk.</conclusio>

### DIFFICULTY SCORE:
0

### Now tag the following petition:
```

Figure 5: English Prompt 1 for the Swedish dataset, with less detailed instructions and one given example output.

```
You are an expert on analysing historical texts. Your task is to identify and label rhetorical sections in Swedish petitions from the 18th century using three specific tags.

### Tags to apply:
1. <salutatio>...</salutatio> - opening formal greeting to the recipient(s) of the petition
2. <petitio>...</petitio> - main request(s) being made
3. <conclusio>...</conclusio> - final phrase(s) of courtesy and/or inferiority, often including a signature

### Core Instructions
1. **Preserve Original Text:** Do NOT add, remove, or change any words, spelling, or punctuation in the original text.
2. **Tag Application:** Only apply tags where the content matches one of the three categories in the schema.
3. **Handle Missing Sections:** Sometimes a tag may be missing, though this should be rare.
4. **Handle Multiple Sections:** Tags may appear more than once, <petitio>...</petitio>, though this should be rare.especially <petitio>, though this should be rare.
5. **Semantics over Syntax:** Your tagging should be guided primarily by semantics, not just grammar. When tagging <petitio>, do not include surrounding clauses or phrases that only provide arguments or background information. Exception: You should include short linking words or phrases that act as anaphoric causal markers, like "therefore" ("därför") and "for this reason" ("av detta skäl") if they directly introduce or conclude the actual request.
6. **Difficulty Score:** At the end, provide an overall score (0-2) for how difficult the text was to tag.
   * **0 (Easy to tag):** All sections are clear and easily identifiable.
   * **1 (Somewhat difficult):** Some sections may be a bit blended or phrasing may be unusual, requiring careful judgment.
   * **2 (Very difficult):** The text is irregular or difficult to interpret, making identification more speculative. The distinction between narrative and request (petitio) can be ambiguous.

### Output Format:
Return only the following two sections. Do not add any explanations, comments, or other text before, between, or after the sections. Use the exact following headings and formatting:

### TAGGED TEXT:
<salutatio>...</salutatio> [any untagged text goes here] <petitio>...</petitio> [any untagged text goes here] <conclusio>...</conclusio>

### DIFFICULTY SCORE:
X

### Example Output:
/.../

### Now tag the following petition:
```

Figure 6: English Prompt 2 for the Swedish dataset, with more detailed instructions and one given example output (though the example text is left out in this figure).

Leveraging RAG for a Low-Resource Audio-Aware Diachronic Analysis of Gendered Toy Marketing

Luca Marinelli[1] **Iacopo Ghinassi**[2] **Charalampos Saitis**[1]
[1]Centre for Digital Music, Queen Mary University of London, UK
[2]College of Computing and Data Science, Nanyang Technological University, Singapore
{l.marinelli, c.saitis}@qmul.ac.uk, iacopo.ghinassi@ntu.edu.sg

Abstract

We performed a diachronic analysis of sound and language in toy commercials, leveraging retrieval-augmented generation (RAG) and open-weight language models in low-resource settings. A pool of 2508 UK toy advertisements spanning 14 years was semi-automatically annotated, integrating thematic coding of transcripts with audio annotation. With our RAG pipeline, we thematically coded and classified commercials by gender-target audience (feminine, masculine, or mixed) achieving *substantial* inter-coder reliability. In parallel, a music-focused multitask model was applied to annotate affective and mid-level musical perceptual attributes, enabling multimodal discourse analysis. Our findings reveal significant diachronic shifts and enduring patterns. Soundtracks classified as energizing registered an overall increase across distinct themes and audiences, but such increase was steeper for masculine-adjacent commercials. Moreover, themes stereotypically associated with masculinity paired more frequently with louder, distorted, and aggressive music, while stereotypically feminine themes with softer, calmer, and more harmonious soundtracks. Code and data to reproduce the results are available on *github.com/marinelliluca/low-resource-RAG*.

1 Introduction

Toy advertisements are a rich site for investigating gendered multimodal discourse, with five decades of research showing persistent and marked gender polarization (Verna, 1975; Johnson and Young, 2002; Marinelli et al., 2024). However, prior studies have relied on manual annotation of relatively small corpora, limiting the ability to conduct large-scale, longitudinal analyses, which are necessary to track the societal impact of evolving stereotypes.

Large language models (LLMs) offer a promising outlook for analyses of large corpora (Xiao et al., 2023; Alonso del Barrio et al., 2024; Gao and Feng, 2025). We extended earlier research on gendered toys marketing by integrating transcript-based thematic coding via RAG and audio-based automatic affective and music tagging. Notably, we computed inter-coder reliability scores to assess the quality of the results.

This study was conducted under self-imposed compute constraints to highlight the benefits of using small open-weight models, rather than relying on pay-walled APIs. There are two key reasons for this approach: first, to ensure reproducibility; and second, to keep computational costs low, making it feasible to replicate this study on consumer hardware. A key issue with pay-walled technology is model deprecation: when models endpoints are retired and become inaccessible, any research based on them becomes immediately non-reproducible. For these reasons, we used smaller open-weight LLMs (4–9 billion parameters) deliberately steering clear of resource-heavy alternatives.

The contributions to the fields of Digital Humanities and Computational Linguistics are manifold. First, this study shows that small open-weight language models combined with RAG can achieve substantial inter-coder reliability, making large-scale discourse analysis feasible on consumer-grade hardware without relying on commercial APIs. Second, it provides a reproducible pipeline that integrates linguistic thematic coding with audio-based affective and musical analysis, enabling large-scale multimodal analysis. Moreover, our empirical findings revealed a constantly evolving *multimodal alignment of gender stereotypes*.

In the following, we briefly introduced the domain of the study, then we surveyed recent work on applications of LLMs for discourse analysis, we then provided details on our data collection, annotation, and RAG pipeline, and finally we presented and discussed the results of the diachronic analysis.

Proceedings of the First Workshop NLP and Language Models for Digital Humanities associated with RANLP 2025, pages 102–111, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-106-6-009

1.1 Toy commercials as gender-based multimodal genres

Across five decades of research, TV advertisements targeted at children have consistently shown marked gender polarization (Verna, 1975; Johnson and Young, 2002; Kahlenberg and Hein, 2010; Marinelli et al., 2024). Differences between feminine-targeted, masculine, and mixed-audience commercials have been registered in sound (voices, background music and sound effects), language, transitions, and camera work, setting, interactions and activities, and colors. Stark gender polarization was observed in both multimodal emotion ratings and perceptual ratings of music in toy adverts (Marinelli et al., 2024). Specifically, masculine-targeted commercials were found to be significantly more aggressive and auditorily abrasive than feminine-targeted adverts.

Music can be imbued with distinct identity dimensions *upon which* ideological discourses are promulgated. Gender is one of these identity dimensions (Dibben, 2002) while androcentrism and heteronormativity are its hegemonic ideological discourses. *Multimodal genres*—which underlie this phenomenon in media portrayals—describe "regular patterns of semiotic choices in multimodal communicative objects and events that are particular to specific communities and cultures" (MODE, 2012). Toy commercials are organized in distinct gender-based multimodal genres. In this work, we explore how multimodal genres change over time, at the intersection of music and language.

2 Recent work on LLMs for discourse analysis

Historically, discourse analysis, including thematic coding, and other forms of qualitative analysis have been seen as domains reserved exclusively for human interpretation (DeJeu, 2025). The introduction of LLMs marks a significant development in qualitative research methodologies. LLMs are well-positioned to assist in qualitative coding, potentially augmenting early analysis, reducing the workload, and expanding the breadth of research corpora through semi-automated coding. Xiao et al. (2023) implemented in-context learning of a pre-compiled codebook with LLMs and achieved fair to substantial agreement with human coders; Gamieldien et al. (2023) employed LLMs to thematically code responses to a physics exam without an initial codebook; Bryda and Sadowski (2024) applied them on podcast interviews to semi-automate the creation of their codebook structure; and Yu et al. (2024) investigated the use of LLMs to automate pragma-discursive corpus annotation of apologies, reporting near to human-level accuracy, although stressing the importance of human oversight.

Curry et al. (2024) have reported negative results for the use of ChatGPT in replication studies that exemplified important tasks in corpus analysis, citing issues of repeatability and replicability tied to its non-deterministic nature (i.e., using its web-based interface). Garg et al. (2024) evaluated the use of LLMs for automated discourse coding in learning analytics on a dataset of questions and responses from secondary school teachers. Even though they obtained promising outcomes through fine-tuning, none met the reliability standards required in their field. Which highlights *the importance of reporting inter-coder reliability measurements*, as opposed to traditional classification metrics that do not account for chance agreement and can lead to the overestimation of performance.

Building from earlier observations of the application and limitations of the use of LLMs in analyzing discourse in corpora, Li and Wang (2024) proposed to improve the prompting method for LLM-based discourse analysis via contextual learning, output formatting, careful task description, and step by step procedure. They reported good results, showing that better prompting can overcome a number of previously described shortcomings.

Successful attempts have also been reported within critical discourse studies. Gao and Feng (2025) employed LLMs to analyze a Hong Kong news corpus and track media attitudes towards China, reporting performance on a par with trained coders. While Alonso del Barrio et al. (2024) showed promising results towards a semi-automated analysis of media content in the complex task of analyzing the framing of TV shows.

3 Methods

3.1 Data collection and annotation

On the 23rd of May 2025 we downloaded 4968 videos from the official YouTube channel of Smyths Toys Superstores, a major UK toy retailer. We paired this sample with another 5614 ads originally downloaded for (Marinelli et al., 2024). Merging these two datasets resulted in many duplicates, which were discarded. Then, only high-quality videos were selected, where ads without audio,

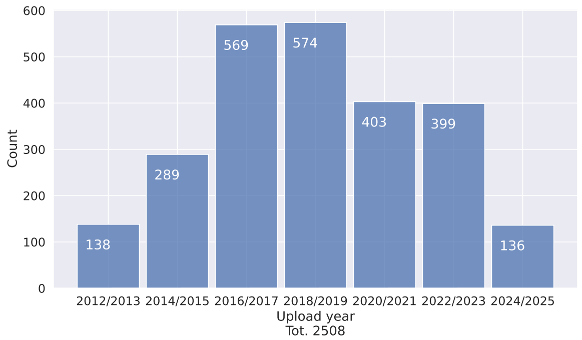


Figure 1: Distribution of the ads by year of upload.

formatted for mobile phones, or with substantial on-screen text were excluded. We then selected commercials where actors' faces were visible, to ensure the internal validity of the gender-targeting construct. Finally, we only kept those commercials that—once transcribed with Whisper large-v3 (Radford et al., 2023)—had transcripts longer than 40 characters. This resulted in **2508**, by and large unique, toy ads spanning a 14-year time frame that were analyzed in this study. Their distribution per *upload* year is shown in Figure 1. In all of the following analyses, the upload years were grouped together two-by-two, as some years had disproportionately less commercials than others.

Ground truth: The present analysis builds on a previous study on the role of music in gendered toy marketing (Marinelli et al., 2024). In the initial ground truth, 606 commercials were manually annotated in terms of their *gender orientation*—i.e., their intended target audience. This was determined by the gender of the presenters and accounted for token representations (Johnson and Young, 2002). In order to determine the inter-coder reliability for this variable, 15% of the commercials was double-coded by two coders independently, and a Krippendorff's alpha of .91 was obtained. Of the 606 commercials, 163 were coded as being targeted at a feminine audience, 149 as targeted at a masculine audience, 200 as targeted at a mixed audience, and 94 did not cast any actors. A pool of 152 musically trained participants rated the soundtracks of the commercials on music-focused scales: Electric/Acoustic, Distorted/Clear, Loud/Soft, Many/Few instruments, Heavy/Light, High/Low pitch, Punchy/Smooth, Wide/Narrow pitch variation, Harmonious/Disharmonious, Clear melody/No melody, Complex/Simple rhythm, Repetitive/Non-repetitive, Dense/Sparse, Fast/Slow tempo, and Strong/Weak beat. A different pool of 151 participants rated the commercials on seven emotion scales: Happy or Delightful, Amusing or Funny, Beauty or Liking, Calm or Relaxing, Energising or Invigorating, Angry or Aggressive, and Triumphant or Awe-inspiring.

In this study, based on insights from (Let Toys be Toys, 2021; de Iulio and Jarrin, 2004), we manually annotated a subset of the masculine, feminine and mixed ads in the ground truth that, once transcribed, contained more than 40 characters. This subset of 467 commercials (which constituted the RAG pool) was manually annotated in terms of the following themes: Domesticity and Nurturing, Fashion and Beauty, Nature and Animals, Love and Tenderness, Magic and Fantasy, Action and Adventure, Fight and Combat, Horror and Monsters, Speed and Racing, and Arts and Crafts.[1] Inter-coder reliability scores (Cohen's kappa) across 467 ads between a human coder (first author) and the RAG pipeline are provided at the end of section 3.3; where due to subpar reliability, three of the ten themes were excluded from the final diachronic analysis.

3.2 Automatic annotation of the soundtracks

We based our multitask transfer-learning model on a previous study performed on the same ground truth (Marinelli et al., 2023). However, differently from that study, we employed CLAP (Contrastive Language-Audio Pretraining) embeddings as audio representation (Elizalde et al., 2024). In addition, we turned the mid-level music-focused descriptors and the emotion scales into binary classifications by binning them on the 50% percentile, and we only kept those scales that performed well above chance in a 10-fold cross validation (i.e., scoring an average F1 above .60). The remaining emotion scales were Happy or Delightful, Beauty or Liking, Calm or Relaxing, and Angry or Aggressive; whilst the mid-level descriptors were Strong beat/Weak beat, Electric/Acoustic, Distorted/Clear, Loud/Soft, Heavy/Light, High pitch/Low pitch, Punchy/Smooth, Harmonious/Disharmonious, and Dense/Sparse. The resulting average F1 for the emotion scales was $.67 \pm .03$, whilst for the mid-level descriptors it was $.75 \pm .05$.

[1]Themes related to consumerism, competitions and sports, science and technology, and fun and play, were initially coded, but were later excluded due to their too broad applicability.

3.3 RAG pipeline

It is well-known that language models are few-shot learners (Brown et al., 2020; Schick and Schütze, 2021) as these models are able to achieve strong performance on many downstream NLP tasks without updating any parameters (i.e., without fine-tuning) by pairing their existing knowledge base and understanding of natural language with few examples or context directly within the prompt. Notably, Liu et al. (2022) found that retrieval-augmented generation (RAG) helps to considerably reduce hallucinations and stabilise performance, by coupling LLMs with an external pre-trained retriever model (Reimers and Gurevych, 2019). With this approach, the examples provided to the LLM are dynamically retrieved so that they only reflect the most relevant labels to the current data point, thereby making up for issues related to the short context windows. In particular, Milios et al. (2023) demonstrated that RAG with off-the-shelf retriever models can deliver robust performance in text classification tasks that involve many labels. In addition, with only 16 examples per class, the clever prompting proposed by Sun et al. (2023), Clue And Reasoning Prompting (CARP), achieved comparable performances, in text classification, to supervised models with 1,024 examples per class.

Themes detection: To perform the detection of the themes, we implemented what could be called a “reverse CARP” technique (see Listing 1) where the model is presented with negative and positive examples of transcripts that are either unrelated or related to the theme under analysis, in order to give it access to patterns and cues associated with the theme. Both negative and positive examples are retrieved via the retriever model (cosine similarity) and each of the positive examples is presented with a list of cues related to the theme under analysis. The model is then given a specific theme definition, the current datapoint transcript, and is asked to determine if cues related to the theme are present. The model is then asked to provide a brief reasoning paragraph, and to return a list of cues related to the theme (or an empty list otherwise). Then, any hallucinated cue not present in the transcript is automatically removed. Finally, a theme is deemed present when the system returned a non-empty list.

Target classification: The classification between feminine, mixed audiences, and masculine-targeted commercials was broken down in five sub-tasks: one vote from each binary subset of the classes (i.e., feminine/masculine, feminine/mixed, mixed/masculine), one vote from the full classification (where the LLMs were presented with all three classes), and one tie-breaking vote from the music-focused model. This is justified under the assumption that the models would be more reliable when comparing only two classes at a time. Once collected all votes, the hard-coded decision logic would either choose the most frequent class, or in case of a tie, choose the final class from the corresponding sub-task.

The prompting structure employed for each sub-task is shown in Listing 2. First the model is presented with a set of examples that illustrate the different possible classes, each with their corresponding themes and transcripts. The model is then instructed to analyze the themes (automatically collected at the previous step) and transcript of the current datapoint, taking into account potential gender stereotypes to determine the target audience. The model is then asked to provide a reasoning paragraph that explains its decision, following a given structure that highlights the relevant themes and their relationship to the target audience. Finally, the model is prompted to return the inferred class, choosing depending on the sub-task, from a predefined set of possible values.

Model	Average F1 across themes (1, 5, 10 examples)		
CohereForAI/c4ai-command-r7b	82 ± 05	$\mathbf{82 \pm 04}$	81 ± 06
microsoft/Phi-3-small-8k-instruct	80 ± 07	79 ± 07	79 ± 07
microsoft/Phi-3.5-mini-instruct	81 ± 05	81 ± 06	$\mathbf{82 \pm 04}$
allenai/Llama-3.1-Tulu-3.1-8B	77 ± 07	77 ± 08	75 ± 08
google/gemma-2-9b-it	81 ± 05	81 ± 06	80 ± 06

Table 1: Preliminary evaluation of the themes detection. The reported deviations are computed across themes.

Preliminary evaluation: All experiments were run on a single NVIDIA A5000 GPU. Different language models between 4 and 9 billion parameters were evaluated on the RAG pool consisting of 467 documents. Considering that the music model was also being trained on the 606 ground truth commercials (thus including data from the RAG pool), we needed to implement a cross-validation algorithm within which we positioned the RAG pipeline. This consisted in excluding the data points of the test set (i.e, fold) from the available RAG pool at each iteration. Which means, that whilst we performed a 10-fold cross-validation for the music model—which on target classification achieved an F1 of $.70 \pm .05$—the RAG pipeline was instead effectively evaluated as a leave-one-out cross-validation,

Model	Full logic (1, 5, 10 examples)			Without music model (1, 5, 10 examples)			3-classes sub-task (1, 5, 10 examples)		
CohereForAI/ c4ai-command-r7b	71 [62, 79]	71 [62, 79]	71 [63, 79]	64 [55, 73]	66 [57, 74]	68 [59, 76]	67 [58, 75]	64 [55, 73]	64 [55, 73]
microsoft/ Phi-3-small-8k-instruct	73 [65, 81]	68 [58, 76]	71 [63, 79]	65 [55, 73]	61 [51, 69]	62 [53, 71]	73 [64, 80]	63 [54, 72]	66 [56, 74]
microsoft/ Phi-3.5-mini-instruct	76 [68, 83]	77 [69, 84]	73 [65, 81]	71 [63, 79]	71 [62, 79]	68 [59, 76]	70 [62, 78]	72 [63, 80]	68 [60, 76]
allenai/ Llama-3.1-Tulu-3.1-8B	77 [69, 84]	78 [70, 85]	77 [69, 85]	75 [67, 83]	74 [66, 81]	74 [65, 82]	75 [66, 82]	75 [66, 82]	77 [69, 84]
google/ gemma-2-9b-it	77 [69, 84]	79 [71, 86]	78 [70, 85]	75 [67, 82]	78 [70, 85]	76 [68, 84]	74 [66, 82]	75 [67, 82]	73 [65, 80]
c4ai (theme) + gemma-2 (target)	//	**80 [72, 87]**	//	//	78 [70, 85]	//	//	75 [66, 82]	//

Table 2: Preliminary evaluation of the target classification, F1 scores with corresponding 95% CI (bootstrapped).

with a slightly reduced RAG pool at each fold.

For the sake of brevity, the results of the theme evaluation are reported as averages across themes in Table 1. The best-performing models were, on a par, Microsoft's Phi-3.5-mini-instruct with 10 examples (effectively, 10 negative and 10 positive examples), and Cohere's c4ai-command-r7b with 5 negative and 5 positive examples. Generally, most models performed similarly well on this task.

The results of the target classification are instead reported in Table 2, where given the higher level of abstraction of this task, we also performed bootstrapping (with replacement, 10k iterations) to provide the corresponding 95% confidence intervals. The observed performance peak at 5 examples *per class* may be due to the limited size of the RAG pool, with more examples leading to an increased proportion of non-relevant examples being presented to the model. Microsoft's Phi-3.5-mini, AllenAI's Llama-Tulu, and Google's Gemma performed similarly well, with Gemma coming out slightly ahead. As our last evaluation, we used Cohere's model to detect the themes, which then were provided as context to Gemma. This resulted in the best-performance, with an F1 score of .80 [.72, .87]. It is with this final combination that we proceeded to analyze the larger corpus of toys commercials.

Listing 1: Themes detection prompt (shortened).

```
# Negative examples: unrelated to {current_theme}
{negative_examples}

# Positive examples: related to {current_theme}
{positive_examples}

# Theme definition
Examples of {current_theme} contain cues referring
to {current_theme_definition}.

# Current datapoint
transcript: {current_transcript}

# INSTRUCTIONS
Based on the theme definition and on the examples
determine if the current datapoint contains cues
about {current_theme}. First, provide a reasoning
paragraph, then return the list of cues. If no
relevant cues are found return an empty list.
```

Listing 2: Target classification prompt (shortened).

```
# EXAMPLES (grouped by class)
{examples}

# Current datapoint
transcript: {current_transcript}
themes: {current_themes}

# Definitions of the collected themes
{current_themes_definitions}

# INSTRUCTIONS
Based on the previous examples, determine the target
audience for the current datapoint. First, choose
only one of the following reasoning structures.

In the current transcript, the themes <theme1>,
<theme2> .. are mainly associated with femininity
so the the target of the toy ad is 'Girls/women'.

or

In the current transcript, the themes <theme1>,
<theme2> .. are mainly associated with masculinity
so the the target of the toy ad is 'Boys/men'.

or ...

Finally, return the above determined value.
Choose only one from: {current_sub-task_classes}.
```

Inter-coder reliability: Before proceeding any further we provided the Cohen's kappa and related 95% confidence intervals (with replacement, 10k repetitions) for each of the themes and for the classification of the gender-based target audience. Due to space constraints, we only focus on the best-performing combination highlighted in Table 1. The theme Domesticity and Nurturing achieved a κ of .72 [.52, .89]; Fashion and Beauty obtained a κ of .76 [.57, .91]; Nature and Animals .67 [.50, .81]; Love and Tenderness .65 [.46, .81]; Magic and Fantasy .77 [.59, .91]; Action and Adventure .45 [.30, .60]; Fight and Combat .71 [.49, .89]; Speed and Racing .78 [.62, .91]; Horror and Monsters .54 [0, 1] (due to numerical errors during bootstrapping); Arts and Crafts .57 [.36, .75]. Considering the traditional threshold of .61—where a score between .61 and .8 indicates *substantial* agreement (Landis and Koch, 1977, p. 165)—we excluded the following themes from the analysis of the results:

Action and Adventure, Horror and Monsters, and Arts and Crafts. Finally, the main logic of the best-performing combination of models achieved a κ of .69 [.57, .79] on the gender target classification.

4 Results

Besides the 467 commercials from the ground truth, 2041 unseen ads were automatically annotated with the previously described pipeline. In the following we reported the results of analyses performed on the hybrid dataset of 2508 commercials.

Diachronic analysis: Spearman's rank correlation coefficients were computed between upload years and the *ratio of positive predictions*. For the themes this number is simply the ratio of commercials where a theme is predicted as present in the transcript. Concerning unipolar affective scales, positive predictions corresponded to commercials with a soundtrack that was predicted as belonging to the upper 50% percentile, while for bipolar mid-level descriptors (e.g. Electric/Acoustic) positive predictions corresponded to the rightmost polarities of the scales, which were binned in the ground truth as the upper 50% percentile.

A minimum count of 20 commercials per bin (at the intersection of control variable and upload year) was imposed. In one case, at the intersection of the theme Love and Tenderness and the years 2024/2025, we dropped said bin from the analysis, as it contained less than 20 commercials, and therefore representativeness could not be guaranteed.

First, no relevant trends were detected across gender targets—that is, across the entire corpus. However, once grouped by predicted gender target, two relevant trends emerged for the predicted themes. Within commercials that were predicted as targeted at a mixed audience, the theme Nature and Animals has been almost steadily decreasing over the last 14 years (Spearman's $\rho = -.86, p = .014$), as reported in Figure 2a. Similarly, although to a lesser extent, as reported in Figure 2b, the theme Fight and Combat has been steadily decreasing within commercials predicted as targeted at a masculine audience ($\rho = -.93, p = .003$). No trends were detected for any theme within commercials targeted at a feminine audience.

No relevant trends were also detected for emotions within gender targets. However, as reported in Figure 3a, commercials classified as targeted to a mixed audience show a negative trend in the ratio of soundtracks with a weak beat ($\rho = -.82$, $p = .023$). Similarly, in Figure 3b, for commercials classified as targeted to a masculine audience, the ratio of those with a weak beat or with a soft soundtrack has been steadily decreasing (both with $\rho = -.93, p = .003$).

Theme	ρ	p	Scale
fight_combat	-.93	.003	Strong beat/Weak beat
	-.89	.007	Loud/Soft
	-.79	.036	Punchy/Smooth
love_tenderness	-.83	.042	Strong beat/Weak beat
	-.89	.019	Electric/Acoustic
magic_fantasy	-.79	.036	Punchy/Smooth
speed_racing	-.86	.014	Loud/Soft
	-.93	.003	Punchy/Smooth
	-.86	.014	Electric/Acoustic
	-.86	.014	Distorted/Clear
	-.79	.036	Strong beat/Weak beat

Table 3: Spearman's coefficients and p-values of themes-wise time trends of the mid-level music descriptors.

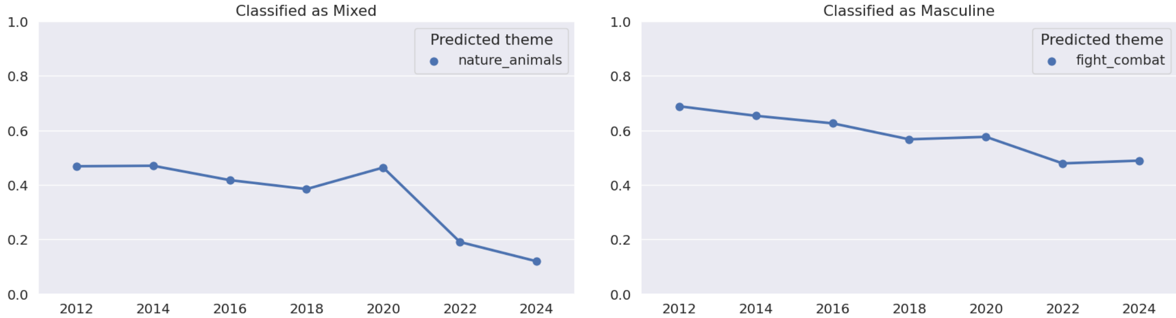
When grouped by theme, no relevant trend was detected in any emotion scale. Instead, several significant correlations emerged from the analysis of mid-level music descriptors within commercials grouped by theme. Which were reported in Figures 4a to 4d and in Table 3, with the corresponding correlation coefficients and p-values.

Interactions between scales and themes: Finally, we explore the interaction between emotions and mid-level descriptors predicted from the soundtracks, and the themes found in spoken language that were annotated with the RAG pipeline. In Figure 5 we reported the ratio of soundtracks—grouped by theme—that were predicted in the upper 50% percentile of each scale.

A clear pattern emerged, where themes stereotypically associated with femininity (de Iulio and Jarrin, 2004) co-occurred with soundtracks that were softer, happier, calmer, more harmonious, lighter, and—with the exception of the theme Fashion and Beauty—with weaker beats, smoother and more acoustic than electric; where such exception is likely a reference to the electronic dance music that accompanies fashion shows. Conversely, themes associated with masculinity were mostly paired with angrier, louder, heavier, more distorted, [...], and more disharmonious soundtracks.

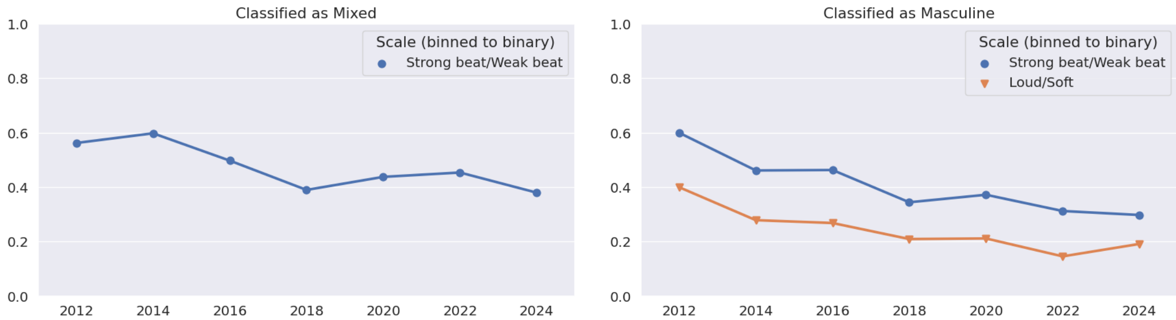
5 Discussion

As evidenced in the results, and as discussed in a previous study (Marinelli et al., 2024), the mid-level music descriptors in the ground truth are collinear. Therefore, trends related to those scales are better interpreted along two latent axes: the



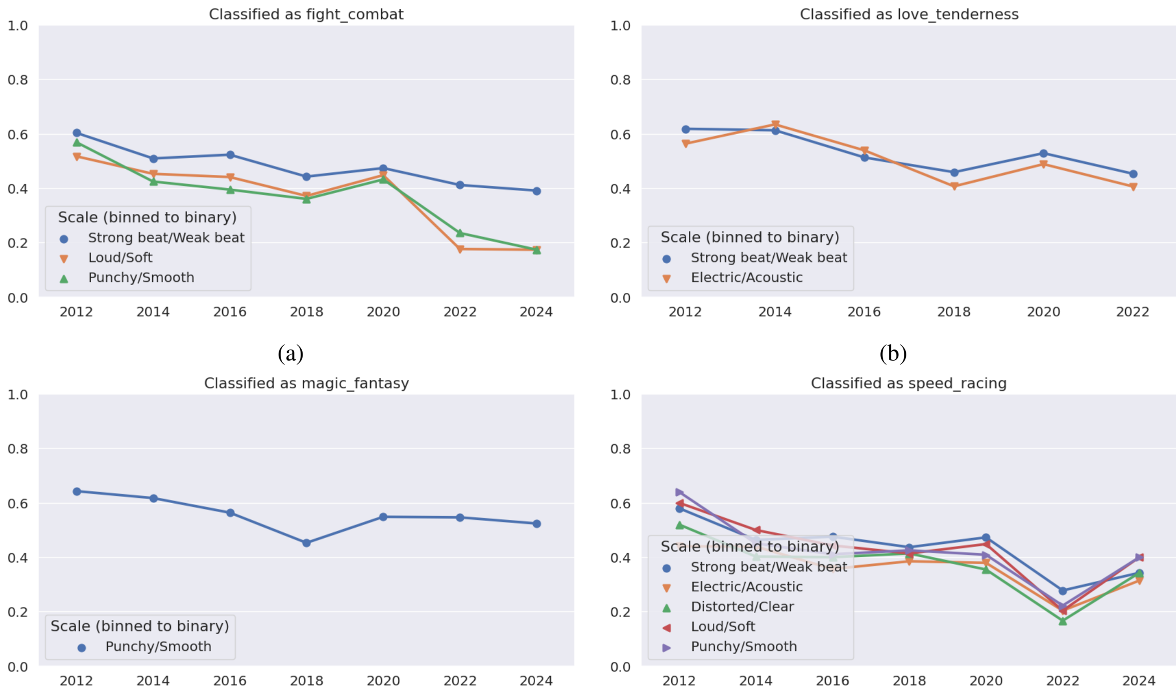
(a) (b)

Figure 2: Ratio of commercials grouped by gender target that contain the predicted theme.



(a) (b)

Figure 3: Ratio of commercials grouped by gender target in the upper 50% percentile of the scales.



(a) (b) (c) (d)

Figure 4: Ratio of commercials grouped by theme in the upper 50% percentile of the scales.

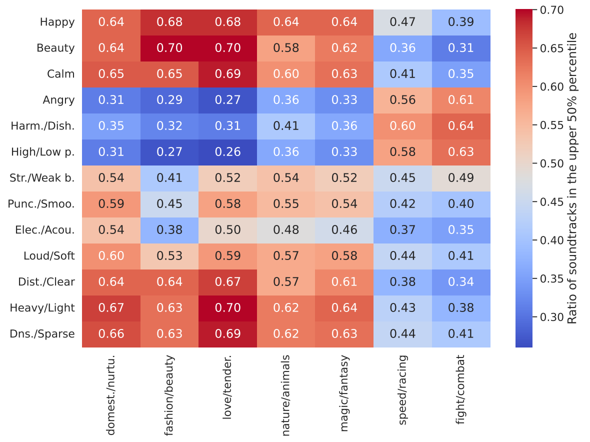


Figure 5: Heatmap showing the ratio of soundtracks grouped by theme (columns) predicted in the upper 50% percentile of each scale (rows).

"Energizing/Soothing" axis and the "Harmonious and clear/Dissonant and distorted" axis.

Concerning the reported diachronic changes, a counter-stereotypical trend was found in the decreasing ratio of masculine-targeted commercials referring to Fight and Combat (Figure 2b). However, this change appeared to be offset by an overall increase of loud soundtracks with a strong beat in masculine-targeted ads (Figure 3b), also reported for mixed-audience commercials (Figure 3a). In particular, ads referring to Fight and Combat (Figure 4a) show a substantial increase in energizing soundtracks. A similar trend appeared also for the theme Love and Tenderness (4b) although to a lesser degree. Moreover, ads referring to Speed and Racing (Figure 4d) show the most pronounced increase in both energizing and dissonant/distorted soundtracks. Therefore, it appears that masculine-adjacent commercials have become more abrasive over the last 14 years, with little to no change to report for feminine-targeted ones.

Tracking these patterns showed that, despite widespread public backlash (Fine and Rush, 2018; Marinelli et al., 2024), toy commercials continue to reinforce traditional gender stereotypes through increasingly polarized multimodal strategies. Specifically, the finding that masculine-adjacent commercials have become more auditorily aggressive, while feminine ones remain largely unchanged, suggests that gender polarization may be widening rather than narrowing. Moreover, diachronic analysis can reveal counter-intuitive trends, such as the decrease of Fight and Combat being offset by increasingly aggressive soundtracks, showing how stereotyped media portrayals can evolve in form while maintaining their underlying polarization.

6 Conclusion

Our best-performing configuration achieved substantial agreement with a human annotator. Specifically, it achieved an average Cohen's κ of .66 across ten themes[2] (average F1 of .82) and a κ of .69 for gender-target classification (F1 of .80) in a leave-one-out cross-validation on 467 ads. The pipeline was then applied on 2041 unseen commercials, for a total of 2508 toy ads spanning 14 years. Evolving patterns emerged in the interactions between language and music in gendered toy advertisements. Our findings reveal a *multimodal alignment of gender stereotypes*, where stereotypically feminine themes in the transcripts co-occur with soft, calm, harmonious soundtracks, while stereotypically masculine themes consistently align with loud, aggressive, and distorted soundtracks. An overall increase of energizing soundtracks was reported for masculine-targeted, mixed-audience commercials, their themes, and even for one traditionally feminine theme (Love and Tenderness). However, such increase was steeper for masculine-targeted commercials and associated themes.

This study had a few limitations. In the ground truth, the gender orientation of the commercials was based on the gender of the actors in each video. However, the pipeline based its decisions on patterns in text and audio which can change over time, undermining the internal validity of this particular inference. Another limitation, which we partially addressed by providing confidence intervals for the performance metrics, is that the RAG pipeline was built using the same set on which it was evaluated.

Future studies might consider testing the pipeline across different cultures, as portrayals may vary between countries. Moreover, adding visual analysis such as colors, character positions, and facial expressions would provide a complete picture of how text, audio, and image cooperate in gendered advertising. Finally, preregistration and other strategies should be considered to ensure that a portion of unseen data is also manually coded by experts and only disclosed to system developers at evaluation time. This is especially relevant when automated annotation is employed on corpora that are orders of magnitude larger than the initial ground truth.

[2]However, three subpar themes were excluded from the final analysis. The remaining themes averaged at a κ of .72.

Acknowledgments

This work was supported by UK Research and Innovation [grant number EP/S022694/1].

Iacopo Ghinassi worked on this study whilst still affiliated to Queen Mary University of London.

References

David Alonso del Barrio, Max Tiel, and Daniel Gatica-Perez. 2024. Human interest or conflict? leveraging llms for automated framing analysis in tv shows. In *Proceedings of the 2024 ACM International Conference on Interactive Media Experiences*, IMX '24, page 157–167, New York, NY, USA. Association for Computing Machinery.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Grzegorz Bryda and Damian Sadowski. 2024. From words to themes: Ai-powered qualitative data coding and analysis. In *Computer Supported Qualitative Research*, pages 309–345, Cham. Springer Nature Switzerland.

Niall Curry, Paul Baker, and Gavin Brookes. 2024. Generative ai for corpus approaches to discourse studies: A critical evaluation of chatgpt. *Applied Corpus Linguistics*, 4(1):100082.

Emily Barrow DeJeu. 2025. Can (and should) llms perform critical discourse analysis? *Journal of Multicultural Discourses*, pages 1–8.

Nicola Dibben. 2002. Gender identity and music. In *Musical Identities*. Oxford University Press.

Benjamin Elizalde, Soham Deshmukh, and Huaming Wang. 2024. Natural language supervision for general-purpose audio representations. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 336–340.

Cordelia Fine and Emma Rush. 2018. "Why does all the girls have to buy pink stuff?" The ethics and science of the gendered toy marketing debate. *Journal of Business Ethics*, 149(4):769–784.

Yasir Gamieldien, Jennifer M Case, and Andrew Katz. 2023. Advancing qualitative analysis: An exploration of the potential of generative ai and nlp in thematic coding. *Available at SSRN 4487768*.

Qingyu Gao and Dezheng (William) Feng. 2025. Deploying large language models for discourse studies: An exploration of automated analysis of media attitudes. *PLOS ONE*, 20(1):1–17.

Ryan Garg, Jaeyoung Han, Yixin Cheng, Zheng Fang, and Zachari Swiecki. 2024. Automated discourse analysis via generative artificial intelligence. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*, LAK '24, page 814–820, New York, NY, USA. Association for Computing Machinery.

Simona de Iulio and Zouha Jarrin. 2004. Toy commercials across Europe. *Young Consumers*, 5(4):39–45.

Fern Johnson and Karren Young. 2002. Gendered voices in children's television advertising. *Critical Studies in Media Communication*.

Susan G Kahlenberg and Michelle M Hein. 2010. Progression on nickelodeon? gender-role stereotypes in toy commercials. *Sex roles*, 62(11):830–847.

J Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33 1:159–74.

Let Toys be Toys. 2021. Who gets to play now? – New research on TV toy ads. Accessed: 2025-19-06.

Bingru Li and Han Wang. 2024. Tacomore: Leveraging the potential of llms in corpus-based discourse analysis with prompt engineering.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Luca Marinelli, György Fazekas, and Charis Saitis. 2023. Gender-Coded Sound: Analysing the Gendering of Music in Toy Commercials via Multi-Task Learning. *24th International Society for Music Information Retrieval Conference (ISMIR 2023)*.

Luca Marinelli, Petra Lucht, and Charalampos Saitis. 2024. A Multimodal Understanding of the Role of Sound and Music in Gendered Toy Marketing. *PLOS ONE*, 19(11):1–32.

Aristides Milios, Siva Reddy, and Dzmitry Bahdanau. 2023. In-context learning for text classification with many labels. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pages 173–184, Singapore. Association for Computational Linguistics.

MODE. 2012. Glossary of multimodal terms. Edited by Jewitt, Carey and Bateman, John. Accessed: 2025-19-06 at https://multimodalityglossary.wordpress.com/genre/.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

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.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text classification via large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8990–9005, Singapore. Association for Computational Linguistics.

Mary Ellen Verna. 1975. The female image in children's tv commercials. *Journal of Broadcasting & Electronic Media*, 19(3):301–309.

Ziang Xiao, Xingdi Yuan, Q. Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer. 2023. Supporting qualitative analysis with large language models: Combining codebook with gpt-3 for deductive coding. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23 Companion, page 75–78, New York, NY, USA. Association for Computing Machinery.

Danni Yu, Luyang Li, Hang Su, and Matteo Fuoli. 2024. Assessing the potential of llm-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apology. *International Journal of Corpus Linguistics*, 29(4):534–561.

Quantifying Societal Stress: Forecasting Historical London Mortality using Hardship Sentiment and Crime Data with Natural Language Processing and Time-Series*

Sebastian Olsen
Amsterdam University College
sebolsen2@gmail.com

Jelke Bloem
University of Amsterdam
j.bloem@uva.nl

Abstract

We study links between societal stress - quantified from 18th–19th century Old Bailey trial records - and weekly mortality in historical London. Using MacBERTh-based hardship sentiment and time-series analyses (CCF, VAR/IRF, and a Temporal Fusion Transformer, TFT), we find robust lead–lag associations. Hardship sentiment shows its strongest predictive contribution at a 5–6 week lead for mortality in the TFT, while mortality increases precede higher conviction rates in the courts. Results align with Epidemic Psychology and suggest that text-derived stress markers can improve forecasting of public-health relevant mortality fluctuations.

1 Introduction

"Great fears of the sickenesse here in the City, it being said that two or three houses are already shut up. God preserve us all!" This entry from Samuel Pepys' diary, written on Sunday 30 April (1665), encapsulates the dread that gripped London during the Great Plague. Accounts such as these show how epidemics are not purely biological phenomena, but events that have a widespread impact on social order. The model of Epidemic Psychology put forward by the sociologist Strong (1990) outlines how societies can be caught up in what he calls a "maelstrom" of impact, altering social structures.

This study investigates whether indicators of societal stress, expressed through historical crime records, relate to fluctuations in total weekly mortality. Such fluctuations are often indicative of public-health crises, but also reflect broader societal vulnerabilities. Researchers such as Alcabes (2009) have explored the social impact of epidemics, with studies primarily focusing on their sociological impacts (Chu et al., 2020; Muthuswamy, 2024). Modern quantitative studies exist, with Kastalskiy et al. (2021) investigating the impact of societal stress on the COVID-19 pandemic. Their research finds that countries able to manage stress can avoid strong second disease waves. These results highlight the potential for investigating similar dynamics in historical contexts. However, there is a distinct lack of quantitative research utilising high-quality historical data. We aim to better understand how societal stress has influenced mortality trends and fill this gap in historical quantitative studies.

Our primary research question is: "How are patterns of crime, judicial outcomes, and the sentiment expressed in criminal trial accounts related to weekly mortality fluctuations in London, reflecting societal stress during periods of crisis?". Our work builds on previous studies that link text sentiment to real-world measures (Bentley et al., 2014) and on quantitative legal history and sentiment analysis (Zhao et al., 2022; Pan et al., 2021).

Research Context and Motivation Modern studies (e.g., Kruspe et al. 2020; Wang et al. 2020) show that public sentiment can shift almost in real time - lockdowns and policy announcements produce immediate mood changes on social media. Such rapid shifts reflect how heightened societal stress and widespread hardship can impair decision-making, fuel policy overreactions, and undermine social stability (FeldmanHall et al., 2015). Historical accounts (for example Bishop de Belsunce on the Great Plague of Marseille; Devaux 2012) suggest comparable social dynamics may have occurred in earlier epidemics. By quantitatively analysing trial texts alongside weekly mortality, we

*Code for this study is available at: https://github.com/Seb-Olsen/ranlp25-hardship-mortality

Proceedings of the First Workshop NLP and Language Models for Digital Humanities associated with RANLP 2025, pages 112–119, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-106-6-010

ask whether spikes in archival hardship language precede or follow mortality surges and whether trial sentiment echoes Strong's "maelstrom" of fear-driven social collapse (Strong, 1990). Applying NLP and time-series forecasting gives us an empirical way to test these dynamics in historical data.

Contributions (1) We construct a hardship sentiment proxy from historical trial texts using MacBERTh and ABSA-style embedding similarity; (2) We quantify lead–lag structure with CCF, VAR/IRF, and predictive tests, separating contemporaneous from delayed associations; (3) We benchmark forecasting with TFT against naïve and seasonal baselines and interpret drivers via variable importance; (4) We connect mortality dynamics with judicial responses, providing quantitative evidence consistent with Epidemic Psychology.

2 Related Work

This research is situated at the intersection of historical NLP, time-series forecasting, and public health. The application of transformer-based models to time-series forecasting is a burgeoning field. The Temporal Fusion Transformer (TFT) (Lim et al., 2021) has shown state-of-the-art performance in various domains like retail and finance, but its application to historical or epidemiological data is less common. Makhonza et al. (2024) successfully applied TFT to modern mortality forecasting, demonstrating its potential. Our work extends this by applying TFT to a uniquely challenging historical dataset, characterised by noise, reporting lags, and non-stationarity.

Furthermore, the use of NLP to extract public health indicators is a well-established practice, with systematic reviews highlighting its application to infectious disease surveillance (Agbehadji and Awad, 2023) and population mental health monitoring (Gkotsis et al., 2024). Specific applications are diverse: foundational studies used sentiment analysis to predict flu outbreaks (Pan et al., 2021) and measure public reaction to health policies (Zhao et al., 2022), while more recent work employs sophisticated transformer models for detailed emotion classification (Cui et al., 2024) and leverages the latest generation of LLMs to analyse health discussions on social media (Zhang et al., 2024). Our study is novel in its application of these techniques to a large-scale historical corpus to create a sentiment-based proxy for societal stress. By linking this proxy to administrative records on mortality and crime, we create a new, quantitative lens through which to study historical societal dynamics.

3 Methodology

3.1 Data Sources

Our primary text source is the Old Bailey Sessions Papers XML corpus (Hitchcock et al., 2023), covering 1678–1849. This corpus was chosen for its detailed trial accounts and precise weekly dating, which mitigates the publication lag found in other historical text corpora and is crucial for time-series analysis. Crucially, its focus on London allowed for direct geographical and temporal alignment with our mortality dataset. The source boasts a 99.99% accurate transcription rate (Hitchcock and Turkel, 2016). We restricted our analysis to 1719–1829 to maximise data utility. We programmatically extracted structured metadata – `offenceCategory`, `verdictCategory`, and `punishmentCategory` – from the records (Hitchcock and Shoemaker, 2006), yielding a total of 138,078 trial observations.

For mortality data, we used the London Bills of Mortality dataset (Smith et al., 2020), which provides weekly mortality counts from 1644–1849. To stabilise variance and mitigate extreme spikes, we log-transformed the weekly death counts, creating the `log_deaths` variable. Izdebski et al. (2022) have previously used such mortality data as a reliable indicator of epidemic impact in pre-modern periods.

3.2 Sentiment Analysis

A key challenge was the linguistic complexity of historical texts and the lack of temporal generalization ability of language models (Verkijk et al., 2025). We leveraged MacBERTh, a BERT model adapted for historical English (1450-1950) (Devlin et al., 2019; Manjavacas and Fonteyn, 2022). Its pre-training on historical corpora provides robustness to spelling variations and helps mitigate semantic drift (Kutuzov and Giulianelli, 2020), making it highly suitable for fine-grained semantic tasks. Recent work has specifically demonstrated the value of using MacBERTh for aspect-based sentiment analysis (ABSA) in literary-historical contexts (Dejaeghere et al., 2024).

Informed by this precedent, we adopted an ABSA approach to quantify hardship within the trial narratives. We generated embeddings for each trial's text and computed reference embeddings for a curated list of hardship-related terms (Table 1). The resulting hardship sentiment score was obtained by cosine similarity between trial embeddings and the hardship reference vectors.

Terms for hardship sentiment		
poor	poverty	necessity
distress	hardship	starve
desperate	ruin	beggar
vagrant	hunger	want

Table 1. Terms used to generate reference embeddings indicative of hardship sentiment.

3.3 Validation and Robustness of the Hardship Measure

We qualitatively validated our hardship measure by auditing a stratified sample of trials. High-scoring texts clearly contained hardship narratives. For instance, one trial described a prisoner who begged a judge for mercy, stating: "What a sad thing will this be for my Wife, who has not a Farthing in the World." In contrast, low-scoring trials were typically terse and factual. To test robustness, we compared our MacBERTh-based scores to a simple keyword-frequency baseline; correlations and lead–lag patterns were similar. This triangulation supports the validity of our construct, though further multi-annotator evaluation remains a priority for future work.

3.4 Time-Series Analysis and Forecasting

To explore the temporal dynamics between our variables, we employed several time-series analyses. Cross-Correlation Functions (CCF) were used to examine lead-lag relationships (Shumway and Stoffer, 2000). We then performed Granger causality tests to assess whether one time series could be used to forecast another (Granger, 1969). We tested for lags from 1 to 12 weeks.

Finally, to model the system's response to shocks, we fitted a Vector Autoregression (VAR) model. A VAR model expresses each variable as a linear function of its own past values and the past values of all other variables in the system (Sims, 1980). From the fitted VAR model, we generated Impulse Response Functions (IRFs), which trace the dynamic effect of a one-time, one-standard-deviation shock in one variable on the future trajectory of another variable, holding other shocks constant. This allows for a deeper analysis of the bivariate relationships.

For the primary forecasting task, we implemented the Temporal Fusion Transformer (TFT) (Lim et al., 2021), a neural network architecture with built-in interpretability. We optimised hyperparameters using the Optuna framework, aiming to minimise Symmetric Mean Absolute Percentage Error (SMAPE). The model was trained using `QuantileLoss` at the 0.1, 0.5, and 0.9 quantiles to estimate prediction intervals. We evaluated its performance against three standard baselines: a Naïve (last-value) forecast, a Seasonal Naïve (same-week-last-year) forecast, and a Historical Average forecast.

4 Results

The final DataFrame contains 5,768 weekly rows and 28 features (aggregated `hardship_sentiment`, crime proportions, conviction/punishment rates, trial counts, a `year_end_spike` indicator, and mortality variables). Trial-level hardship scores excluded 5,549 very-short trials (<30 words), i.e., $\approx 4.02\%$, to improve reliability (see Figure 1).

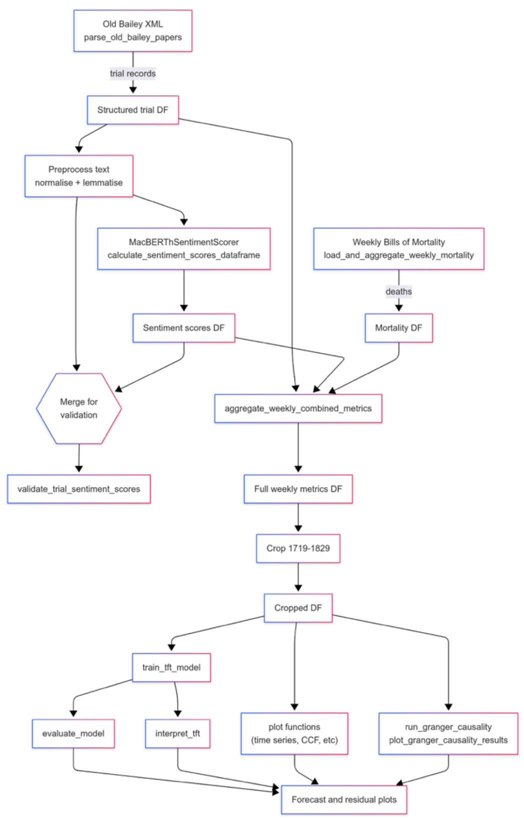


Figure 1. Pipeline for sentiment analysis and mortality data forecasting.

4.1 Exploratory Data Analysis

The weekly death count (Figure 2) shows a gradual decrease over the period, likely due to public health improvements (Porter, 1991). The prominent 'jitters' are artefacts of historical reporting, where parishes aggregated death counts at year-end (Figure 3). The conviction rate (Figure 4) was steady around 60%, but spiked to 97% between 1791–1793, possibly reflecting fears following the French Revolution (Eastwood, 1995). Across the series mean weekly deaths decline from $\approx$ 500 in the early 1700s to $\approx$ 420 by 1829, motivating the log-transform used throughout. The dataset therefore combines long-run decline with irregular, year-end aggregation spikes that we explicitly model.

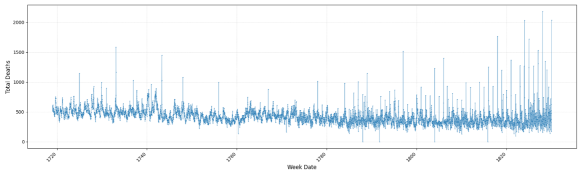


Figure 2. Weekly Death Rates (1719–1829).

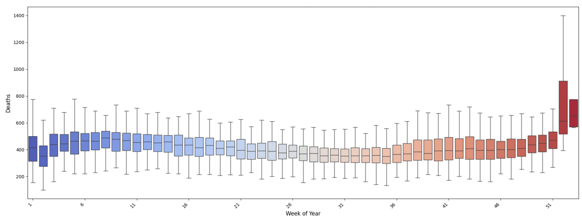


Figure 3. Distribution of Deaths by Week (1719–1829).

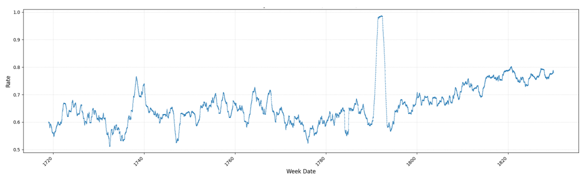
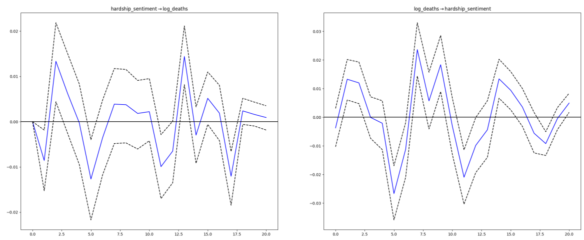


Figure 4. Weekly Conviction Rate (1719–1829).

Analyses of trial and verdict counts around 1788–1792 show that the conviction-rate spike was not driven by fewer trials, suggesting a genuine increase in conviction likelihood during that period.

4.2 Shock Analysis with VAR/IRF

VAR/IRF analysis shows that an unexpected shock in hardship sentiment leads to a statistically significant increase in the growth rate of `log_deaths` about two weeks later, followed by oscillatory effects (Figure 5a). Conversely, a shock in `log_deaths` prompts smaller but statistically significant responses in hardship sentiment (Figure 5b).



(a) Hardship → Deaths IRF (b) Deaths → Hardship IRF

Figure 5. Impulse response functions (95% CIs).

4.3 Lead–lag associations and predictive tests

The cross-correlation plot (Figure 6) shows a modest peak when `hardship_sentiment` leads `log_deaths` by 4–6 weeks, and a stronger, significant correlation when deaths lead sentiment by 17–20 weeks. A five-year rolling correlation (Figure 7) reveals a highly dynamic relationship, with a strong positive peak (>+0.6) around the 1740 "Great Frost" (Engler et al., 2013).

Granger causality tests (Figure 8) for `hardship_sentiment` causing `log_deaths` were not statistically significant. However, tests show that `log_deaths` significantly Granger-causes the conviction rate, particularly with a lag of two or more weeks (Figure 9).

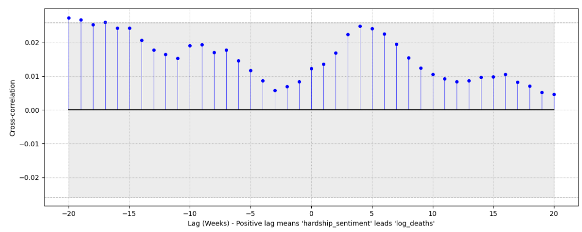


Figure 6. Cross-correlation of hardship and deaths.

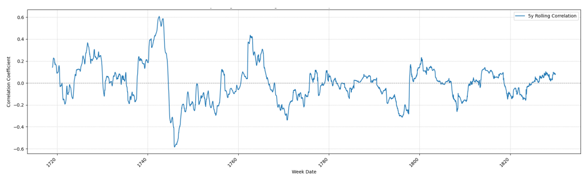


Figure 7. Five-year rolling correlation.

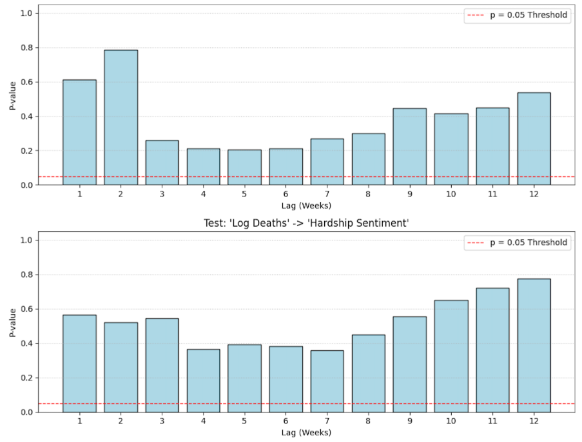


Figure 8. Granger Test: Hardship → Deaths.

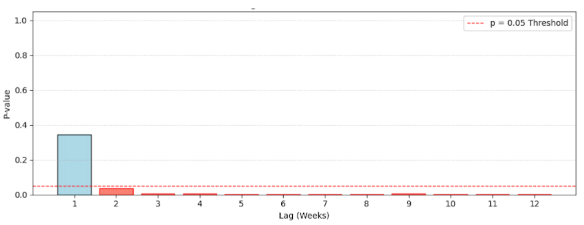


Figure 9. Granger Test: Deaths → Conviction Rate.

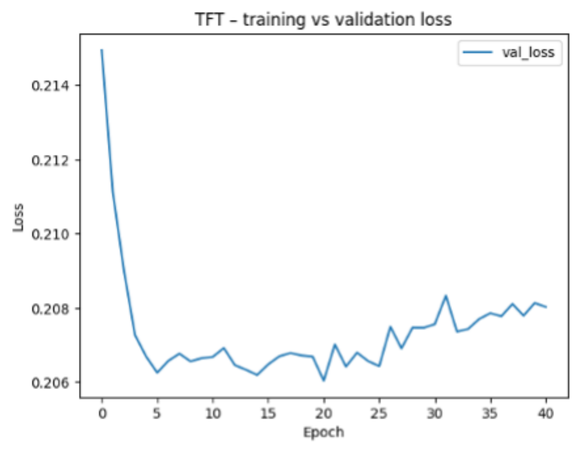
4.4 TFT Model Performance and Error Analysis

The TFT model outperformed all baselines (Table 2), achieving an MAE of 115.9. An error analysis (Figure 10) reveals the TFT's robustness. The validation loss curve (Figure 10a) shows that early stopping at epoch 20 prevented overfitting. The forecast visualization (Figure 10b) shows that the TFT model successfully captures the overall mortality trend and year-end spikes, unlike the baselines. The Seasonal Naïve model performs worst (SMAPE: 43.8%), likely because it overfits to past seasonal patterns and is brittle to the irregular shocks that drive historical mortality.

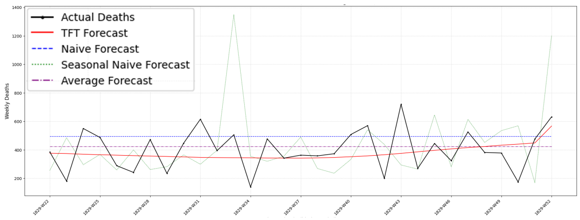
Model	MAE	MSE	SMAPE (%)	RMSE
TFT	115.9	20396.4	30.9	142.8
Naive	136.5	28921.7	34.3	170.1
Seasonal Naive	191.5	68216.0	43.8	261.2
Average	118.4	20562.2	31.0	143.4

Table 2. Forecasting performance of TFT and baselines.

The model's interpretability features highlight which signals drive forecasts. Among decoder variables, the `year_end_spike` indicator was most influential ($\approx 45\%$; Figure 11a). For encoder variables, hardship sentiment with a six-week lag was the strongest predictor ($\approx 16\%$; Figure 11b).

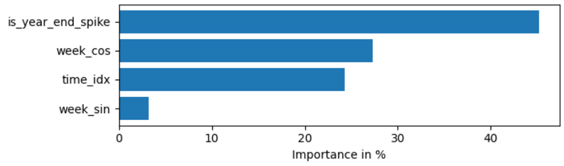


(a) TFT Validation Loss (early stop at epoch 20).

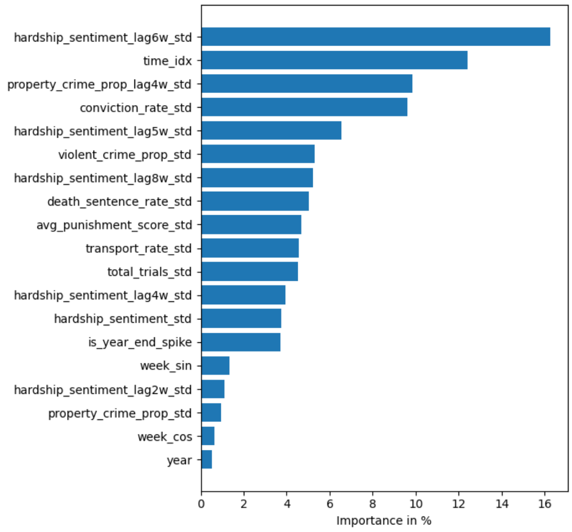


(b) Actual vs. TFT and Baseline Forecasts - Validation Period

Figure 10. TFT model performance and forecast visualization.



(a) TFT – Decoder Variables Importance.



(b) TFT – Encoder Variables Importance.

Figure 11. TFT model interpretability: feature importance.

5 Discussion

This research finds that patterns of crime, judicial outcomes, and sentiment in trial accounts are significantly related to weekly mortality in historical London. The finding that hardship sentiment lagged by six weeks is the top encoder variable for the TFT model strongly supports the hypothesis that hardship has a tangible, though delayed, relationship with mortality. This delay may be attributable to the time it takes for chronic stress and worsening living conditions to physically impact individuals, a conclusion supported by sociological research on unemployment and immune dysfunction (Milner et al., 2013; Balakin et al., 2025).

The relationship is multifaceted and bidirectional. While contemporaneous correlation is weak ($r = 0.01$), time-lagged analyses reveal a more complex story. The CCF, VAR/IRF, and TFT results all point to hardship sentiment having its greatest impact on mortality with an approximate six-week delay. Furthermore, mortality increases appear to precede rises in hardship sentiment by about 17-20 weeks, and Granger-cause a rise in conviction rates after a lag of two weeks. This indicates a bidirectional lead–lag pattern wherein public health crises influence public sentiment and judicial responses.

The finding that increased deaths Granger-cause higher conviction rates (Figure 9) raises complex questions about judicial objectivity, corroborating conclusions from Jedwab et al. (2021). Was the judicial system swayed by public fear, leading to a search for 'scapegoats' during periods of high mortality? Was it used by authorities to restore order after periods of high societal stress? These questions merit further research into the response of the legal system to drivers of societal stress.

These results contribute directly to Strong's (1990) concept of the "maelstrom", where disease outbreaks are followed by suspicion and moral controversy. As Strong notes, "friends, family and neighbours may be feared... the world may be turned upside down" (1990, p. 252). Our findings build on this by showing that hardship sentiment and judicial metrics can act as quantitative predictors of future weekly mortality. Methodologically, this offers a new way to track this historical effect and serves as a strong use case for applying historical language models like MacBERTh to extract meaningful, predictive signals from complex textual archives.

5.1 Limitations and Future Work

One key limitation is the quality of 18th-century data reporting. Spikes in weekly deaths are often artifacts of year-end data aggregation. While our `year_end_spike` feature helps mitigate this, future work could also benchmark our ABSA-style approach against other sentiment analysis techniques. A second challenge lies in the historical texts themselves, which contain spelling variations and semantic drift that are only partially mitigated by models like MacBERTh.

Another limitation is the use of total weekly mortality as a proxy for epidemic periods. This aggregate measure includes deaths from other causes potentially influenced by societal stress (e.g., malnutrition). Future work with more disaggregated, cause-specific death records could differentiate the impact of hardship more clearly. The findings of this study are specific to London from 1719-1829; generalizing these links to other contexts or modern data requires caution and further investigation to validate the relationships.

A final limitation concerns causal interpretation. While lead–lag structure is suggestive, our analyses cannot rule out unobserved confounders (e.g., climate, policing intensity). We therefore interpret findings as predictive associations only.

6 Conclusion

This study explored the temporal relationships between societal stress indicators and weekly mortality in 18th and 19th-century London. Our analysis revealed several key findings: (1) hardship sentiment, particularly with a 5-6 week lag, is a significant leading predictor of weekly mortality fluctuations; (2) patterns of crime also demonstrate predictive value; and (3) a bidirectional pattern is observed, with rising mortality improving forecasts of subsequent conviction rates.

The demonstrated link between text-derived hardship and mortality suggests the potential utility of similar sentiment analysis techniques on modern, large-scale textual data (e.g., social media, news reports) for early-warning systems of societal distress. By providing a quantitative lens on Epidemic Psychology, this work underscores the far-reaching societal impacts of public health crises and opens new avenues for both historical inquiry and contemporary policy.

References

Israel Eben Agbehadji and Alhassan Awad. 2023. A systematic review of machine learning and deep learning on social media for public health surveillance of infectious diseases. *Social Network Analysis and Mining*, 13(1):52.

Philip Alcabes. 2009. *Dread: How Fear and Fantasy Have Fueled Epidemics from the Black Plague to Avian Flu*. PublicAffairs.

E. Balakin, K. Yurku, M. Ivanov, A. Izotov, V. Nakhod, and V. Pustovoyt. 2025. Regulation of stress-induced immunosuppression in the context of neuroendocrine, cytokine, and cellular processes. *Biology*, 14(1):76.

R. Alexander Bentley, Alberto Acerbi, Paul Ormerod, and Vasileios Lampos. 2014. Books average previous decade of economic misery. *PLoS ONE*, 9(1):e83147.

Irene Y. Chu, Parisha Alam, Heidi J. Larson, and Li Lin. 2020. Social consequences of mass quarantine during epidemics: a systematic review with implications for the COVID-19 response. *Journal of Travel Medicine*, 27(7).

Yubao Cui, Xin Li, Shuai Ma, and Qi Zhang. 2024. Roberta-based multi-label emotion classification for public health social media data. *Applied Sciences*, 14(5):2075.

Tanne Dejaeghere, Pushpinder Singh, Els Lefever, and Julie Birkholz. 2024. Exploring aspect-based sentiment analysis methodologies for literary-historical research purposes. In *Proceedings of the 4th Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)*.

Christian A. Devaux. 2012. Small oversights that led to the great plague of Marseille (1720–1723): Lessons from the past. *Infection, Genetics and Evolution*, 14:169–185.

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*, pages 4171–4186.

David Eastwood. 1995. E. P. Thompson, Britain, and the French revolution. *History Workshop Journal*, 39:79–88.

Steven Engler, Franz Mauelshagen, Jürg Luterbacher, and Johannes P. Werner. 2013. The irish famine of 1740–1741: famine vulnerability and "climate migration". *Climate of the Past*, 9(3):1161–1179.

Oriel FeldmanHall, Candace M. Raio, Jonathan T. Kubota, Michael G. Seiler, and Elizabeth A. Phelps. 2015. The effects of social context and acute stress on decision making under uncertainty. *Psychological Science*, 26(12):1918–1926.

Georgios Gkotsis, Alison O'Mara-Eves, Stan Zammit, Samuel R Chamberlain, David Gunnell, Trevor Thompson, and James Thomas. 2024. Natural language processing for public mental health surveillance: A scoping review. *JMIR Mental Health*, 11:e48275.

Clive W. J. Granger. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438.

Tim Hitchcock and Robert Shoemaker. 2006. Digitising history from below: The Old Bailey proceedings online, 1674-1834. *History Compass*, 4(2):193–202.

Tim Hitchcock, Robert Shoemaker, Clive Emsley, Sharon Howard, and Jessica McLaughlin. 2023. The Old Bailey proceedings online, 1674–1913. Version 9.0.

Tim Hitchcock and William J. Turkel. 2016. The Old Bailey proceedings, 1674–1913: Text mining for evidence of court behavior. *Law and History Review*, 34(4):929–955.

Adam Izdebski, Piotr Guzowski, Roman Poniat, Luca Masci, J. Palli, C. Vignola, M. Bauch, C. Cocozza, R. Fernandes, F. C. Ljungqvist, and 1 others. 2022. Palaeoecological data indicates land-use changes across Europe linked to spatial heterogeneity in mortality during the black death pandemic. *Nature Ecology & Evolution*, 6(3):297–306.

Remi Jedwab, Allan M. Khan, Jonathan Russ, and Esha D. Zaveri. 2021. Epidemics, pandemics, and social conflict: Lessons from the past and possible scenarios for COVID-19. *World Development*, 147:105629.

Igor A. Kastalskiy, Elena V. Pankratova, Evgeny M. Mirkes, Victor B. Kazantsev, and Alexander N. Gorban. 2021. Social stress drives the multi-wave dynamics of COVID-19 outbreaks. *Scientific Reports*, 11(1):2021.

Anna Kruspe, Matthias Häberle, Iona Kuhn, and Xiao Xiang Zhu. 2020. Cross-language sentiment analysis of European twitter messages during the COVID-19 pandemic. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

Andrey Kutuzov and Mario Giulianelli. 2020. UiO-UvA at SemEval-2020 task 1: Contextualised embeddings for lexical semantic change detection. ArXiv preprint arXiv:2005.00050.

Bryan Lim, Sercan Ö. Arık, Nicolas Loeff, and Tomas Pfister. 2021. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764.

B. Makhonza, N. Mogodi, and R. Mbuvha. 2024. Mortality forecasting using temporal fusion transformers. *SSRN Electronic Journal*.

Enrique Manjavacas and Lauren Fonteyn. 2022. Adapting vs. pre-training language models for historical languages. *Journal of Data Mining & Digital Humanities*.

Allison Milner, Andrew Page, and Anthony D. LaMontagne. 2013. Long-term unemployment and suicide: A systematic review and meta-analysis. *PLoS ONE*, 8(1):e51333.

V. V. Muthuswamy. 2024. Epidemics and society: A historical lens on public health and community resilience. *Journal of Natural Science, Biology and Medicine*, 15:58–71.

Wenju Pan, Ru Wang, Wenting Dai, Guiping Huang, Chang Hu, Wei Pan, and S. Liao. 2021. China public psychology analysis about COVID-19 under considering sina weibo data. *Frontiers in Psychology*, 12.

Roy Porter. 1991. Cleaning up the Great Wen: public health in eighteenth-century London. *Medical History*, 35(S11):61–75.

Robert H. Shumway and David S. Stoffer. 2000. *Time Series Analysis and Its Applications*, 1st edition. Springer, New York.

Christopher A. Sims. 1980. Macroeconomics and reality. *Econometrica*, 48(1):1–48.

R. S. Smith, R. D. Davenport, and G. N. Newton. 2020. London weekly bills of mortality, 1644-1849.

P. Strong. 1990. Epidemic psychology: a model. *Sociology of Health & Illness*, 12(3):249–259.

Stella Verkijk, Piek Vossen, and Pia Sommerauer. 2025. Language models lack temporal generalization and bigger is not better. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20629–20637.

Shihan Wang, Marijn Schraagen, Erik Tjong Kim Sang, and Mehdi Dastani. 2020. Dutch general public reaction on governmental COVID-19 measures and announcements in twitter data.

Zian Zhang, Yixuan Sun, Martha Effting, Iris van der Vegt, and A. J. Ton Klein. 2024. Assessing the utility of ChatGPT for public health surveillance: a case study of social media analysis on heat-health discussions. *Frontiers in Public Health*, 12:1358607.

Shu Zhao, Lin Chen, Yang Liu, Ming Yu, and H. Han. 2022. Deriving anti-epidemic policy from public sentiment: A framework based on text analysis with microblog data. *PLoS ONE*, 17(8):e0270953.

Exploring Language in Different Daily Time Segments Through Text Prediction and Language Modeling

Kennedy Roland
St. Francis Xavier University
Antigonish, NS, Canada
x2022gds@stfx.ca

Milton King
St. Francis Xavier University
Antigonish, NS, Canada
mking@stfx.ca

Abstract

Temporal-aware language models have proved to be effective over longer time periods as language and its use changes, but little research has looked at how language use can change at different times of the day. We hypothesize that a person's usage of language varies at different times of day. We explore this concept by evaluating if models for language modeling and next word prediction improve their performance when considering the time of day. Specifically, we explore personalized temporal-aware models for next-word prediction and language modeling and compare them against baseline models, including non-temporal-aware personalized models. Specifically, our proposed model considers which of the 8, 3-hr daily time segments that a text snippet was written during for a given author. We found that our temporal-aware models tend to outperform temporal-agnostic models with respect to accuracy and perplexity.

1 Introduction

Language models are often trained on large amounts of text from many different people but do not necessarily consider the time that the text was written. In this work, we tailor a general pre-trained language model (GPT2) (Radford et al., 2019) toward a single person and the time of day that their text was written. The intuition behind this is that we believe that the same individual's use of a language varies throughout different times of the day and we can use models to explore that. For example, a person's text snippets in morning could differ from their text snippets in the evening. As far as we have found, there has been no prior work considering the use of language at different times of day. Improving the performance of a language model through considering the time of day that the text was written, could potentially assist authorship attribution models (Fabien et al., 2020), although additional experiments are required to validate this.

Classifiers have been found to have improved performance when tested on the same time period they were trained on compared to other intervals, both annually and seasonally (Huang and Paul, 2018). Considering the month a post was made in has been shown to have an impact on document classification using a pre-trained BERT model when looking at Reddit posts to classify which political subreddit they belonged to (Röttger and Pierrehumbert, 2021). Models trained on an even shorter time frame, like day of the week, have also been seen to outperform temporal-agnostic models for tasks like word sense disambiguation (Wei and King, 2024). We now aim to look at the potential of time impacting next-word prediction and language modeling in an even shorter duration. Many applications that support sharing text often include timestamps with text and therefore it is reasonable to consider this type of data in real scenarios. Due to the size of our dataset, we had segmented time of day into 8, 3-hour segments, but acknowledge that the size of the time segments could be a hyperparameter that could be tuned to select the best performing time range size.

2 Related Work

Although there is a lot of previous work looking at language models with a temporal context, relatively low amounts of research have focused on authors speaking differently at various times in the context of improving next-word prediction and language modeling. Some research showed that there is a benefit in using temporal data for various other tasks including document classification (Huang and Paul, 2018; Röttger and Pierrehumbert, 2021). It's suggested that classifiers perform better when they are applied to the time period they were

Proceedings of the First Workshop NLP and Language Models for Digital Humanities associated with RANLP 2025, pages 120–125, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-106-6-011

trained on, both on a seasonal level and by year (Huang and Paul, 2018). They recommend training a classifier from the most current chronological samples instead of randomly in order to get the best performance (Huang and Paul, 2018). Word sense disambiguation models — models that are tasked with assigning a sense to a word in context — have also been shown to perform better when tailored toward temporal segments (Wei and King, 2024).

Rosin et al. (2021) proposed tempoBERT, a model that considers the context of time and used that alongside prepended text indicating the year that it was written to initially help with time based facts; they also explored semantic change and sentence time prediction. They found that by using smaller language models, they were able to produce state-of-the-art results, outperforming the larger models (Rosin et al., 2021). Rosin and Radinsky (2022) worked on temporal attention by creating a matrix of the input and embedding the time vector into that, allowing for the language model to consider time without changing the input at all, so there isn't a need to prepend text based on what year it is. King and Cook (2020) applied a similar technique, which they refer to as priming, where they input tokens from text from the author to an LSTM-based language models before being evaluated on some text for testing. They evaluated their models using adjusted perplexity (a variation of perplexity), accuracy@k, and accuracy@k given the first c number of characters from the target token.

Another contribution that involves temporal information is tempLAMA, a dataset which contains queries with time-sensitive answers, making it a good tool to test temporal language models (Dhingra et al., 2022). They used tempLAMA to train language models and found similar results, where language models that consider time performed better than time-agnostic models that were trained on more text and temporal-aware models have consistently performed better on these time-sensitive questions compared to time-agnostic models (Dhingra et al., 2022).

There has been work and discussion around keeping human nature in the field of natural language processing. Hovy and Yang (2021) suggest that there are many things that change the way the same person will write text such as who the receiver of the message is, what the event/occasion is or, based on the topic. They suggest that we need to get closer to social understanding of humans to better language prediction (Hovy and Yang, 2021). We agree with the idea that social modeling can benefit with the addition of more context and our work focuses on exploring if time of day has any impact on this.

Soni et al. (2022) expanded on the idea that we need human social understanding and had a similar idea to our hypothesis in the sense that humans will use language differently in various situations. They segmented things by the human state and were able to make the model aware of what state of being the individual was in, personifying a bit of the machine part of natural language processing (Soni et al., 2022). In our case, the different states can be comparable to what time of day the author is writing during.

3 Dataset

We randomly selected 50 subreddits[1] from a list of popular communities[2]. We then extracted the 25 most recent posts from each of those subreddits. We reselected a new subreddit if the originally selected subreddit primarily consisted of images. We collected all public comments and posts from the users that were authors of each selected post from the original list of subreddits, which includes comments and posts from other subreddits. The intent of gathering text from different subreddits by the same user was to increase topic diversity for a given user. We removed any authors that had less than 25 posts or comments total (across all subreddits, selected originally and otherwise) from the dataset. We converted the times of all posts and comments from UTC into our local time (ADT). This was done because there was no associated location and time zone info publicly available and it is important to mention that we do not have access to the actual time of day for the timezone of the author when the text was posted. Therefore, our analysis and experiments will be considering the time relative to each individual author and we cannot compare the same time segment across authors. We then separated all posts and comments from each author into time segments (3-hour periods beginning at 12AM ADT). This resulted in 8 different time segments within a day. Authors that had not posted across a minimum of two time segments were removed.

[1] Subreddits are topic-specific forums or sub-sites where users can post and comment on the social media platform, Reddit.

[2] https://www.reddit.com/best/communities/1/

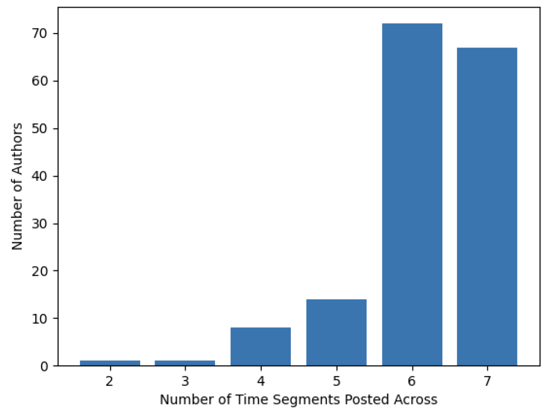


Figure 1: The distribution of the number of time segments the authors had posted across.

This resulted in 163 different authors and 1,008 author/time segment combinations across those authors — up to 8 time segments per author. The number of time segments per author can be seen in Figure 1). Most authors had posted across 6 or 7 time segments, with no authors posting across all eight segments. There are very few authors that only posted in 2 or 3 time segments. There are 2.34 posts and 101.01 comments per author on average, with an average length of 36.56 words. All text included in this dataset is in English. Text that was detected to be in multiple languages were removed from the dataset before ensuring they met the minimum post and time segment requirements.

4 Experimental Setup

The tasks that we focus on are language modeling and next-word prediction. The same models will be applied to both tasks. The models we used were not trained specifically on temporal-aware data. Each model varies by selecting different types of text to prepend the text to be used for testing (or not prepending any text in one case seen in Section 4.2). From hereon, we refer to every piece of text the author has written as a post, regardless of whether it was a post or comment.

4.1 Models

Due to the lack of location-based information on posts, and therefore timezone information, each model is tailored toward a time segment for a given author, instead of similar time segments across authors. This is to capture that the timestamp for the text is relative to the author's time of day. For example, morning in one location could be night in another location, but the existing information with the text does not provide enough location-based context to determine the region-specific time of day. Therefore, we can't have a model tailored towards text from a specific time segment from multiple authors. We used GPT2 (Radford et al., 2019) with a 1024 token limit from Huggingface [3] as our general pretrained language model and its corresponding tokenizer.

Each of the following models use GPT2 and we discuss the variations of the models in the following subsections. Each model will be given text to evaluate their language modeling capabilities and next-word prediction.

4.2 *Nothing*

In this model, we use GPT2 without any modifications. Specifically, we pass each post on its own to GPT2. We refer to this model as *Nothing* in the later sections.

4.3 *Random*

In this model, each given post is prepended with text selected randomly from any of the other authors in our dataset and any relative time segment. We refer to this model as *Random* in the later sections. For example, if the test text was from author A in time segment 7, then the *Random* text prepended to that could be from any time segment (1-8) and any author that isn't author A.

4.4 *Author*

In this model, each given post is prepended with text from the same author and different time segments to the post that was being evaluated. For example if the test text is from author A in time segment 7, then the *Author* text prepended to it would be from author A and any available time segment other than 7. We refer to this model as *Author* in the later sections.

4.5 *Temporal*

In this model, each post is prepended with text from the same author and the same time segment. The text can be from different days, but it is the same daily time segment relative to the given author. For example, if the test text is from author A and time segment 7, *Temporal* text prepended to it would include all other posts from time segment 7 for

[3] https://huggingface.co/docs/transformers/en/model_doc/gpt2#openai-gpt2

author A. We refer to this model as *Temporal* in the later sections.

We chose these experimental setups to explore the influence of time in the models. For example, if someone were to write in the same manner regardless of the time of day, we would expect the improvement in performance to be similar between *Author* and *Temporal*. However, if *Temporal* outperforms *Author*, then we would expect that it is due to same person changing their manner of writing throughout the day. *Random* was chosen so that we could test if simply adding more prepended text could help as much as the personalized models. Since this dataset was limited and token length can vary within time segments of an author, we recorded the length of text used for the *Temporal* model and limited the amount of text for *Author* and *Random* to have the same token length so that results would not be affected by one model simply having more text to benefit from.

5 Results

In this section, we discuss our results from the experiment.

5.1 Evaluation Metrics

The performance of Each of the models were measured in terms of perplexity per post and accuracy at k, with k ranging from 1 through 5. The Perplexity score per post was averaged to be one value for each author/time segment combination. The accuracy at k for every post was summed and divided by the sum of all tokens in a time segment to be the percentage of tokens accurately predicted in that time segment. These values were then compared across all four of the models for each time segment.

For every time segment for every author, each post had a perplexity value and accuracy@k — the number of tokens accurately predicted within the top k predictions – for each value 1-5. Accuracy@k lends itself as a more extrinsic evaluation and resembles the desired performance on systems with next-word prediction, such as messaging applications. For example, the application recommends k words as a prediction when writing text.

Both the values for perplexity and accuracy@k from our models were then compared against all our other models to see how frequently one model produced a better value (lower perplexity or greater accuracy@k) than the model it was being compared against.

5.2 Evaluation Results

Table 1 shows each model's weighted average perplexity and accuracy@k. The weighted average incorporates the number of tokens in each post. The weighting was done so that the impact of perplexity scores was relative to the length of a post, for example, a short post could dominate the overall score if it performed exceptionally poorly on it. *Temporal* consistently outperforms all other considered models on both perplexity and accuracy@k. Although showing relatively large improvements over *Author* for perplexity, the difference with respect to accuracy@k is marginal. *Author* consistently outperformed *Nothing* and *Random*, which shows that personalizing models is beneficial.

To directly compare between models, we calculate the percentage of instances that one model outperforms another model in Table 2. The values in the table represent the percentage of all instances that one model (row) beat the other model (column) in regards to average perplexity. This further shows that the temporal-aware model, *Temporal* more often outperforms the other temporal-agnostic models with respect to perplexity.

Similarly, Table 3 shows the percentage of time segments that one model outperformed another model on next-word prediction with respect to accuracy@1. Table 4 and Table 5 also show next-word prediction results for a model comparison at k being equal to 3 and 5, respectively. The values in these table show that *Temporal* is more likely to outperform the other temporal-agnostic models than be outperformed by them. Interestingly, *Author* outperforms *Nothing* more often than *Temporal* outperforms *Nothing*.

Random outperforming *Nothing* on most instances shows that more text typically does help a model improve accuracy in next-word prediction and lowers perplexity. However, seeing that both *Temporal* and *Author* beat *Random* a majority of the time supports the idea that performance can be improved with human context, simply giving the model more text doesn't have as significant of an impact as targeting that text to represent the human who it is testing on. Author-specific text is better than random posts, but if you can get time segment info, that does tend to perform better, supporting our hypothesis.

Our results show that the type of text provided to the model can influence the performance, which is demonstrated by *Author* and *Temporal* outperform-

	Nothing	Random	Author	Temporal
Perplexity	198.21	180.61	160.31	**129.14**
% Acc@1	0.25	0.27	0.29	**0.30**
% Acc@2	0.34	0.36	0.38	**0.39**
% Acc@3	0.39	0.41	0.44	**0.45**
% Acc@4	0.43	0.45	0.47	**0.48**
% Acc@5	0.45	0.48	0.50	**0.51**

Table 1: The overall weighted average perplexity (lower is better) and accuracy@k (higher is better).

	Nothing	Random	Author	Temporal
Nothing		17.76	8.23	7.54
Random	82.24		13.59	12.4
Author	91.77	82.14		37.8
Temporal	92.46	83.13	57.94	

Table 2: The percentage of time that one model (row) outperformed the other (column) with regards to perplexity.

	Nothing	Random	Author	Temporal
Nothing		14.88	4.27	4.76
Random	84.92		9.62	6.75
Author	95.44	83.13		34.42
Temporal	95.04	85.62	55.06	

Table 4: The percentage of time that one model (row) outperformed the other (column) with regards to accuracy@3.

	Nothing	Random	Author	Temporal
Nothing		12.3	3.97	4.56
Random	87.5		8.33	6.25
Author	95.83	83.83		33.43
Temporal	95.24	86.21	55.95	

Table 5: The percentage of time that one model (row) outperformed the other (column) with regards to accuracy@5.

ing *Random*. This finding supports existing work regarding personalized models. Lastly, the manner in which a person uses a language during different times of day is potentially captured by our model and presented as *Temporal* outperforming *Author*.

6 Conclusions

At the root of human nature is communication, and much of that exists in the context that language is used. In this work, we focused on the time segment that the text was written in for our context. We examined if the usage of language for an individual author changes at different times of day. To explore this phenomena, we compared the performance of different models with respect to language modeling (perplexity) and next-word prediction (accuracy@k). Our *Temporal* model outperformed the other temporal-agnostic models on both perplexity and accuracy@k. This demonstrates the difference in the usage of language for an individual author at different times of day (time segments). Furthermore, in a direct comparison, *Temporal* outperformed the other temporal-agnostic models more than half the time for all accuracy@k and perplexity on each segment. Our results also reinforce the benefit of personalized models, since both the personalized temporal-aware and temporal-agnostic models outperform both non-personalized models. An additional benefit to the use of prepending text for our models is that it is not relatively computationally expensive and it does not require high-end GPUs as our experiments were conducted primarily on a computer with modest hardware.

	Nothing	Random	Author	Temporal
Nothing		25.50	5.26	5.85
Random	74.11		8.73	9.03
Author	94.44	82.34		38.19
Temporal	93.85	83.33	51.49	

Table 3: The percentage of time that one model (row) outperformed the other (column) with regards to accuracy@1.

Unfortunately, the dataset could not access the user's location data, and therefore we could not compare time segments across authors as we are unable to know what time it is relevant to their timezone, which could allow us to explore if people use language differently at different times of day regardless of the specific author. For example, exploring if people generally use language differently in their morning compared to their evening. This would be a reasonable and interesting direction for future work.

7 Ethical Considerations

All text was acquired through Reddit's public API and anything posted on private subreddits was not included in this research.

8 Acknowledgements

This research was supported by the Alley Heaps Undergraduate Research Internship through St. Francis Xavier University.

References

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

Maël Fabien, Esau Villatoro-Tello, Petr Motlicek, and Shantipriya Parida. 2020. BertAA : BERT fine-tuning for authorship attribution. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 127–137, Indian Institute of Technology Patna, Patna, India. NLP Association of India (NLPAI).

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

Xiaolei Huang and Michael J. Paul. 2018. Examining temporality in document classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 694–699, Melbourne, Australia. Association for Computational Linguistics.

Milton King and Paul Cook. 2020. Evaluating approaches to personalizing language models. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2461–2469, Marseille, France. European Language Resources Association.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Guy D. Rosin, Ido Guy, and Kira Radinsky. 2021. Time masking for temporal language models. *CoRR*, abs/2110.06366.

Guy D. Rosin and Kira Radinsky. 2022. Temporal attention for language models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1498–1508, Seattle, United States. Association for Computational Linguistics.

Paul Röttger and Janet Pierrehumbert. 2021. Temporal adaptation of BERT and performance on downstream document classification: Insights from social media. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2400–2412, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nikita Soni, Matthew Matero, Niranjan Balasubramanian, and H. Andrew Schwartz. 2022. Human language modeling. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 622–636, Dublin, Ireland. Association for Computational Linguistics.

Yuchen Wei and Milton King. 2024. Sense of the day: Short timeframe temporal-aware word sense disambiguation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14676–14686, Torino, Italia. ELRA and ICCL.

Identifying Severity of Depression in Forum Posts using Zero-Shot Classifier and DistilBERT Model

Zafar Sarif[1], Sannidhya Das[2], Abhishek Das[1], Md Fahin Parvej[3], Dipankar Das[3]
[1]Aliah University, Kolkata
[2]St. Xavier's College, Kolkata
[3]Jadavpur University, Kolkata
zsarifau@gmail.com

Abstract

This paper presents our approach to the RANLP 2025 Shared Task on "Identification of the Severity of Depression in Forum Posts." The objective of the task is to classify user-generated posts into one of four severity levels of depression: subthreshold, mild, moderate, or severe. A key challenge in the task was the absence of annotated training data. To address this, we employed a two-stage pipeline: first, we used zero-shot classification with *facebook/bart-large-mnli* to generate pseudo-labels for the unlabeled training set. Next, we fine-tuned a DistilBERT model on the pseudo-labeled data for multi-class classification. Our system achieved an internal accuracy of 0.92 on the pseudo-labeled test set and an accuracy of 0.289 on the official blind evaluation set. These results demonstrate the feasibility of leveraging zero-shot learning and weak supervision for mental health classification tasks, even in the absence of gold-standard annotations.

1 Introduction

In the 21st Century, mental health has become a pressing global concern, with depression identified as one of the most prevalent and disabling mental disorders. The World Health Organization (WHO) estimates that more than 280 million people globally suffer from depression, with significant impacts on quality of life, social functioning, and productivity (WHO, 2023). Depression can affect individuals of all ages and backgrounds, with symptoms often developing as early as childhood or adolescence. Left untreated, depression can lead to severe consequences, including self-harm and suicide (Friedrich, 2017). Despite growing awareness, stigma around mental illness persists, discouraging many individuals from seeking professional help in a timely manner (Clement et al., 2015).

Depression is not a binary condition but exists on a continuum of severity. It is typically categorized into subthreshold, mild, moderate, and severe levels, each requiring different clinical interventions (American Psychiatric Association, 2013). Diagnosis traditionally involves psychological evaluations conducted by trained professionals through interviews or standardized questionnaires. However, the subjective nature of symptoms, reluctance to disclose emotional distress, and limited access to mental health services—especially in low-resource settings—often delay diagnosis and treatment (Patel et al., 2018).

With the rapid growth of online platforms, individuals increasingly turn to anonymous forums for sharing mental health concerns. These platforms have become safe spaces for seeking peer and expert support. However, the sheer volume of posts makes manual monitoring by clinicians or moderators infeasible. To address this challenge, Machine Learning (ML), Deep Learning (DL) and Natural Language Processing (NLP) techniques offer promising solutions. By automatically analyzing linguistic patterns in user posts, ML and DL models can assess emotional states and estimate the severity of depression, enabling timely intervention and resource prioritization (Chancellor et al., 2019).

We have participated in the RANLP 2025 shared task on "Identification of the Severity of Depression in Forum Posts." Our findings of the task are presented through this paper. The objective of the task is to classify forum posts into four severity levels: subthreshold depression (*label 0*), mild (*label 1*), moderate (*label 2*), and severe depression (*label 3*). A notable challenge is the absence of annotated training data. To overcome this, we first employed zero-shot classification

Proceedings of the First Workshop NLP and Language Models for Digital Humanities associated with RANLP 2025, pages 126–132, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-106-6-012

using the *facebook/bart-large-mnli* [1] model to generate pseudo-labels. This model is a version of the BART model (Lewis et al., 2019) that has been fine-tuned on the MultiNLI (MNLI) database. The generated pseudo-labels are then used to fine-tune a DistilBERT [2] model, optimized for multi-class classification. Finally, we have evaluated our system on the organizers' evaluation data set. Our submission achieved a decent accuracy, demonstrating the potential of semi-supervised approaches in mental health NLP tasks.

2 Related Work

The automatic detection and classification of depression severity using digital platforms has garnered substantial attention in recent years. Traditional clinical diagnosis relies heavily on self-report questionnaires such as the Hamilton Depression Rating scale (HAM-D) and the Patient Health Questionnaire (PHQ-9), which, despite their clinical validity, are limited by subjectivity and inaccessibility for real-time monitoring (Nease et al., 2002). Consequently, researchers have explored computational methods leveraging textual, vocal, and neurophysiological signals to identify not only the presence of depression but also its severity.

Multimodal approaches have proven particularly effective in assessing depression. The authors (Stepanov et al., 2018) employed speech, facial expression, and linguistic features to predict PHQ-8 scores, concluding that behavioral cues from speech were the most reliable predictors of depression severity, surpassing visual and linguistic signals in accuracy. Similarly, an article (Dibeklioglu et al., 2017) demonstrated that combining facial and head movement dynamics with vocal prosody using autoencoders achieved robust severity classification, particularly for moderate and severe depression categories.

Language-based detection methods have also shown promise. The authors (Kabir et al., 2023) proposed a clinically inspired framework (DepTweet) to label Twitter posts using DSM-5 and PHQ-9 criteria. Their annotated dataset of over 40,000 tweets allowed for the training of models such as BERT and DistilBERT to classify posts across multiple severity levels, highlighting the potential of social media text in real-world depression assessment.

Neurophysiological research has added another dimension to severity estimation. Study in a paper (Liu et al., 2023) identified neurobiological correlates of depression severity in first-episode major depressive disorder using gamma-band EEG responses. They observed that 40 Hz and 60 Hz Auditory Steady State Responses (ASSRs) significantly correlated with clinical severity, suggesting these measures as potential diagnostic biomarkers. Similarly, another work (Mahato et al., 2020) demonstrated the effectiveness of EEG-derived features like wavelet energy and asymmetry measures in both detection and severity scaling of depression, achieving high classification accuracy using Support Vector Machine (SVM) based models.

While many prior works have relied on clinically labeled data or multimodal inputs, challenges remain in applying such techniques to forum or social media text, where class labels are often unavailable. To address this, zero-shot learning has emerged as a viable solution. In the absence of labeled data, zero-shot classification models such as *facebook/bart-large-mnli* have been utilized to generate pseudo-labels, which are then used to fine-tune more efficient downstream models like DistilBERT. This two-stage approach forms the foundation of our methodology, enabling the construction of supervised models from unlabeled depression forum datasets.

Collectively, these studies underline the effectiveness of machine learning models in detecting depressive symptoms and classifying their severity across various modalities and data sources. Our work contributes to this growing field by applying zero-shot learning and transformer fine-tuning techniques in a low-resource, text-only setting aligned with real-world use cases.

3 Datasets

The dataset for the RANLP 2025 Shared Task on 'Identification of the Severity of Depression in Forum Posts' consists of two main components: an unlabeled training dataset and an evaluation dataset intended for system testing. Both datasets contain user-generated content collected from mental health-related online forums.

[1] https://huggingface.co/facebook/bart-large-mnli

[2] https://huggingface.co/distilbert/distilbert-base-uncased

	user	text	predicted_severity
4516	SteveLC777	The first step to getting help. I have been o...	subthreshold depression
4517	Dagebow	Slip sliding away. Hi all. New to the forum ...	mild depression
4518	Easy_D	Slip sliding away. So I'm writing this to giv...	subthreshold depression
4519	thadeedz	My depression situation - anyone out there wit...	severe depression
4520	ccdep71	Lost. I have been suffering depression and an...	severe depression

Figure 1: Training dataset after generating pseudo-labels by using zero-shot classification.

The training data, distributed as a tab-separated file titled *Training_data_D-severity.tsv*, comprises 4536 entries. Each entry includes three textual fields: a user identifier (*user*), a short post heading (*title*), and a longer description (*question*). To prepare the data for model training, we merged the *'title'* and *'question'* fields into a single input string *'text'*. This dataset does not include any severity labels, which required us to generate pseudo-labels through zero-shot classification (Figure 1). The data reflects typical user-generated texts from mental health forums, featuring informal language, emotional expressions, and variability in length and structure.

The evaluation dataset, provided in a CSV file titled *evaluation_textonly.csv*, contains 5189 user posts. Each entry consists of a single column (*text*), representing the full body of a forum post. Unlike the training data, the evaluation set does not include any identifiers or structured fields such as title or question, and the gold labels indicating severity levels are not released to participants. Instead, predictions submitted by participants are scored externally by the organizers against the hidden gold labels. This setup reflects a real-world application scenario, where systems are expected to classify the severity of depression in anonymous, unstructured forum posts without access to contextual or metadata cues.

4 Methodology

Our approach to the shared task involves two-major phases - (1) generating pseudo-labels for the unlabeled training data using zero-shot classification, and (2) training a supervised text classification model using a fine-tuned transformer architecture on the pseudo-labeled data (Figure 2).

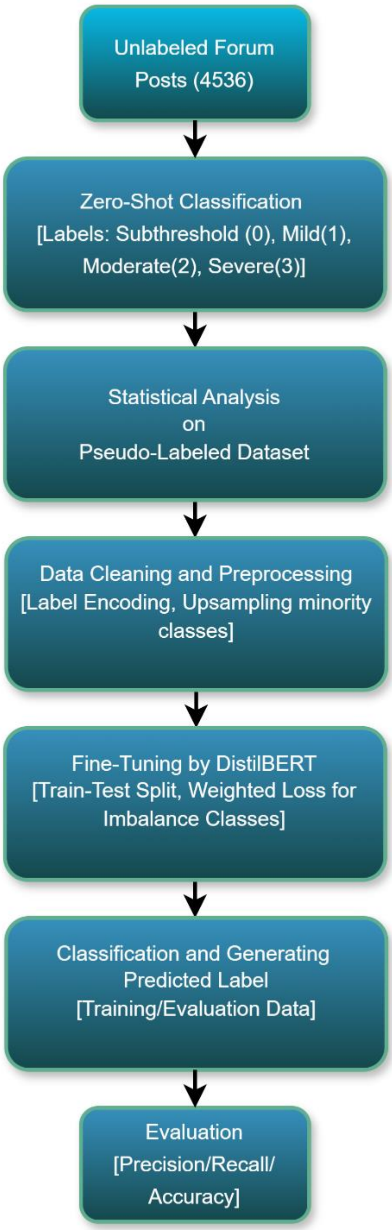


Figure 2: Pipeline of Proposed Approach.

4.1 Pseudo-label Generation

The training data was provided without severity labels. To overcome this limitation, we employed the *facebook/bart-large-mnli* model from Hugging Face Transformers for zero-shot classification (Yin et. al., 2019; Schopf et. al., 2022). For each forum post, the title and question fields were concatenated to form a complete input. We defined four candidate labels — subthreshold depression *(label 0)*, mild depression *(label 1)*, moderate depression *(label 2)*, and severe depression *(label 3)* — and allowed the model to assign the most probable label to each post based on natural language inference. The predictions were stored and used to create a pseudo-labeled dataset.

4.2 Statistical Analysis of Data

After label generation, we have conducted some statistical analysis on data to understand the behavior and relationships among the variables.

From the label distribution graph (Figure 3), it's clearly understandable that the data has imbalance and for model fitting we need to apply some data-imbalance handling techniques. The dataset contains 63.16% severe cases, which means most of the users have a serious need for some consultation and help to deal with their problems. Severe depression is majority and subthreshold is minority class here.

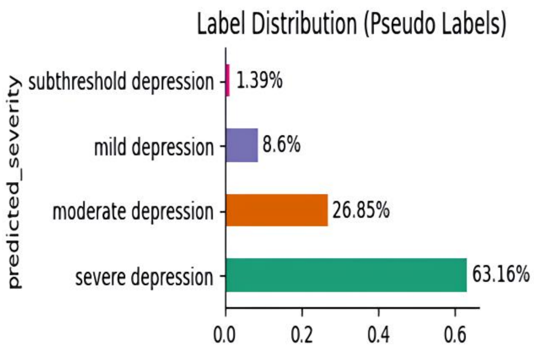


Figure 3: Label Distribution (Training Data).

A distribution of text lengths with various class shows that most posts fall between 100–400 words (Figure 4). A steep drop-off after 500 words is observed, but some posts go up to 1900+ words also (outliers). This long tail implies that a small number of users write significantly longer posts. That's why, text with ~512 tokens are considered during preprocessing to avoid undue influence from outliers and ensure model stability.

Another important observation from the boxplot (Figure 5) is that - median text lengths are longer for severe and mild categories than for moderate and subthreshold categories. The greatest number of extreme outliers falls into the serious

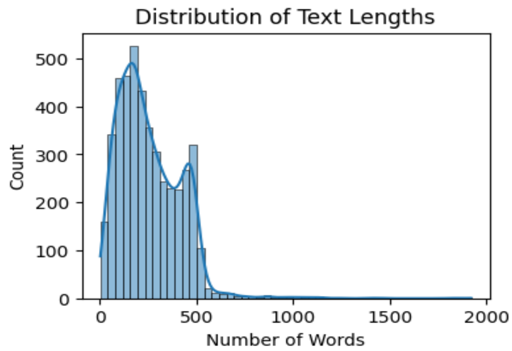


Figure 4: Distribution Plot.

group, indicating that those who suffer from severe depression might typically make longer messages. The lengths of subthreshold entries are often shorter and less varied. The degree of severity and verbosity may be related. More expressive or in-depth narratives may be linked to higher severity. To deal with this doubt we apply Chi-Squared test and got p-value = 4.42e-22 < 0.05, which clearly indicates that the earlier observation that verbosity varies across severity levels and is not due to chance.

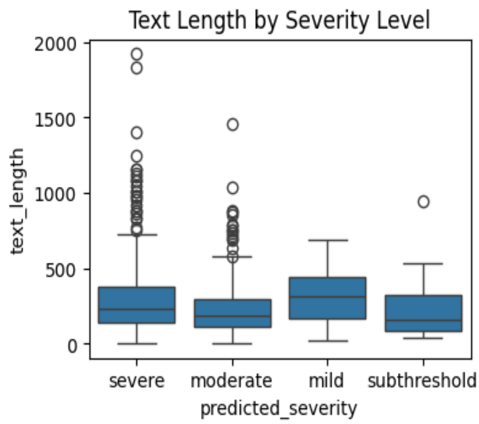


Figure 5: Box Plot.

4.3 Data Preprocessing and Resampling

After zero-shot labelling, we cleaned the dataset by removing rows with missing values and incorrect predictions. A *LabelEncoder* was applied to map the textual labels to numeric values (0–3). To address label imbalance (Figure 3) — where the majority class was 'moderate depression' — we applied random up-sampling (Gosain and Sardana, 2017; Chai et. al., 2025) to the minority classes to improve class representation in training.

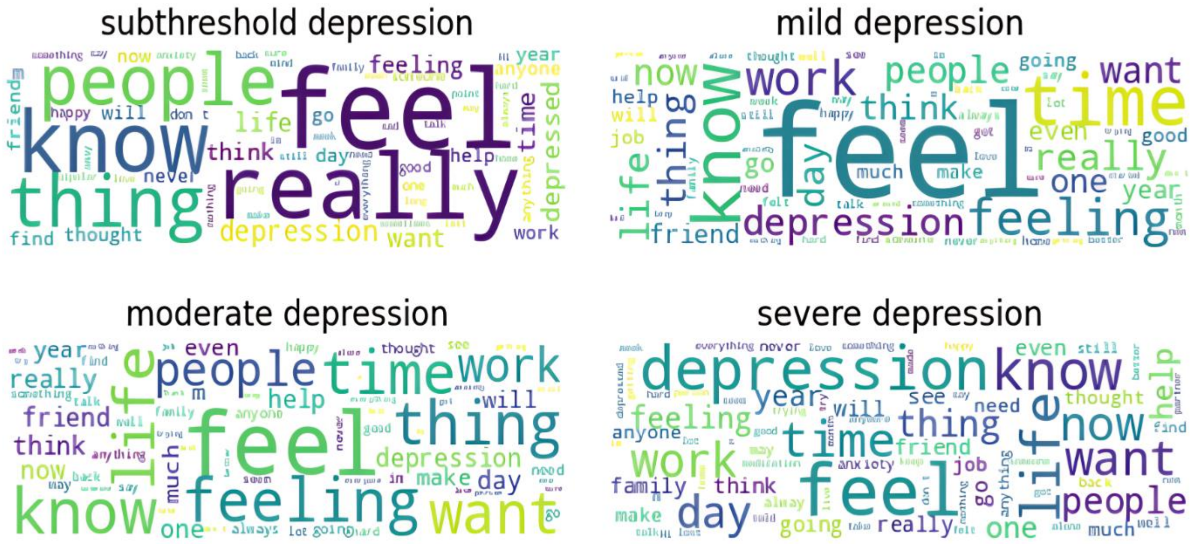


Figure 6: Few word clouds for word-level understanding of classes.

4.4 Fine-Tuning with DistilBERT

For the classification model, we fine-tuned the *distilbert-base-uncased* transformer (Sanh et. al., 2019) using the Hugging Face Trainer API. The data was split into training (75%), validation (15%), and test (10%) sets. Tokenization was performed with a maximum sequence length of 128 tokens. We also computed class weights to balance the loss function during training and implemented a custom DistilBERT subclass using a weighted *CrossEntropyLoss* function. Training was conducted for 10 epochs with early stopping (patience = 2). We used mixed-precision training (fp16=True) to optimize performance and avoid GPU memory overflow. Evaluation was conducted at the end of each epoch using weighted precision, recall, F1 score, and accuracy as metrics.

5 Result Analysis

This section presents a detailed analysis of our model's performance, covering both the internal evaluation on pseudo-labeled data and the external assessment on the official shared task evaluation set. The results illustrate the strengths and limitations of using zero-shot pseudo-labeling combined with fine-tuned transformer models for depression severity classification.

As mentioned, the datasets (training and evaluation) provided for the task are fully unlabeled, which makes the task complicated one. The pseudo labels generated by zero-shot classification approach are the base of the training our model and which is not fully proof. A manual checking is done by us for a subsample of training dataset and also for the predicted results on the evaluating dataset. Though zero-shot performs well, there are several factors such as genuineness, biasness, acceptability of forum data etc. which have affected the labelling and classification. Using forum data in research needs to proper ethical guidelines and here for our task the organizers have taken care of the same too. Ultimately these have hampered the overall accuracy or performance of our model.

5.1 Performance Metric

To evaluate the performance of our depression severity classification model, we employed multiple standard classification metrics including accuracy, precision, recall, and F1 score. For the shared task submission, overall accuracy was used as the official scoring metric, calculated by comparing predicted labels with gold-standard labels on the hidden evaluation set.

5.2 Result on Training Dataset

To validate our training pipeline, we performed an internal evaluation using a held-out test set comprising 10% of the pseudo-labeled training data. The training-validation-test split was 75%-15%-10%, with class labels derived from a zero-shot classifier. We have tried various Machine Learning algorithms like Random Forest, Support Vector Machine, XGBoost etc. but transformer model DistilBERT has outperformed all Machine Learning (ML) algorithms.

	precision	recall	f1-score	support
mild depression	0.97	0.99	0.98	289
moderate depression	0.88	0.93	0.91	289
severe depression	0.91	0.83	0.87	361
subthreshold depression	0.92	0.94	0.93	289
accuracy			0.92	1228
macro avg	0.92	0.92	0.92	1228
weighted avg	0.92	0.92	0.92	1228

Figure 7: Classification Report on Training Dataset.

The model, fine-tuned on this data with class-balanced loss, demonstrated strong predictive performance. We have achieved accuracy 0.92 (Figure 7). The confusion matrix (Figure 8) provides insight into per-class behavior. The model performed most reliably on the 'moderate' and 'severe' depression classes, which were better represented in the pseudo-labeled training data. Confusions were observed primarily between 'subthreshold' and 'mild' categories, suggesting overlap in their linguistic patterns. Despite the noisy supervision, the model learned useful class-discriminative features.

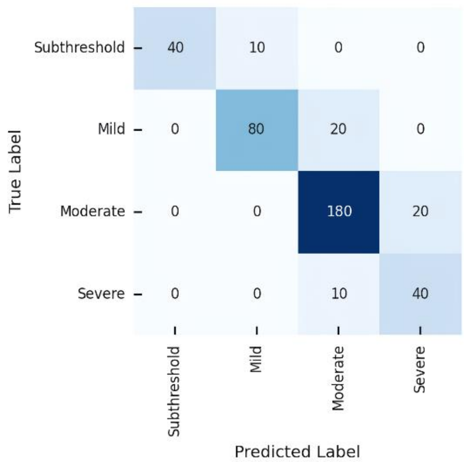


Figure 8: Confusion Matrix.

5.3 Result on Evaluation Dataset

For the official evaluation, we applied the trained model to 5189 unseen forum posts provided in the *evaluation_textonly.csv* file. The true labels were hidden by the organizers, and predictions were evaluated externally upon submission. We have submitted our prediction in *test.predictions* file with our model's predicted output. These predictions are finally evaluated with the original label or gold label by the organisers. Our model achieved final accuracy 0.289.

While this result is significantly lower than internal test accuracy, it is consistent with expectations for weakly supervised learning. The discrepancy highlights two main issues: first, label noise in the training set due to the reliance on zero-shot classification; second, a possible distributional shift between pseudo-labeled and official test data. Despite these limitations, the result establishes a baseline for depression severity classification using a fully unsupervised training pipeline.

6 Conclusion and Future Work

In this paper, we presented a two-stage system for classifying the severity of depression in forum posts. Our approach was designed to address the lack of annotated training data by first generating pseudo-labels through zero-shot classification using *facebook/bart-large-mnli*. These pseudo-labels were then used to fine-tune a DistilBERT model for multi-class classification of depression severity. Despite the absence of gold-labelled training data, our system achieved a respectable internal accuracy of 92.1% on the pseudo-labelled test set. When evaluated on the official blind test data provided by the organizers, the model reached an accuracy of 0.289. These results highlight both the potential and the limitations of using zero-shot learning and weak supervision in sensitive tasks such as mental health assessment. The low score on the official evaluation set is primarily attributed to the inherent noise in the pseudo-labels generated by the zero-shot model, as well as possible distributional differences between the training and test sets. Nonetheless, the pipeline demonstrates that transformer-based models can be trained effectively even in severely label-scarce environments when combined with smart initialization strategies.

For future work, we plan to explore the integration of semi-supervised learning techniques, such as self-training or consistency regularization, to improve label reliability and model robustness. We also aim to incorporate uncertainty estimation to identify and filter out low-confidence predictions during pseudo-label generation. Additionally, extending the model to leverage user history or temporal context could provide more nuanced understanding of depressive states. Finally, we propose to evaluate our approach on multilingual or cross-platform datasets to assess generalizability across diverse user communities and linguistic expressions of distress.

References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (5th ed.)*. Washington, DC: Author.

Chancellor, S., & De Choudhury, M. (2019). Methods in predictive techniques for mental health status on social media: A critical review. *NPJ Digital Medicine*, 2(1), 43.

Clement, S., Schauman, O., Graham, T., Maggioni, F., Evans-Lacko, S., Bezborodovs, N., ... & Thornicroft, G. (2015). What is the impact of mental health-related stigma on help-seeking? A systematic review of quantitative and qualitative studies. *Psychological Medicine*, 45(1), 11-27.

Friedrich, M. J. (2017). Depression is the leading cause of disability around the world. *JAMA*, 317(15), 1517.

Patel, V., Saxena, S., Lund, C., Thornicroft, G., Baingana, F., Bolton, P., ... & UnÜtzer, J. (2018). The Lancet Commission on global mental health and sustainable development. *The Lancet*, 392(10157), 1553-1598.

World Health Organization (WHO). (2023). *Depression*. https://www.who.int/news-room/fact-sheets/detail/depression

Stepanov, E. A., Lathuilière, S., Chowdhury, S. A., Ghosh, A., Vieriu, R. L., Sebe, N., & Riccardi, G. (2018). Depression severity estimation from multiple modalities. *2018 IEEE International Conference on e-Health Networking, Applications and Services (Healthcom)*.

Dibeklioglu, H., Hammal, Z., & Cohn, J. F. (2017). Dynamic multimodal measurement of depression severity using deep autoencoding. *IEEE Journal of Biomedical and Health Informatics*.

Kabir, M., Ahmed, T., Hasan, M. B., Laskar, M. T. R., Joarder, T. K., Mahmud, H., & Hasan, K. (2023). DEPTWEET: A typology for social media texts to detect depression severities. *Computers in Human Behavior*, 139, 107503. https://doi.org/10.1016/j.chb.2022.107503:contentReference[oaicite:9]{index=9}

Liu, S., Liu, X., Chen, S., Su, F., Zhang, B., Ke, Y., Li, J., & Ming, D. (2023). Neurophysiological markers of depression detection and severity prediction in first-episode major depressive disorder. *Journal of Affective Disorders*, 331, 8–16. https://doi.org/10.1016/j.jad.2023.03.038:contentReference[oaicite:10]{index=10}

Mahato, S., Goyal, N., Ram, D., & Paul, S. (2020). Detection of depression and scaling of severity using six-channel EEG data. *Journal of Medical Systems*, 44, 118. https://doi.org/10.1007/s10916-020-01573 y:contentReference[oaicite:11]{index=11}

Nease, D. E., Klinkman, M. S., & Volk, R. J. (2002). Improved detection of depression in primary care through severity detection. *The Journal of Family Practice*, 51(12), 1065–1070. https://www.researchgate.net/publication/10940294:contentReference[oaicite:13]{index=13}

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Yin, W., Hay, J., & Roth, D. (2019). Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *arXiv preprint arXiv:1909.00161*.

Schopf, T., Braun, D., & Matthes, F. (2022, December). Evaluating unsupervised text classification: zero-shot and similarity-based approaches. In *Proceedings of the 2022 6th International Conference on Natural Language Processing and Information Retrieval* (pp. 6-15).

Gosain, A., & Sardana, S. (2017, September). Handling class imbalance problem using oversampling techniques: A review. In *2017 international conference on advances in computing, communications and informatics (ICACCI)* (pp. 79-85). IEEE.

Chai, Y., Xie, H., & Qin, J. S. (2025). Text Data Augmentation for Large Language Models: A Comprehensive Survey of Methods, Challenges, and Opportunities. *arXiv preprint arXiv:2501.18845*.

Recall Them All: Long List Generation from Long Novels

Sneha Singhania
MPI for Informatics
ssinghan@mpi-inf.mpg.de

Simon Razniewski
ScaDS.AI & TUD
simon.razniewski@tu-dresden.de

Gerhard Weikum
MPI for Informatics
weikum@mpi-inf.mpg.de

Abstract

Language models can generate lists of salient literary characters for specific relations but struggle with long, complete lists spanning entire novels. This paper studies the non-standard setting of extracting complete entity lists from full-length books, such as identifying all 50+ friends of Harry Potter across the 7-volume book series. We construct a benchmark dataset with meticulously compiled ground-truth, posing it as a challenge for the research community. We present a first-cut method to tackle this task, based on RAG with LLMs. Our method introduces the novel contribution of harnessing IR-style pseudo-relevance feedback for effective passage retrieval from literary texts. Experimental results show that our approach clearly outperforms both LLM-only and standard RAG baselines, achieving higher recall while maintaining acceptable precision.

1 Introduction

Motivation and Problem. Analyzing literary texts often involves entity markup and the extraction of relations between characters (Piper et al., 2021; Bamman et al., 2024). For example, to discover narrative patterns in contemporary or historical fantasy stories, a tool should track character movements across locations and label them by role or sentiment (Wilkens et al., 2024). Similarly, cultural studies on gender roles in fiction across different epochs and regions (Kejriwal and Nagaraj, 2024) need labeling of character types and relationships.

To this end, tools for NER/NED and relational IE (RE for short) must be adapted to the specifics of literary language and narrative structure. In this paper, we focus on the task of RE: identifying subject-predicate-object (SPO) triples in fictional narratives, where S and O are named entities, and P is a binary relation such as parent, family, friend, or opponent.

There is ample work on RE, based on deep neural networks (Han et al., 2020; Zhao et al., 2024). Recent methods employ LLMs for encoding input texts (Josifoski et al., 2022; Ma et al., 2023; Xu et al., 2024). These models are sequence-to-sequence taggers: given text T and target subject S, they identify candidate objects O appearing in T, tag cue words for relation P, and classify each SPO candidate as valid or invalid. The key limitation is that texts are short—often single paragraphs, commonly from Wikipedia. Thus, there are only a few O candidates, and the task reduces to classification: mapping SO candidates onto none, one, or more P.

RE methods perform well when the S and O entities are salient, the text T is short, and the language style can be learned upfront via training on Wikipedia or fine-tuning on a specific corpus. However, when the input spans an entire book, pretraining has limited value and fine-tuning is infeasible due to the lack of annotated data. Also, the desired outputs would include long-tailed O's that appear only a few times over hundreds of pages. In contrast to the $SO \rightarrow P$ approach of standard RE, we cast this underexplored task as $SP \rightarrow \{O\}$: given subject S and relation P, extract/generate a long—ideally complete—list of objects O that stand in relation P with S. This goal entails two major research questions:

RQ1: How well are LLMs performing on this challenge? How much value is added by running LLMs in RAG mode?

RQ2: How can the outcome of LLM/RAG methods be further enhanced? How can we boost recall without losing too much in precision?

To illustrate the problem, consider enemies/opponents of Michael Corleone in *The Godfather* books by Mario Puzo. Figure 1 shows book excerpts with cues about McCluskey, Sollozzo, Roth, Tommasino and Fabrizzio being in this list (which,

Proceedings of the First Workshop NLP and Language Models for Digital Humanities associated with RANLP 2025, pages 133–142, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-106-6-013

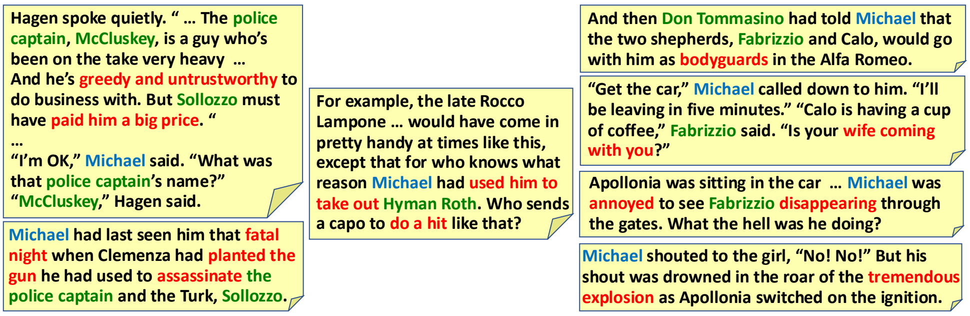


Figure 1: Example for the problem of long lists from long narratives. For the subject "Michael Corleone", we aim to extract all 40 enemies/opponents, appearing in the books.

according to sources like fan wikis, has 40 people). We observe three cases: easy (left), hard (middle), and challenging (right). The easy cases are salient entities that are frequently mentioned—extracting them needs only one or two informative passages. The hard cases arise for entities that appear infrequently (like Hyman Roth, who is a minor figure in the books); here, the issue is finding the "needle in the haystack". Finally, the most challenging cases involve vague and terse cues for the predicate, requiring deeper inference over multiple, possibly scattered, passages—such as identifying Fabrizzio as the culprit behind the car bomb attack on Michael Corleone's wife.

Approach and Contributions. We devise a novel methodology to address this challenging task. Our method, called **L3X** (**L**M-based **L**ong **L**ist e**X**traction), operates in two stages:

Stage 1: Recall-oriented Generation. An LLM is prompted with a subject and predicate from a book, to generate a long list of candidate objects by various prompts, including RAG mode, with passages retrieved from the book text. In contrast to mainstream RAG, we retrieve a large number of passages (e.g., 500 for a given SP pair) and judiciously select the best ones for prompting.

Stage 2: Precision-oriented Scrutinization. Given a high-recall list of object candidates, we devise a classifier to corroborate or prune objects.

Since we tackle an unexplored task, we construct and release a new dataset for evaluation and as a resource for the NLP community. The data comprises 11 books or book series, with 16,000 pages total. It covers 8 relations of long-tailed nature (friends, opponents etc.). To use the copyrighted texts, purchasing the e-books is required.

Salient contributions of this work are: (1) the new task of extracting a long list of objects for a given subject and relation from book-length narrative texts; (2) L3X methodology for this task, based on retrieval-augmented LLMs and combining information-retrieval (IR) techniques with LLM generation; (3) experiments with a new benchmark, showing that L3X outperforms LLM-only and LLM-RAG baselines, with an in-depth analysis of strengths and limitations of different methods. The dataset, licensing details, code and experimental results are available at https://anonymous.4open.science/r/l3x-9E4A.

2 L3X Methodology

Figure 2 gives an overview of the L3X components and the data flow between them. The first four steps form the recall-oriented stage 1, the fifth step is for the precision-oriented stage 2. The following presents the full L3X pipeline. Baselines and L3X variants are derived from specific configurations (Sec 4) and given in experimental results (Sec 5).

2.1 Passage Retrieval

Long texts, like books, are chunked into short passages of 15 sentences, totaling up to 1000 characters. We create all overlapping passages (i.e., with shared sentences) to ensure that sentences with co-references stay connected to named entities in their proximity. Since books often contain extended direct speech, which may omit explicit speaker names, we enrich each passage with *mentions of people and locations* from the *preceding* 10 passages. This metadata annotation ensures that relevant named entity information from prior chunks remains accessible within the current passage.

On the large pool of enriched passages, indexed for efficient retrieval, we select the open-sourced and effective Contriever (Izacard et al., 2022), a BERT-based dense neural IR method fine-tuned on MS-MARCO dataset[1]. The query vector is con-

[1] https://github.com/facebookresearch/contriever

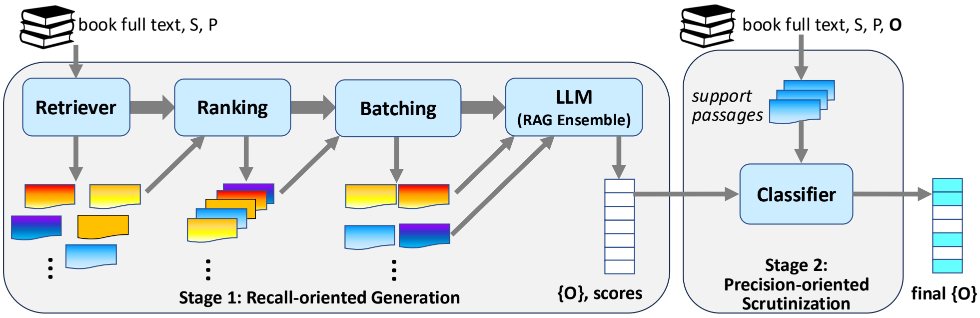


Figure 2: Overview of the L3X methodology.

structed from the SP pair; an example is: "enemies of Michael Corleone." Moreover, paraphrases of P and alias names are included, such as "opponents rivals Don Michael" for ensemble mode.

2.2 Passage Ranking

Default Ranking (def). The default passage ranking is given by the retriever scores. For a given SP pair, formulated as a natural language query, the dense retriever ranks top-d passages based on cosine similarity to the query vector.

Amplification (amp). For this novel *re-ranking* of passages, we employ the IR principle of *pseudo-relevance feedback* (Zhai, 2008). After extracting O lists from the initially selected passages (see Subsection 2.4), we assess the passage quality based on the number of distinct O's the passage yields. The best s passages are assumed to provide good cues about relation P in surface form (with hyper-parameter s). The averaged embedding vectors for these *high-yield passages* are the reference against which all retrieved passages are re-ranked. The *amp* technique works in two alternating steps and iterates them as follows:

1. For each SP pair, we consider the previously generated O values and the best s high-yield passages: those from which the LLM could extract the most objects.
2. All retrieved passages are re-ranked by the retriever's scoring model based on combining the original query (about SP) with the selected high-yield passages. The now highest-ranking passages go into the next batch of O extraction via LLM prompting, where each batch consists of b passages (e.g., $b = 4$).

For scoring, we utilize the retriever for computing cosine similarity of passages to a refined query: convex combination of the original query vector and the sum of top-s support passages' vectors: $\mathbf{E}(Q') = \alpha\mathbf{E}(Q) + (1-\alpha)\sum_{i=1}^{s}\mathbf{E}(S_i)$ with embeddings $\mathbf{E}(\,)$ and hyper-parameter α.

2.3 Passage Batching

For the high cost (or even infeasibility) of augmenting a large number of passages (e.g., all 500 retrieved ones) into an LLM, we group the passages into smaller batches of size b (a hyper-parameter; typical values being 2, 4, or 6), by default in descending ranking order. Alternatively, passages can be batched using one of two criteria:

- **Named Entity Overlap (neo):** passages with a large overlap in named entity mentions;
- **Passage Similarity (sim):** passages whose embeddings have a high cosine similarity.

For *neo*, we compute Jaccard similarity using minhash sketches of entity sets, while *sim* uses embedding vectors computed by the retriever. Both strategies process a priority queue of passages as follows: for each rank r (starting with highest, r=1), find the b-1 most related passages from lower ranks (r'>r) to form a batch and prompt the LLM. Mark all the batch passages as "done" and proceed with the next lower rank (r'>r), which is not yet "done".

2.4 Prompt-based Object-List Generation

We append passages into the prompt for RAG-based list generation. As LLMs have limits on input context (and GPU memory demands increase with prompt length), we divide the top-k passages (ranked by retriever scores) into batches of b passages each (e.g., k=20, b=4 gives 5 batches). The O values generated from batch-wise processing are combined by their union for high recall.

Prompts can be zero-shot or few-shot, with the latter including a small set of demonstration examples for in-context inference. The examples explicitly mention SP appearing in books disjoint from the dataset, along with their complete O lists. In **single-prompt** mode, the LLM uses only the best of these formulations. In **ensemble** mode, for each relation, we manually prepare five prompt templates, and repeat the LLM-based generation with all templates. The final O list is the union of the O values generated across all runs. We focus on the **few-ensemble** setting for the main results.

2.5 Classifier to Enhance Precision

In the precision-oriented scrutinization stage, we leverage the fact that, unlike in the first stage, we have lists of candidate objects. This allows us to identify the passages from which the corresponding SPO triples were extracted or where they appear.

Scoring of O Candidates. Each LLM call returns a list with a score for the entire list, no scores for individual objects. However, with batch-wise LLM calls and the ensemble with different prompts, we can derive a total score for each O candidate (for a given SP), by a weighted occurrence frequency:

$$\text{score}(O) = \sum_{\text{batch}_i} \exp\left(\text{score}_{\text{M}}(L_i)\right) \times \mathbf{I}_i(O)$$

where $\mathbf{I}_i(O)$ is an indicator variable set to 1 if O occurs in the output list L_i for the ith batch of passages, and zero otherwise. score_{M} is the LLM log probability. This can then be used for direct pruning by thresholding on scores.

Default Thresholding (thr). The simplest scrutinizing technique is to prune O candidates below a a specified cut-off point in the ranked list of per-O scores. As the score distribution is often skewed, we do not truncate by score value, but set the cut-off point to be the tth quantile of the cumulative score distribution, with the default setting t=0.8.

Support Passages as Evidence. While stage 1 starts with SP only, stage 2 has O candidates. This allows us to *search the full book* for snippets that contain cues for the entire SPO triple. For each SPO, we retrieve the top-p passages, termed *support passages*. These passages differ from the high-yield passages used by *amp* in stage 1, as they are retrieved afresh for each SPO. For retrieval, we generate passage embeddings using the retriever's text-to-vector model. The vectors are compared against embeddings of the concatenated SPO strings, including SO alias names and paraphrases of P, using cosine similarity.

Predicate Classifier (pred). The collection of support passages, for all SO with the same relation P, can be used to learn an embedding for P cues, sort of a "mini-LM" for P. The intuition is that support passages with indicative phrases, such as "life-or-death combat with", "deeply hates" or "I will destroy you" (in direct speech), can collectively encode a better signal for P. To construct the classifier, we perform the following steps:

1. For each O, retrieve top-p support passages, and encode them into embedding vectors.
2. Identify the top-ranked O values with score(O) above a threshold ω.
3. Using the top-ranked O, combine the per-O passage vectors by a weighted sum, with score(O) as weights, to obtain a single P-vector.
4. Each SO pair under scrutiny (O below the threshold ω) is tested by comparing the vector of the top-p support passages for this SPO candidate against the P-vector computed using steps 1 to 3.
5. Accept an SO pair if the cosine similarity between the embeddings is above threshold θ.

We construct a *pred* classifier for each SP pair, in a completely self-supervised manner. All classifiers share hyper-parameters ω, p and θ; these are tuned via withheld train/dev data with SPO ground-truth, but without any supervised passage labels.

3 Dataset

Extracting long O lists from long books is a new task, with no suitable datasets available. We constructed a new dataset of books and ground-truth O lists for SP pairs. We selected eleven book series[2], discussed on community websites[3]. These fan sites feature lists and infoboxes from which we derived SPO ground-truth with high confidence (with manual curation). Book length goes up to 10K passages in epic series like A Song of Ice and Fire.

Since entities often appear under multiple surface forms, we manually constructed an alias name dictionary. On a per-book basis, we ensured that certain first names, last names, or nicknames were uniquely identifiable, e.g., "Daenerys" is unique but "Targaryen" is ambiguous. So for this entity, aliases include "Dany", "Daenerys Targaryen", "Daenerys Stormborn", but not "Targaryen". LLM outputs like "Targaryen" alone are counted as false.

Our dataset comprises 764 distinct SP pairs for 8 relations. In total, it covers ca. 5,300 entities, referenced under ca. 12,000 alias names. While the S entities are prominent book characters, their associated O lists are long and dominated by rarely men-

[2] A Song of Ice and Fire Series, Godfather Series, Harry Potter Series, Outlander Series, Little Women, Malibu Rising, Pride and Prejudice, Steve Jobs, The Girl with the Dragon Tattoo, Wuthering Heights, The Void Trilogy

[3] www.cliffsnotes.com, www.bookcompanion.com, www.fandom.com

tioned, long-tail entities. To highlight the gap with standard RE, we examined the Wikidata knowledge graph (KG) for triples involving the 30 Harry Potter characters used as target S. While the KG includes most of our predicates, it lacks substance beyond metadata (e.g., featured-in-media, library IDs). It is also extremely sparse: it lists only 2 of Harry's enemies, compared to 50+ in our ground truth—a trend consistent across other subjects and relations.

Relation Difficulty. The chosen 8 predicates include 3 *easy relations* with a limited number of O values (parent, child, and sibling) and 5 *hard relations* with long O lists (family, friend, opponent, placeHasPerson (i.e., people being at a place), and hasMember (i.e., members of orgs. or events).

4 Experimental Setup

Evaluation Metrics. By the design rationale of L3X, we use different metrics for stage 1 and stage 2. For the recall-oriented stage 1, the obvious measure of interest is **Recall**: the fraction of ground-truth object (O) values correctly generated. For stage 2, neither precision nor recall alone reflect our objective, and F1 would merely be a generic compromise. Instead, we aim to achieve high recall while keeping precision at an acceptable level. Therefore, our key metric—computed from the final ranked lists— is **Recall@PrecisionX (R@Px)**, where x is the precision to be guaranteed (e.g., x being 50% or, ideally, 80%). R@Px metric reflects the need for high-coverage outputs worthwhile for downstream applications such as tool-supported literature analysis, while avoiding too many errors as these entail manual curation. For both stages, we also report *precision* values and the *precision-recall area under the curve, AUC.*

All reported numbers are *macro-averaged percentage* scores, computed in three steps. For each SP pair, we first compute the precision and recall of the generated O list against the ground-truth. These are then averaged across all SP pairs for each P. Finally, the results are averaged across all relations.

System Configurations and Baselines. The L3X methodology comes with options for components and configurations. For our main experiments, we operate in the few-ensemble prompt setting and focus on the following choices:

- **LLM-only**: directly prompting the LLM without passages (Subsection 2.4). For stage 2, we apply *thr* pruning with t=0.8.
- **RAG:** restricting L3X to the Retriever (Subsection 2.2) and LLM prompting, leads to standard RAG, with *def* passage ranking at stage 1 and *thr* (t=0.8) for stage 2.
- **L3X-amp-thr:** a configuration with *amp* for re-ranking, and *thr* (t=0.8) for pruning.
- **L3X-amp-pred:** a configuration with *amp* for re-ranking, and the *pred* classifier in two variants: *pred(g)* with globally tuned hyper-parameters and *pred(p)* with predicate-specific hyper-parameter values (see below).
- **L3X-amp-neo:** a configuration with *neo* batching and a pruning classifier (*thr* or *pred*).
- **L3X-amp-sim:** a configuration with *sim* batching and a pruning classifier (*thr* or *pred*).

Hyper-Parameters. L3X includes several tunable hyper-parameters; Optimal values are identified using withheld train/dev data. To this end, we split the entire dataset into three folds (30:20:50), via stratified sampling over books and SP pairs, ensuring equal representation of varying O-list lengths. For each S in train/dev, the complete O list is taken in the ground-truth to avoid information leakage into the test set. Hyper-parameters are tuned via grid search, maximizing the recall metric in stage 1 and R@P50 metric in stage 2. This is done in two modes: a single *global* value per hyper-parameter, or *per-predicate* values, specific to each P.

5 Results

We present the main findings on the long list generation task.

5.1 RQ1: Performance of LLMs and RAG

Table 1 reports macro-averaged results for the LLM-only setting with three widely used models (GPT-3.5[4], Llama3.1-8B, Llama3.1-70B[5]) and RAG results with the best of these (Llama3.1-70B). All are in few-shot ensemble mode, and all use *thr* (t=0.8) for stage 2.

LLM-only performance is poor, achieving less than 50% recall after stage 1, with mediocre precision. Llama-70B and GPT-3.5 perform comparably, while Llama-8B substantially lags behind. In RAG mode (with *def* ranking of passages), results improve: 84% recall after stage 1, but precision stays low even after *thr*-based scrutinization. The best R@P50 number is 40.2%. As a reference, we estimate an oracle upper bound of 88% by counting

[4]platform.openai.com/docs/models/gpt-3.5-turbo

[5]huggingface.co/meta-llama/Llama-3.1-8B, huggingface.co/meta-llama/Llama-3.1-70B

the distinct O values from ground truth that appear in at least one of the retrieved top-500 passages.

The insight here is that LLMs can recall only a fraction of O's from pre-training, and add many false positives. Equipped with book passages, the recall is improved, but false positives remain a major challenge for this very difficult task.

5.2 RQ2: Added Value of L3X Configurations

Table 2 compares different L3X configurations, contrasting them with Llama3.1-70B model in RAG mode. Adding smart re-ranking (*amp*) and batching to RAG pays off very well, and the sophisticated classifier (*pred*) also enhances scrutiny. After stage 1, the recall by L3X variants is similar to the RAG, but we observe a notable improvement in AUC, reaching 27.5%. This signals a higher concentration of true positives among the top-ranked O values—an important asset for stage 2.

The L3X *amp* method for iterative re-ranking achieves the biggest boost over the RAG baseline: moving R@P50 close to 50% and R@P80 to around 36%. Combining it with one of the two batching techniques does not add value, as *amp* by design is already judicious in picking its batches. Replacing the *thr* pruning with the sophisticated *pred* classifier further enhances the performance a bit. Again, drill-down by predicate shows higher gains for some of the hard P, indicating potential for more. The influence of hyper-parameter tuning for *pred* is discussed below.

The bottom line is that L3X *amp* adds substantial benefits over LLM-only and standard RAG methods, highlighting the crucial role of judicious passage ranking. The final R@P results—reflecting the benefit/cost ratio for downstream usage—are promising, but still fall short of being fully satisfying. This emphasizes the challenging nature of the new task explored in this paper.

Hyper-parameter Tuning for pred. The *pred* classifier has three hyper-parameters. Setting their values by global grid search with train/dev data in a self-supervised manner leads to the best results, with ω=20, p=5, and θ=0.75. As various P exhibit different characteristics, we would expect further gains with per-predicate grid search. Indeed, this led to rather different predicate-specific values, e.g., for Sibling, the best values are ω=10, p=2, θ=0.9, but for Friend we get ω=50, p=1, θ=0.55. Nevertheless, *pred(p)* did not achieve significant improvements over the globally tuned variant *pred(g)*. We attribute this to the fact that the simpler configurations are already close to the best possible outputs given the inherent difficulty of the task.

Comparison to Other Classifiers. We explored two alternative approaches for stage 2 scrutinization. First, we used the LLM itself to *elicit its own confidence* (Wang et al., 2023). For each SPO, we included the support passages along with all named entities into the prompt for in-context inference: "Given this information, is [SPO] a correct statement?". This approach performed poorly. For example, with the L3X *amp*, it has 46.6% precision, 44.7% recall, 19.0% AUC, 31.7% R@P50 and 20.4% R@P80.

Second, we evaluated *standard relational IE methods* that classify an SO pair to a given predicate P. We fine-tuned two state-of-the-art models—GenIE (Josifoski et al., 2022) and DREEAM (Ma et al., 2023)—on our train/dev folds. However, both models performed very poorly, with recall below 5% and precision no higher than 10%. This highlights the difficulty of our task: these models were trained on Wikipedia-style text, very different from long and complex fiction.

6 Discussion

6.1 Drill-Down and Sensitivity

Predicate Drill-Down. While results are macro-averaged over all relations, some predicates are easier than others (see Section 3). We analyzed performance per predicate using the best configurations after stage 1 and 2. Stage 1 recall is fairly consistent across predicates (75-90%), but stage 2 R@P numbers vary widely:"easy" relations with short, well-defined lists perform well, while "hard" relations—those with longer lists and vaguer cues—show a significant drop. As expected, Opponent is the most difficult predicate, where even our best method reaches only ca. 32% of R@P50. Full per-predicate scores after both stages is in Table 3.

Entity Popularity. We further analyzed performance by splitting ground-truth O entities in the test set into *head* and *tail* groups, based on their frequency in the book. Entities above the 75th percentile were labeled as head, the rest as tail. This results in four combinations: (easy P, head), (easy P, tail), (hard P, head), and (hard P, tail). We observe that L3X-amp consistently outperforms standard RAG across all four cases. However, in the most challenging setting—hard P with tail O—performance drops sharply.

LLM	Config	Stage 1			Stage 2 (thr, t=0.8)				
		P	R	AUC	P	R	AUC	R@P50	R@P80
GPT-3.5	LLM-only	43.6	43.9	21.3	**41.5**	31.4	17.4	25.4	20.2
Llama-8B	LLM-only	21.2	31.9	11.5	21.2	27.8	10.5	16.9	12.4
Llama-70B	LLM-only	34.1	47.7	20.5	37.0	39.5	20.5	31.4	21.9
Llama-70B	RAG	12.0	84.3	22.9	14.6	**82.8**	**22.7**	**40.2**	**26.1**

Table 1: Results for LLM-only and RAG in few-shot ensemble mode.

L3X Config	Stage 1			Stage 2				
	P	R	AUC	P	R	AUC	R@P50	R@P80
RAG-thr	12.0	84.3	22.9	14.6	82.8	22.7	40.2	26.1
amp-thr	13.7	83.6	27.5	16.0	81.0	27.4	48.6	35.9
amp-neo-thr	14.1	83.4	27.1	16.7	**81.6**	26.9	47.7	35.4
amp-sim-thr	14.1	83.4	27.1	16.0	80.5	26.3	47.0	33.8
amp-pred(p)	13.7	83.6	27.5	**23.5**	77.3	28.0	48.7	36.2
amp-pred(g)	13.7	83.6	27.5	20.4	80.4	**28.1**	**49.7**	**36.5**
amp-neo-pred(p)	14.1	83.4	27.1	22.1	76.4	27.4	48.0	35.4
amp-neo-pred(g)	14.1	83.4	27.1	19.8	79.9	27.6	48.7	35.7

Table 2: Results for L3X configurations with Llama-70B in few-shot ensemble mode.

Relation	Stage 1						Stage 2							
	RAG			L3X-amp			RAG-pred(g)				L3X-amp-pred(g)			
	P	R	AUC	P	R	AUC	P	R	AUC	R@P50	P	R	AUC	R@P50
parent	25.9	75.6	25.5	29.8	76.2	26.0	38.2	71.4	27.3	57.7	42.7	76.2	27.9	61.3
children	20.7	86.5	27.6	19.6	82.5	32.1	33.7	83.9	28.1	60.9	30.7	82.0	36.5	72.3
sibling	26.9	87.2	34.9	36.6	86.2	47.7	38.2	85.8	38.0	65.4	45.3	85.2	47.7	79.4
avg. Easy P	24.5	83.1	29.3	28.7	81.6	35.3	36.7	80.4	31.1	61.3	39.6	81.1	37.3	71.0
family	5.5	79.8	25.2	5.1	79.8	33.3	11.5	77.2	25.2	34.0	12.4	76.3	33.1	44.8
friend	7.1	85.4	20.1	7.5	85.5	24.1	11.7	80.3	19.7	27.1	11.7	80.3	23.8	35.5
opponent	4.1	80.8	17.7	4.4	81.1	18.9	7.2	73.6	17.6	29.3	8.6	74.3	18.9	32.4
hasMember	2.6	89.0	16.5	2.8	86.6	20.8	5.5	82.1	16.5	25.7	5.5	83.4	20.7	32.5
placeHasPer	3.0	89.8	15.4	3.4	90.7	16.8	5.6	82.0	14.7	30.3	6.0	85.3	16.5	39.6
avg. Hard P	4.5	85.0	19.0	4.7	84.7	22.8	8.3	79.0	18.7	29.3	8.8	79.9	22.6	37.0
avg. All P	12.0	**84.3**	22.9	13.7	83.6	27.5	18.9	79.5	23.4	41.3	**20.4**	80.4	**28.1**	**49.7**

Table 3: Drill-Down Recall Results by Predicate for Stage 1 and Stage 2.

Role of LLM's Parametric Memory. To assess the influence of pre-training, we compared LLM-only to L3X-amp on the Void Trilogy—a series with minimal Web coverage, for which we invested great effort in compiling ground truth. Results show that LLM-only fails completely on this case: 12% recall and just 5% precision, whereas L3X-amp gets 82% recall after stage 1, and 34.1% R@P50 and 38.3% R@P80 with *thr* in stage 2.

Sensitivity of Hyper-Parameters. We conducted extensive experiments to assess the sensitivity of hyper-parameters, specifically the no. of top-k retrieved passages and batch size b. We observe that increasing k improves recall, but with diminishing returns and higher LLM cost. Batch size b matters more when k is large—larger batches boost recall by providing more context, but also increases prompt length and cost. Reducing the number of retriever query reformulations hurts both recall and R@P, highlighting the value of query diversity.

6.2 Error Cases.

We observed recurring error types and discuss three of the most notable cases.

Hallucinations. LLM calls often return huge lists of O's, including names that do not occur in the respective books. Even in RAG mode, the LLM does not necessarily restrict its outputs to entities present in the input passages—a case of *unfaithful* generation. To quantify the effect, we compute the no. of generated O's that do not appear in the respective book, normalized by the total no. of generated O values. Hallucination rates after stage 1 were: LLM-only: 55.3%, RAG: 51.7%, L3X-amp: 40.7%, L3X-amp-neo: 38.1%. Hallucinations include made-up names and non-entity phrases. This underlines the importance of stage 2 scrutinization.

Confusing Predicates. LLMs generate valid O values that are related to subject S, but under the wrong relation P. A notable case is when O belongs to a different ground-truth predicate Q ($\neq P$) (e.g., Dumbledore appearing as Harry Potter's parent instead of friend), We computed the #P$\times$#P confusion matrix, counting Os generated under P when their true relation is Q. With L3X-amp-pred(g), we observed a 60:30:10 ratio: correct TPs, predicate-confused TPs, and false positives (FPs). This shows that most SO pairs are reasonable, but predicate accuracy at high recall remains a challenge.

Missing True Positives in the Low Ranks. The majority of TPs are at high ranks, followed by a long tail of mostly FPs but sprinkled with TPs at lower ranks. To assess how well stage 2 recovers *low-ranked* TPs, we use the R@P50 cut-off rank to count the missing TPs below this threshold—those misclassified as false negatives. Even with our best methods, about 16% of all the ground-truth O values fall into this low-rank, missed-TP category.

7 Related Work

Relation Extraction. A core task in IE is extracting the relation P between two entities, subject S and object O, where P comes from a predefined set of predicates. State-of-the-art methods (Han et al., 2020; Wang et al., 2020; Cabot and Navigli, 2021; Josifoski et al., 2022; Ma et al., 2023) typically operate on single passages using a multi-label classifier or sequence tagger. Recent works (Zhao et al., 2024; Xu et al., 2024) have advanced the scope of the extractors' input under the theme of "long-distance IE", extending beyond single sentences or passages. However, techniques like graph neural networks or LLM-powered generative IE are geared for short news or encyclopedic texts, and cannot cope with book-length texts. Even the popular document-level benchmarks, DocRED (Yao et al., 2019) and REBEL (Cabot and Navigli, 2021), limit inputs to single Wikipedia paragraphs.

Retrieval-Augmented Generation. LLMs excel in QA and IE tasks by drawing from their extensive parametric knowledge (pretrained over massive contents), particularly with few-shot in-context inference (Zhao et al., 2023; Minaee et al., 2024). However, they are still susceptible to hallucinations, especially for long-tail entities and facts (Ji et al., 2022). To improve the overall task accuracy, recent work has focused on integrating relevant text snippets into in-context prompts through the RAG paradigm (Lewis et al., 2020; Guu et al., 2020; Cai et al., 2022; Asai et al., 2023; Wang et al., 2023; Gao et al., 2023). However, the effectiveness of RAG crucially depends on the retriever policy, and the case of long novels has not been studied so far.

Information Extraction from Books. Prior works (Bamman et al., 2019; Stammbach et al., 2022; Chang et al., 2023) pursue LLM-supported IE about characters from fiction books. But these methods focus on NER-like generation of single names from single passages. Bamman et al. (2024) considers usage of LLMs for cultural analytics, and Piper and Bagga (2024) shows LLMs for characterizing and annotating narrative texts. None of these works addresses full-fledged RE from entire books.

8 Conclusion

We introduced the task of extracting long lists of objects from long documents, and proposed the L3X methodology, comprising LLM prompting, retrieval augmentation, passage re-ranking and batching, and classifier-based pruning. Extensive experiments demonstrate that L3X significantly outperforms baselines in both recall and R@P. Our best performing L3X configuration *amp-pred(g)*, leveraging pseudo-relevance feedback and a tuned classifier, achieves remarkable performance of ca. 85% recall and ca. 37% R@P80 on full-length books. However, drill-down analyses by relation and entity popularity reveal substantial gaps in the hard cases. This highlights the core challenge of our task: while scattered textual cues across long books may be intuitive for humans, they remain difficult for AI systems, including LLMs, to reliably detect and extract.

References

Akari Asai, Sewon Min, Zexuan Zhong, and Danqi Chen. 2023. Retrieval-based language models and applications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)*, pages 41–46, Toronto, Canada. Association for Computational Linguistics. Tutorial materials at "http://acl2023-retrieval-lm.github.io/".

David Bamman, Kent K Chang, Li Lucy, and Naitian Zhou. 2024. On classification with large language models in cultural analytics. *Computational Humanities Research Conference (CHR)*.

David Bamman, Sejal Popat, and Sheng Shen. 2019. An annotated dataset of literary entities. 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 2138–2144, Minneapolis, Minnesota. Association for Computational Linguistics.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. REBEL: relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 2370–2381. Association for Computational Linguistics.

Deng Cai, Yan Wang, Lemao Liu, and Shuming Shi. 2022. Recent advances in retrieval-augmented text generation. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 3417–3419. ACM. Tutorial materials at "https://jcyk.github.io/RetGenTutorial/".

Kent Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023. Speak, memory: An archaeology of books known to ChatGPT/GPT-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7312–7327, Singapore. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *CoRR*, abs/2312.10997.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Xu Han, Tianyu Gao, Yankai Lin, Hao Peng, Yaoliang Yang, Chaojun Xiao, Zhiyuan Liu, Peng Li, Jie Zhou, and Maosong Sun. 2020. More data, more relations, more context and more openness: A review and outlook for relation extraction. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, AACL/IJCNLP 2020, Suzhou, China, December 4-7, 2020*, pages 745–758. Association for Computational Linguistics.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Wenliang Dai, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55:1 – 38.

Martin Josifoski, Nicola De Cao, Maxime Peyrard, Fabio Petroni, and Robert West. 2022. GenIE: Generative information extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4626–4643, Seattle, United States. Association for Computational Linguistics.

Mayank Kejriwal and Akarsh Nagaraj. 2024. Quantifying gender disparity in pre-modern english literature using natural language processing. *Journal of Data Science*, 22(1).

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Youmi Ma, An Wang, and Naoaki Okazaki. 2023. DREEAM: Guiding attention with evidence for improving document-level relation extraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1971–1983, Dubrovnik, Croatia. Association for Computational Linguistics.

Shervin Minaee, Tomás Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *CoRR*, abs/2402.06196.

Andrew Piper and Sunyam Bagga. 2024. Using large language models for understanding narrative discourse. In *Proceedings of the 6th Workshop on Narrative Understanding*, pages 37–46.

Andrew Piper, Richard Jean So, and David Bamman. 2021. Narrative theory for computational narrative understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana,* *Dominican Republic, 7-11 November, 2021*, pages 298–311. Association for Computational Linguistics.

Dominik Stammbach, Maria Antoniak, and Elliott Ash. 2022. Heroes, villains, and victims, and GPT-3: Automated extraction of character roles without training data. In *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*, pages 47–56, Seattle, United States. Association for Computational Linguistics.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Jiayang Cheng, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, Yidong Wang, Linyi Yang, Jindong Wang, Xing Xie, Zheng Zhang, and Yue Zhang. 2023. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *CoRR*, abs/2310.07521.

Difeng Wang, Wei Hu, Ermei Cao, and Weijian Sun. 2020. Global-to-local neural networks for document-level relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 3711–3721. Association for Computational Linguistics.

Matthew Wilkens, Elizabeth F Evans, Sandeep Soni, David Bamman, and Andrew Piper. 2024. Small worlds: Measuring the mobility of characters in english-language fiction. *Journal of Computational Literary Studies*, 3(1).

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. 2024. Large language models for generative information extraction: a survey. *Frontiers Comput. Sci.*, 18(6):186357.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. Docred: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 764–777. Association for Computational Linguistics.

ChengXiang Zhai. 2008. Statistical language models for information retrieval: A critical review. *Found. Trends Inf. Retr.*, 2(3):137–213.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *CoRR*, abs/2303.18223.

Xiaoyan Zhao, Yang Deng, Min Yang, Lingzhi Wang, Rui Zhang, Hong Cheng, Wai Lam, Ying Shen, and Ruifeng Xu. 2024. A comprehensive survey on relation extraction: Recent advances and new frontiers. *ACM Comput. Surv.*, 56(11):293:1–293:39.

Exploring the Limits of Prompting LLMs with Speaker-Specific Rhetorical Fingerprints

Wassiliki Siskou[1,2] and **Annette Hautli-Janisz**[2]
[1] University of Konstanz, Germany
[2]University of Passau, Germany
wassiliki.siskou@uni-passau.de
annette.hautli-janisz@uni-passau.de

Abstract

The capabilities of Large Language Models (LLMs) to mimic written content are being tested on a wide range of tasks and settings, from persuasive essays to programming code. However, the question to what extent they are capable of mimicking human conversational monologue is less well-researched. In this study, we explore the limits of popular LLMs in impersonating content in a high-stakes legal setting, namely for the generation of the decision statement in parole suitability hearings: We distill a linguistically well-motivated rhetorical fingerprint from individual presiding commissioners, based on patterns observed in verbatim transcripts and then enhance the model prompts with those characteristics. When comparing this enhanced prompt with an underspecified prompt we show that LLMs can approximate certain rhetorical features when prompted accordingly, but are not able to fully replicate the linguistic profile of the original speakers as their own fingerprint dominates.

1 Introduction

Recent research on LLM alignment shows that depending on the task, LLMs can mimic or imitate human language to an extent that the generated content is indistinguishable from or even surpasses the quality of human language. Mimickry is an intermediate step towards impersonation, the latter assuming that an agent not only copies general human behavior, but pretends to be a specific person and acts accordingly. In this paper we show that we can nudge LLMs towards impersonation, but that there remains a gap between actual human and generated content. We do so by crafting speaker-specific rhetorical fingerprints that we first use as prompt enhancements and then employ as means to identify the differences between human and generated content.

The setting in which we test this is sensitive: we use anonymized parole suitability hearing transcripts from California and task the model with generating the decision statement of the presiding commissioner. By distilling a rhetorical fingerprint of the commissioner across multiple hearings, we compare the effect of prompting several models with the fingerprint-enhanced prompt and their performance when prompted with a general prompt not containing the fingerprint. The experiments show that all LLMs seem to have their own linguistic fingerprint from which they do not deviate even if prompted so. Additionally, prompting the models to replicate the style they observe in a given text, does not succeed, as their own fingerprint remains more dominant.

2 Related Work

Recent studies explore how effectively LLMs can mimic human-like behavior in different aspects. For example, Milička et al. (2024) task different versions of OpenAI's GPT to impersonate children between two and six years old. Their findings show that the models are able to adapt their linguistic behavior to the developmental stage expected from them. Salewski et al. (2023) observe a boost in performance by prompting the LLM to act as a domain expert, but they also identify the reproduction of gender, age and racial biases in the model's output. Herbold et al. (2024) show that LLMs can impersonate politicians to the extent that the model responses are judged more authentic, relevant and coherent than the actual human responses.

To the best of our knowledge, there has been no work on how well LLMs perform in mimicking human-like speech with rhetorically enhanced prompts. Recently, several studies have focused on the capabilities of LLMs to emulate human writing styles by looking into coarse and more fine-grained linguistic analysis (Bhandarkar et al., 2024; Alhafni et al., 2024). Bhandarkar et al. (2024) test the performance of 12 pre-trained LLMs for stylistic rewriting, by instructing them to mimic the author's

Proceedings of the First Workshop NLP and Language Models for Digital Humanities associated with RANLP 2025, pages 143–154, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-106-6-014

writing style together with shallow guided instructions regarding different linguistic features. While their results show that current models are able to replicate author style to some extent, they are not capable of producing text that is fully indistinguishable from that of the original author.

In a more recent approach, Dinu et al. (2025) tested how good LLMs are at imitating writing styles by prompting it to complete an author's unfinished novel. While LLMs perform acceptably in mimicking the literary style, their quality was not assessed as being as good as the human written ones.

In this present study, we build on this line of research. We extend the focus from imitating writing styles to simulate spoken language, by using detailed rhetorical prompts. We distill the linguistic characteristics of each person and enhance the prompts dynamically to simulate specifically tailored spoken natural language dialogue.

3 The data

3.1 Parole Suitability Hearings

In California, an inmate's potential to be reintegrated into society despite serving a life sentence is assessed by one presiding and one deputy commissioner during parole suitability hearings (PSHs). After an hour-long interview with the inmate and their attorney, the presiding commissioner communicates the decision, taking into account the inmate's answers, a review of the rehabilitation plan, psychological assessments and disciplinary records.

Typically, decision statements follow a structured scheme, including an introduction, the announcement of the final decision, a discussion of the mitigating and aggravating factors, such as the institutional behavior of the inmate and the life crime itself. In case of a parole denial, commissioners may give recommendations for improvement. Additionally, they are required to set a denial length, which determines when the inmate is eligible to reappear before the parole board. While these elements are consistently covered by all commissioners, each commissioner may change the order of covering those parts in their statements or may choose to discuss one factor more in detail than others. Rhetorically, the decision statement has to establish authority by keeping a professional tone, at the same time signaling empathy and a reasoned judgment. We incorporated all these structural requirements in our prompt design to ensure alignment of the generated decision statements with the content observed in actual parole hearing decision statements.

3.2 The PSH v1.0 corpus

The dataset that underlies the present study, PSH v1.0, comprises 100 parole hearing transcripts that we requested from the California Department of Corrections and Rehabilitation (CDCR)[1]. We employ the anonymization model of Itani et al. (2024) to remove any instances of names, locations and age-related information to ensure no personal details of any individual involved in the parole hearings is leaked.

For PSH v1.0 we select two female and two male presiding commissioners with 25 transcripts per commissioner. The PDF files range between 37 and 162 pages (8,171 pages in total) and contain the verbatim transcripts of the hearing. The first section of the transcript contains all content said during the interview of the parole hearing. Altogether, this section amounts to 1,297,488 words in PSH v1.0 (excluding punctuation and numbers), with a range of 4,141 to 29,278 words per transcript.

The second section of each transcript contains the decision statements. As we are only interested in the statement provided by the presiding commissioner, we remove all utterances by other speakers. These include mainly interruptions by inmates, translations and supplementary remarks made by the deputy commissioners. The human presiding commissioner statements are between 890 and 4,049 words long and have not been shown to the LLMs tested in this study. We use those decision statements to first distill the rhetorical fingerprint of each commissioner and then to compare the original statements with the LLM-generated statements.

4 Rhetorical Fingerprints

4.1 The dimensions

To assess the relevant rhetorical characteristics of human decision statements, we conduct a manual analysis of 20 decision statements to identify key rhetorical features that are across all four presiding commissioners. This set of linguistic features represents the collective speech style of the presiding commissioners overall, as well as their individual

[1]https://www.cdcr.ca.gov/bph/psh-transcript/

speech style. Deriving both the collective and the individual fingerprint allows us to (1) create an individual linguistic profile to incorporate into the prompt and (2) to conduct a systematic comparison of authentic commissioner statements and the LLM-generated counterparts. The following features are taken into account:

Sentence complexity This feature gives a measurement of the syntactic complexity employed by the presiding commissioners when formulating the sentences. The score is calculated by counting the number of clausal modifiers, conjuncts, adverbial clauses, clausal complements, clausal subjects, and parataxes in each sentence, based on the dependency tag given to each token by SpaCy (Honnibal and Montani, 2017). We then average the complexity over all sentences.

Lexical diversity To assess the lexical diversity, e.g. how much variety and complexity there is in the statements, we use the measure of textual lexical diversity (MTLD) (McCarthy and Jarvis, 2010). For implementation we use the module provided by Shen (2022). Unlike Type-Token-Ratio (Chotlos, 1944), MTLD is length-independent and measures how many words are needed before the Type-Token-Ratio falls below a predefined threshold. Due to the difference in text length between original and AI-generated commissioner statements, we use MTLD for reasons of comparability. An MTLD score is calculated for each of the 25 decision statements and then averaged, resulting in an overall measure per commissioner.

Discourse markers We expect a coherent line of discourse and argumentation in a legal context such as parole hearings. Discourse markers such as *because*, *therefore*, and *however* help to link evidence and conclusions and contribute to the perception of fairness and transparency. We measure the construction of reasoned decisions by counting the occurrence of discourse markers listed in the PDTB resource (Prasad et al., 2008). For aggregating the information, we divide the number of discourse markers across all 25 decision statements by the total number of words spoken by each commissioner.

Nominalizations Nominalizations are known to abstract the responsibility and obscure agency (Fairclough, 2001). They are therefore attributed to an authoritative and bureaucratic tone. Although they are usually attributed to formal written language (Siskou et al., 2022), the manual analysis of the transcripts suggests that they are also relevant in the current context. We estimate the preference for nominalizations by counting nouns ending on *-tion*, *-ment*, *-ance*, etc. across all 25 decision statements and dividing them by the total wordcount per commissioner.

Modals Modal verbs like *must*, *should*, *could* encode power (Fairclough, 2001) and are often used by commissioners to frame parole decisions. Depending on the type and frequency of usage they may convey obligation and institutional authority or empathy. Building on the wordlist for modality used by Herbold et al. (2023), we added a few more modal verbs and adverbs to evaluate the degree of assertiveness in the commissioner's speech style. Modals are aggregated in the same way as nominalizations.

Pronoun usage Pronouns are an important linguistic feature, signaling how the presiding commissioners relate to the inmates and their role in the hearing. We distinguish two dimensions: First, the addressing of the inmate either with *you* versus the reference with *he* (there are no female prisoners in PSH v1.0), latter signaling are more distanced tone. Second, pronouns used when the presiding commissioners refer to themselves (e.g., the more personal *I*) versus a more collective reference (e.g., *we*). Each pronoun version is aggregated in the same way as nominalizations and modals are.

Jargon In institutional settings, such as parole hearings, legal jargon conveys authority, but does also exclude and confuse people who are not familiar with the domain. To compile a domain-specific wordlist of legal terms that are common in the context of parole hearings, we extracted all nouns in the corpus that occurred in at least 3 different decision statements. We then manually went through this list and selected only abbreviations and parole hearing specific and crime related terms. We proceeded in the same way for bigrams and trigrams of consecutive nouns. After the statement generation, we repeated this process to expand the list of jargon used by the LLMs. The final list mainly consists of abbreviations for rehabilitation programs, statutory references, as well as references to forms that inmates can request to file. We normalize jargon usage by dividing the frequency count by the total number of spoken words.

4.2 Initial findings

Overall, we attribute high sentence complexity, high lexical diversity, as well as a frequent use of nominalizations, modals, jargon, indirect references to the inmate (by using third person singular pronouns), and collective self-reference (first person plural) to an authoritative tone. Addressing the inmate directly and framing the decision in the first person singular are considered as empathetic language. The use of discourse markers indicates a reasoned judgment.

Commissioners might choose to alternate between direct and distanced references to the inmate. In our dataset, we see a frequent use of second-person singular pronouns (*you, your*), to directly address the inmate either throughout the entire decision statement or only when providing recommendations to the inmate directly for future hearings. However, some commissioners completely avoid direct engagement with the inmate. In this case, addressing the inmate by using third-person singular pronouns (*he, she, the inmate*) establishes a more distanced tone and enhances the power distance between commissioners and the inmate.

Similarly, depending on the rhetorical intent of the statement, they choose between pronouns that frame the decision as a collective agreement between commissioners or those that signal the commissioner's personal opinion. To emphasize the collective nature of the board's decision, most commissioners use first-person plural pronouns (*we, our*). Through point-of-view distancing ((Brown and Levinson, 1987, p. 204-206); (Locher, 2004, p. 130)) the speaker puts the focus on the idea that the decision resulted of panel deliberation and therefore can distance themselves from individual responsibility. Phrases like e.g. *"Subsequent growth [...] and increased maturity, um, while incarcerated, as we've reflected on this, we didn't find much."* highlight the institutional nature of the decision and signal unity of the commissioners in decision-making. Some commissioners might use first-person singular (*I, my*) pronouns when announcing the decision. As this breaks from the institutional neutrality, it is rather unusual. When it does appear, it typically is used to signal strong personal conviction, as in *"You had a rule violation, a pattern, um, back in and, uh, as I looked through your history [...]"*.

4.3 Standardization and visualization

To normalize the linguistic features, we first calculate their relative frequency for each presiding commissioner across all of their 25 original decision statements. Frequency counts of each feature are aggregated and divided by the total word count per commissioner.

As the features selected for the linguistic evaluation do not share the same distribution, we standardize the normalized feature values with z-scores for better comparability using the `pandas` library. This allows us to observe differences in language use and in units of standard deviations from the mean use across all commissioners, giving insight into how individual presiding commissioners differ from the group norm in their feature use. Figure 1 shows an example of a fingerprint visualization.

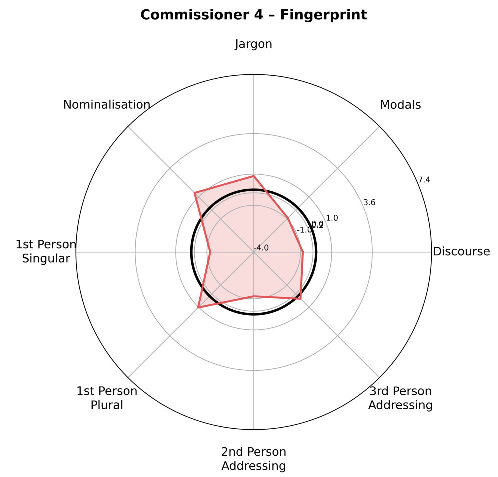


Figure 1: Rhetorical fingerprint across eight dimensions. Lexical Diversity and Sentence Complexity were excluded from this visualization.

The axes of the radarplots represent the rhetorical features. The grey lines indicate the z-score scale. The thick grid line indicates a z-score of 0. Values below or above the thick grid line indicate a lower or higher usage of this particular feature compared to the average usage across all presiding commissioners. For instance, according to Figure 1, commissioner 4 uses more nominalizations, first person plural and jargon than their colleagues, with a z-score of 1 (or higher) – indicating that their jargon usage is at least one standard deviation above the group average. In contrast, they are using less modals, discourse markers and first person singular than their colleagues. Although not included in the Figure above, the z-scores from Table 1 show that compared to their colleagues, commissioner 4 prefers statements with a higher lexical diversity as well as more complex sentences.

This information is used for an enhanced prompting of the models with a speaker-specific rhetorical fingerprint, turning the z-scores into natural language text. The details are discussed in the following section.

5 Prompt engineering

5.1 Assembling the system prompt

In the system prompt, we instruct the models to impersonate an experienced presiding commissioner in a Californian parole hearing and to generate a decision statement about whether to grant or deny parole to the inmate. The decision must be based on the information given in the transcript (provided in the user prompt) and on California state laws and policies. A description of the parole process that is publicly available on the official website of the Board of Parole Hearings[2] and that explains the general factors that need to be considered to assess the risk of reintegration into society (c.f., Section 3.1) is also included in the system prompt. To ensure realistic output, we instruct the models to deliver a spoken statement. We also emphasize the importance of professionalism and factual grounding to prevent the LLMs of inventing details to the case. We explicitly prohibit headings and bullet points. The exact system prompt can be found in Appendix A.1.

5.2 Assembling the user prompt

While the system prompt provides the more general information about how we expect the LLMs to behave as presiding commissioners, the user prompt provides more detail on the linguistic characteristics expected in the outputs as well as some structural guidance (e.g., by providing introductory phrases for the decision statements and instructions on what to discuss in the statement itself).

The rhetorical fingerprint is assembled in a building-block manner with static feature descriptions and commissioner-specific prompt sections. First, we define the usage categories. Previous studies (Sun et al., 2023) show that LLMs tend to underperform when prompted to use specific features with a hard-restricted frequency. We therefore turn the z-scores from the rhetorical fingerprint into natural language sentences and provide those in the user prompt. The z-scores are converted into four categories, namely ‘strong’, ‘frequent’, ‘rare’, and ‘avoided’ feature usage by the following heuristics:

- **Strong usage**: if $z_f > 1$
- **Frequent usage**: if $0 < z_f \leq 1$
- **Rare usage**: if $-1 < z_f \leq 0$
- **Avoided usage**: if $z_f \leq -1$

Second, we add static explanations for each linguistic feature in the fingerprint and add a usage instruction based on the previously distilled fingerprints for each commissioner. The user prompt additionally provides the transcript and placeholders for metadata concerning age and gender of the inmate for each case. We also provide two typical opening lines that we take from the original transcripts and instruct the LLMs to not use section headings or bullet points. To avoid hallucinations, we include a section that demands the models to only rely on facts given in the transcript. In the end we arrive at a commissioner-specific impersonation prompt, an example of which can be found in Appendix A.2.

To test whether these precise linguistic instructions improve the rhetorical alignment, we mirror the fingerprint prompt with a simplified version of the user prompt, including only the general information about parole hearings. Under this condition, we remove the information about the speaker specific fingerprints and task the LLMs to mirror the language style of the presiding commissioner by drawing on the language patterns in the transcript without any guidance on linguistic features and style. We refer to this condition as primed-by-corpus.

5.3 Prompting parameters

During the prompt engineering phase, we test the performance of all LLMs on a transcript that is not included in the final corpus. We tested the performance of user and system prompts multiple times in an iterative way. Swapping information between the system and the user prompt did not result in any notable difference in response quality. The temperature is set to 0.3 for stylistic consistency after testing for multiple other temperature settings. The three state-of-the-art models, namely GPT-4o, GPT-4.1 and DeepSeek R1, were accessed and prompted via their respective API by using the same system

[2] https://www.cdcr.ca.gov/victim-services/parole-process/

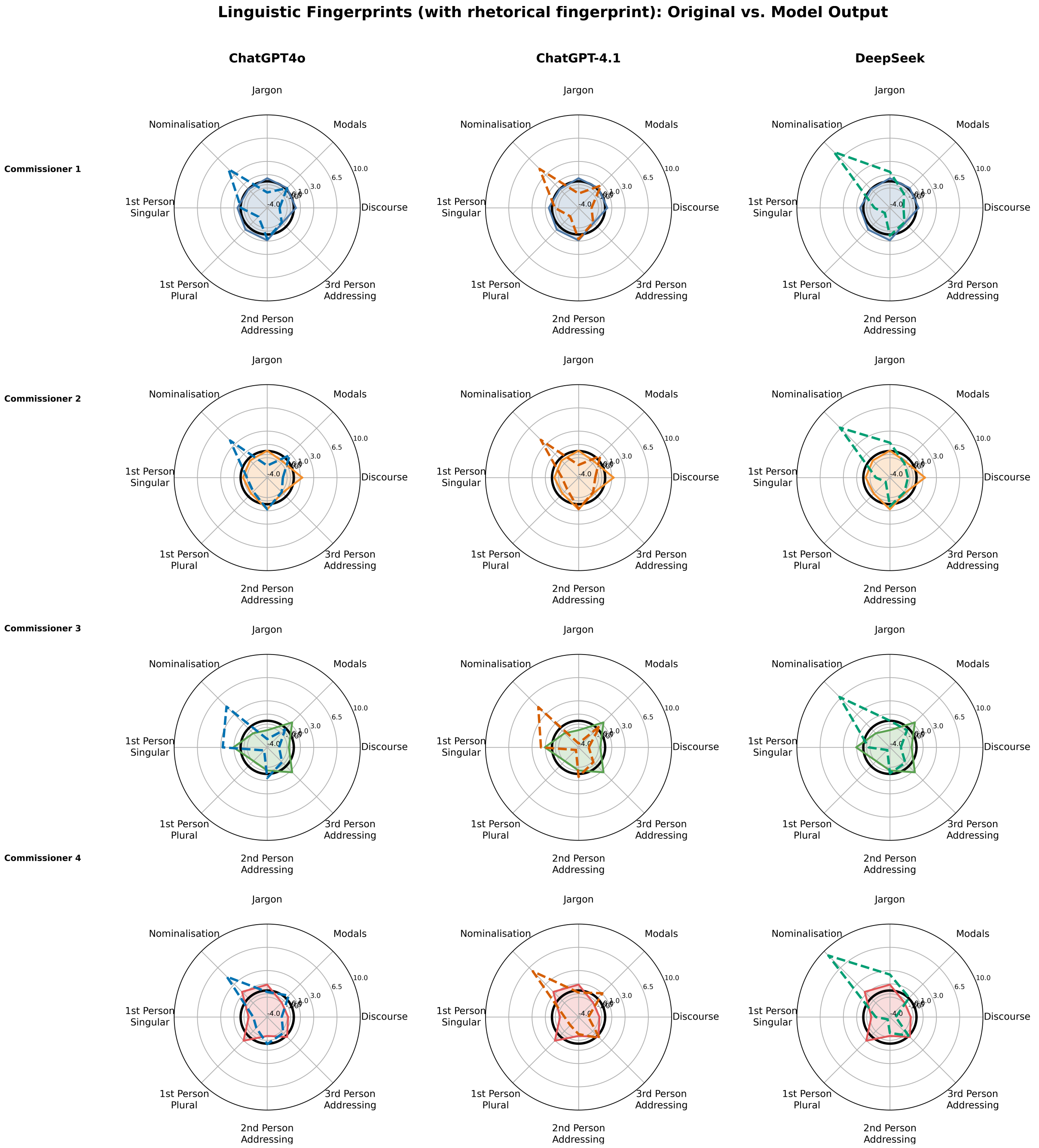


Figure 2: Comparison of rhetorical fingerprints when prompted with a commissioner's rhetorical fingerprint. Original vs. Generated statements. Solid lines indicate original commissioner fingerprints. Dotted lines indicate the fingerprint of the respective models. The thick grid line indicates a z-score of 0.

and user prompts, with and without commissioner-specific rhetorical fingerprints. Despite setting the output token parameter to the highest possible for each model, we observe that all three models give responses that are way below their maximum token output limit.

6 Results

The comparison in this paper is two-fold: First, we identify the rhetorical differences that hold between human and generated, impersonated content. Second, we investigate whether an enhanced prompt with a rhetorical fingerprint yields responses with a higher level of impersonation than a 'plain' prompt with general instructions.

Regarding the first question, we compare the rhetorical approximation of the generated statements with the original commissioners' rhetorical patterns. To this end we calculate the z-scores for each of the selected linguistic features between

Comm.	Original				ChatGPT-4o								ChatGPT-4.1								DeepSeek							
					byCorpus				fingerprint				byCorpus				fingerprint				byCorpus				fingerprint			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Lexical Diversity	-0.55	-0.32	-0.61	1.48	16.92	17.71	16.21	16.39	21.32	22.80	21.25	21.88	5.82	6.06	6.29	5.76	16.57	17.36	16.56	15.67	37.80	38.35	40.20	38.68	40.49	40.28	39.52	36.46
Sentence Complexity	0.04	0.16	-1.31	1.10	25.32	27.67	26.35	27.61	30.02	31.76	33.30	33.66	17.52	19.02	18.24	16.52	23.25	27.20	25.26	27.13	28.23	30.18	28.64	29.96	32.09	30.48	31.19	34.91
Discourse Markers	0.31	1.27	-0.73	-0.85	-2.23	-2.29	-2.09	-2.35	-2.14	-1.65	-2.21	-1.83	-0.93	-0.66	-1.00	-0.73	-2.05	-1.49	-2.47	-2.52	-1.95	-1.62	-2.72	-2.16	-1.95	-1.28	-2.36	-3.13
Modals	0.22	-0.62	1.31	-0.90	-0.18	0.28	-0.09	-0.30	0.22	0.78	-0.20	0.57	-0.83	-0.44	-0.41	-1.10	0.77	0.61	0.56	1.07	-1.06	-1.03	-0.23	-1.20	-1.01	-0.92	-0.36	0.12
Jargon	0.46	0.06	-1.41	0.88	-2.26	-2.75	-2.50	-2.03	-1.68	-2.19	-2.76	-0.21	0.49	-0.62	-1.33	0.27	-1.83	-2.10	-3.41	-0.30	3.16	2.66	1.01	3.18	1.43	1.26	0.01	2.40
Nominali-zation	-0.008	-0.35	-1.00	1.36	3.90	3.61	4.11	4.20	4.24	3.91	4.66	4.59	3.28	3.37	4.03	3.43	4.30	4.04	4.57	5.75	7.42	6.73	7.03	6.47	7.90	6.64	6.73	9.10
3rd Person Addressing	-0.79	-0.76	1.31	0.24	-0.91	-0.92	-0.91	-0.91	-0.91	-0.91	-0.84	-0.60	-0.90	-0.91	-0.83	-0.88	-0.90	-0.90	-0.85	0.45	-0.91	-0.89	-0.88	-0.83	-0.91	-0.91	-0.72	-0.05
2nd Person Addressing	0.88	0.78	-0.50	-1.15	0.77	0.83	0.57	0.71	0.65	0.77	0.61	0.07	1.22	1.25	0.97	1.12	0.70	0.74	0.44	-1.43	0.22	0.34	0.21	0.04	0.18	0.32	-0.08	-1.60
1st Person Plural	0.63	-0.58	-1.08	1.04	-1.37	-1.13	-1.54	-1.36	-2.13	-1.03	-3.35	-1.69	-2.74	-2.75	-2.73	-2.60	-2.22	-1.56	-3.45	-2.25	-3.08	-3.21	-3.16	-3.01	-2.92	-3.09	-3.44	-3.52
1st Person Singular	0.48	-0.35	1.08	-1.21	-1.48	-1.47	-1.50	-1.47	-0.05	-0.94	2.69	-1.90	-0.85	-1.01	-0.61	-0.94	-0.43	-1.50	1.66	-1.90	-1.81	-1.81	-1.76	-1.79	-1.68	-1.90	-0.79	-1.87

Table 1: Comparison of z-scores for all features across original and LLM-generated outputs

original and generated statement by first normalizing the frequency count for each feature in the generated statements. To calculate the z-scores we use the mean and standard deviations of each feature calculated from the original commissioner statements. Using the original mean and standard deviation metrics establishes the baseline against which the generated decisions statements are compared. The resulting z-scores show to which degree the generated statements deviate from the original statements: Positive z-scores indicate a stronger usage compared to the commissioner average, negative values reflect underuse (in terms of standard deviations). An overview of the performance of each LLM for the primed-by-corpus and rhetorical fingerprint scenario can be found in Table 1.

Figure 2 shows the resulting radar plots of the rhetorical fingerprint prompts in comparison to the original fingerprint visualization[3]. The axes of the radarplots represent the individual features. The grey lines indicate the z-score scale. The thick grid line indicates a z-score of 0. Columns represent LLMs, while lines represent the individual commissioners. Each LLM and commissioner is color-coded. Solid lines represent the scores in the rhetorical fingerprint of the original commissioners, while the dotted lines show the rhetorical approximation of the generation models.

In the following we discuss the dimensions in the fingerprint in terms of how the models deviate rhetorically from the original commissioners rhetorical patterns.

6.1 Lexical diversity and sentence complexity

From a procedural point of view, high lexical diversity and/or sentence complexity makes parole hearing decision statements difficult to understand, going against the guideline that the hearings should be accessible and easy to understand by the inmates. The original decision statements exhibit a relatively stable usage of minimal lexical diversity and simple sentence structures, indicating that commissioners are mindful about making their statements accessible. In our dataset, Commissioner 4 is the only one showing an elevated z-score for lexical diversity (1.48) and sentence complexity (1.10).

The analysis of the LLM-generated statements shows that all three LLMs highly deviate in lexical diversity and sentence complexity compared to the original decision statements (see Table 1), across impersonated commissioners and conditions in the user and system prompt. GPT-4.1 shows the lowest z-score for lexical diversity (5.76) and sentence complexity (16.52) for Commissioner 4 in the primed-by-corpus condition. All other models exceed these z-scores substantially (z-scores range from 5.76 to 40.20 for Lexical Diversity and 16.52 to 34.91 for sentence complexity). DeepSeek demonstrates the highest z-scores for both features, indicating that even elaborate prompting does not help to mitigate this behavior. Taken together, this indicates that LLMs are insensitive to prompts when it comes to aligning spoken content in terms of its lexical diversity and sentence complexity. This is probably due to the underlying training data being mostly written language.

Due to the emerged non-alignment in terms of lexical diversity and sentence complexity, we exclude both dimensions from the radar plots in Figure 2 to prevent those features from skewing the plots.

[3]The visualizations for the primed-by-corpus condition can be found in Appendix B.

6.2 Nominalizations and jargon

Nominalizations and a frequent use of domain-specific jargon are attributed to written communication and make the content of the statement inaccessible to individuals who are not familiar with legal language. In the original statements we observe a variety of jargon and nominalization preference patterns. What we observe consistently is that commissioners who are using more jargon also use more nominalizations than their colleagues and vice versa. The plots in Figure 2 indicate a consistent underuse of jargon in both GPT models prompted with rhetorical fingerprints, unless they are prompted to use jargon strongly (Commissioner 4). When prompted with the more general prompt without the rhetorical fingerprint, GPT-4o continues to underperform, while GPT-4.1 seems to approximate the linguistic behavior of the original commissioner and thus infers the degree of usage of this feature. DeepSeek consistently overuses jargon.

Additionally, all three models show a strong preference for using nominalizations, suggesting a strong bias towards formal and written language, likely due to their training data. A similar observation has been made by McGovern et al. (2025), who show that LLMs exhibit a high usage of nouns in their responses. This behavior cannot be mitigated by prompting and holds across all models and conditions. Prompting for strong usage of nominalizations even triggers the models to use more nominalizations than they already do (see DeepSeek for Commissioner 4, where z-score was 6.47 for primed-by-corpus and 9.1 for fingerprint condition).

6.3 Modal verbs and discourse markers

Modal verbs and discourse markers are important features for parole decision statements as they help to convey authority and coherence by expressing obligations and transparency about the reasoned judgment. The commissioners in our dataset either use modal verbs frequently or rarely. The same applies for discourse markers, but Commissioner 2 is the only one showing a strong preference to use them. The analysis of the generated statements shows that all models tend to underuse modal verbs. Only GPT-4o fully replicates the behavior of Commissioner 1 when prompted with the rhetorical fingerprint instruction to frequently use modal verbs.

Discourse markers are consistently underused by all models and all commissioners across prompting conditions. This indicates that LLMs show limited sensitivity to these features. We attribute this to the fact that LLMs might interpret discourse markers as filler words which can be dropped without affecting the semantic structure of the generated text. Overall, we can conclude that even when explicitly prompted with specific usage instructions all LLMs show limitations in their ability to replicate reasoning structures and modality for most cases.

6.4 Pronouns

Addressing the inmate in the third person singular ('he', 'she',) even if they are present in the same room conveys authority and manifests power. We only see a preference for third person singular addresses with Commissioners 3 and 4, whereas Commissioners 1 and 2 prefer to address the inmates directly by using the more personal 'you'. By looking at the radar plots for all commissioners in Figure 2, we see a pattern of preferring second person singular addresses across all models. Prompting them for strong indirect inmate addressing does not yield the expected results (see model performance for Commissioner 4). GPT-4.1 and DeepSeek follow this instruction to a very small extent, but only if prompted for strong usage. We attribute this behavior to our prompt context which explicitly asks for conversational tone. The underlying dialogue training data for each model is very likely coming from written online communications (e.g. Reddit), where indirect addresses are uncommon. LLMs may therefore infer that they are speaking with the inmate, instead of about them.

Regarding the use of pronouns when referencing either themselves as individuals or as a collective, we observe that all models underuse the first person plural. We suspect here that our prompts are being misinterpreted by the models, which are unaware of the fact that the presiding commissioner is the representative of the parole hearing panel. They therefore default to the individual self-referencing pronoun 'I' (which is also more likely to be overrepresented in their training data).

6.5 LLM-specific fingerprint

All models exhibit their own model-specific linguistic characteristics, independent of the prompt. This suggests that some features are inherent to the model's own rhetorical style and are therefore not adjustable by prompting at all or only to a small degree. This is particularly evident when comparing the radar plots of GPT-4o and GPT-4.1 under the rhetorical fingerprint condition: The overall shape of the fingerprint remains nearly identical across these two model versions when prompted with the speaker-specific rhetorical fingerprints, suggesting that linguistic characteristics are inherited across model versions. For the DeepSeek model we barely see any change in linguistic behavior between the two prompting conditions. Only when prompted for strong usage of nominalizations we see a minor adjustment in feature intensity. Nevertheless, the instruction of using first person plural pronouns gets ignored completely, which reinforces our suggestion of model-specific rhetorical fingerprints.

7 Conclusion

In this study we test whether state-of-the-art LLMs can be pushed to impersonate humans when prompted with linguistically informed fingerprints. Our findings from testing three off-the-shelf models in generating parole hearing decision statements, a high-stake setting, underscore the current rhetorical limits of LLMs in mirroring human-like behavior as they fail to deviate from their own model-specific rhetorical fingerprint.

The next step in this endeavor is a more detailed investigation of the effect of rhetorically enhanced prompting and the outcome of these hearings (whether parole is granted or not and under which conditions). A significant amount of more detailed analyses is required in order to show the limits of applying LLMs in sensitive and high-stake settings like the present one.

Limitations

Feature selection. In this study we only consider a limited number of linguistic features, which does not cover the full complexity of human rhetorical characteristics. We restricted the prompts to stylistic characteristics, which may oversimplify human language.

Domain. We test the ability of three LLMs on a very specific domain. Future work should look into the performance of LLMs when mimicking human-like speech in different domains.

Model selection. At the time of writing, we covered the three most popular LLMs, all of which do not openly disclose their training data. Therefore, we do not know whether parole hearing transcripts were included in the model's training data. More recent models may differ in performance.

Ethics Statement

While this paper focuses on the linguistic capabilities of LLMs to reproduce a certain rhetorical fingerprint, we are aware of the potential risks associated with generating human-like institutional dialogue. This study intends to assess the stylistic approximation capabilities of LLMs within a controlled research setting. It is not intended to support or encourage the use of LLMs for deceptive or harmful applications, especially in legal settings. We also do not propose to actually use them in parole hearing evaluations. For reasons of data protection, we will not publish the original, nor the anonymized transcripts. However, we will provide a list of the individual hearings upon request, making it possible to interested researchers to request the exact same files from the CDCR for replication purposes.

Acknowledgments

The work reported on in this paper was funded by the Deutsche Forschungsgemeinschaft (DFG – German Research Foundation) under Germany‘s Excellence Strategy – EXC-2035/1 – 390681379 as part of the project “Inequality in Street-level Bureaucracy: Linguistic Analysis of Public Service Encounters”.

References

Bashar Alhafni, Vivek Kulkarni, Dhruv Kumar, and Vipul Raheja. 2024. Personalized text generation with fine-grained linguistic control. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 88–101, St. Julians, Malta. Association for Computational Linguistics.

Avanti Bhandarkar, Ronald Wilson, Anushka Swarup, and Damon Woodard. 2024. Emulating author style: A feasibility study of prompt-enabled text stylization with off-the-shelf LLMs. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 76–82, St. Julians, Malta. Association for Computational Linguistics.

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Studies in Interactional Sociolinguistics. Cambridge University Press.

John W. Chotlos. 1944. IV. A statistical and comparative analysis of individual written language samples. *Psychological Monographs*, 56(2):75–111. Place: US Publisher: American Psychological Association.

Anca Dinu, Andra-Maria Florescu, and Liviu Dinu. 2025. Analyzing large language models' pastiche ability: a case study on a 20th century Romanian author. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 20–32, Albuquerque, USA. Association for Computational Linguistics.

Norman Fairclough. 2001. *Language and Power*. Language in social life series. Longman.

Steffen Herbold, Annette Hautli-Janisz, Ute Heuer, Zlata Kikteva, and Alexander Trautsch. 2023. A large-scale comparison of human-written versus ChatGPT-generated essays. *Scientific Reports*, 13(1):18617.

Steffen Herbold, Alexander Trautsch, Zlata Kikteva, and Annette Hautli-Janisz. 2024. Large language models can impersonate politicians and other public figures.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Abed Itani, Wassiliki Siskou, and Annette Hautli-Janisz. 2024. Automated anonymization of parole hearing transcripts. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 115–128, Miami, FL, USA. Association for Computational Linguistics.

Miriam A. Locher. 2004. *Power and Politeness in Action*. De Gruyter Mouton, Berlin, New York.

Philip M. McCarthy and Scott Jarvis. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2):381–392.

Hope Elizabeth McGovern, Rickard Stureborg, Yoshi Suhara, and Dimitris Alikaniotis. 2025. Your large language models are leaving fingerprints. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 85–95, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Jiří Milička, Anna Marklová, Klára VanSlambrouck, Eva Pospíšilová, Jana Šimsová, Samuel Harvan, and Ondřej Drobil. 2024. Large language models are able to downplay their cognitive abilities to fit the persona they simulate. *PLOS ONE*, 19(3):1–25.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. 2023. In-context impersonation reveals large language models' strengths and biases. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Lucas Shen. 2022. LexicalRichness: A small module to compute textual lexical richness.

Wassiliki Siskou, Laurin Friedrich, Steffen Eckhard, Ingrid Espinoza, and Annette Hautli-Janisz. 2022. Measuring plain language in public service encounters. In *Proceedings of the 2nd Workshop on Computational Linguistics for Political Text Analysis (CPSS-2022) Potsdam, Germany*.

Jiao Sun, Yufei Tian, Wangchunshu Zhou, Nan Xu, Qian Hu, Rahul Gupta, John Wieting, Nanyun Peng, and Xuezhe Ma. 2023. Evaluating large language models on controlled generation tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3155–3168, Singapore. Association for Computational Linguistics.

A Prompts

A.1 System prompt

This is the general system prompt that we used for all 4 commissioners under all conditions:

```
    You are an experienced Parole Commissioner from
the Board of Parole Hearings in California, deciding
whether to grant or deny parole to inmates.  You
need to make informed parole decisions based on
California state laws and policies, including
the guidelines of the California Board of Parole
Hearings.
Parole proceedings are not to decide guilt or
innocence.  The Board of Parole Hearings accepts as
fact the guilty verdict imposed by the courts.  The
purpose of a parole proceeding is to determine if or
when an inmate can be returned to society.  Under
normal circumstances, the panel or the Board shall
set a release date unless it determines that the
gravity of the crime (offense), or the timing and
gravity of current or past convictions, requires
a more lengthy period of incarceration to ensure
public safety.

In general, some of the factors considered by
the panel and which are discussed in the proceeding
include:
    • counseling reports and psychological
      evaluations
    • behavior in prison (i.e., disciplinary notices
      or laudatory accomplishments)
    • vocational and educational accomplishments in
      prison
    • involvement in self-help therapy programs that
      can range from anti-addiction programs for
      drugs and alcohol to anger management
    • parole plans, including where an inmate would
      live and support themselves if they were
      released
    After reading the transcript provided by the
user, your task is to decide whether the inmate
should be granted or denied parole and deliver your
decision as spoken dialogue, mirroring a natural,
ongoing conversation in the hearing room.  Do not
use headings or bullet points in your statement.

Remain professional and consistent with the tone
and format expected of official parole decision
statements.  However, because you will be delivering
this decision as spoken dialogue, adapt the
formality to reflect a real parole hearing's
conversational flow.  You are not allowed to use
bullet points or headings in your statements.  You
must provide detailed, professional, fair, and
well-reasoned responses.  Avoid bias, stereotypes,
prejudice, or speculation.  Refer only to the facts
```

and background details included in the transcript. If some details are missing, acknowledge them rather than inventing information.

A.2 User prompt

This is an example of a commissioner prompt (commissioner 4). Passages written in blue were only included under the rhetorical fingerprint condition. Passages written in violet were only included under the primed-by-corpus condition. Passages written in black appear in both conditions.

```
   You will read a transcript of a Californian
parole hearing and act as the presiding
commissioner.
   After reading the transcript, your task is to
decide whether the inmate should be granted or
denied parole and then deliver your decision as
spoken dialogue, mirroring a natural, ongoing
conversation in the hearing room.
   These are the overall style requirements:
    • Conversational tone: avoid enumerations,
      headings, or overly formal written structures.
      Instead, formulate your response as if it
      were spoken in a parole hearing. You may
      use occasional pauses and conversational
      transitions to make it flow naturally.
    • Commissioner style: You are speaking as the
      presiding commissioner. Please adapt your
      response to reflect the commissioner's typical
      language style, including tone, sentence
      structure, level of formality, and use of
      hesitation markers. Your statement should
      feel authentic to a commissioner's usual way
      of delivering decisions.
Do not label sections in your final text, but
address these points in a conversational and
detailed manner:
    • Introduction: Set the context of the hearing.
      You can use these opening lines to do so:
      "Today's date is [MONTH] [DAY], [YEAR]. The
      time is approximately [TIME] AM. All parties
      who were present before have returned." or
      "Today's date is [MONTH] [DAY], [YEAR]. The
      time is approximately [TIME] PM. We're back in
      the matter of Mr. ..."
    • Decision: Clearly state whether the inmate is
      granted or denied parole.
    • Evaluation and Reasoning: Discuss both
      aggravating and mitigating factors that
      influence your decision.
    • Recommendations (if parole is denied):
      Specify the denial length and explain the
      reasons for setting that length. You can set
      3, 5, 7, 10 or 15 years of length, depending
      on the severity of each case. Offer detailed
      suggestions for what the inmate could do to
      improve the likelihood of a positive outcome
      at a future parole hearing (e.g. additional
      programming, self-improvement efforts, insight
      development).
    • Clarify (only if parole is granted): Clarify
      that this decision is not final and will be
      subject to further review by the Governor.
      Explain that the inmate will be formally
      notified in writing once a final decision is
      made.
   After reading the transcript, your task is to
decide whether the inmate should be granted or
denied parole and deliver your decision as spoken
dialogue, mirroring a natural, ongoing conversation
in the hearing room. Do not use headings or bullet
points in your statement.
   Below are the key linguistic features you may
use, along with usage instructions. Each feature
includes a Usage Category that can be set to any of
the following:
    • avoid: Do not use this feature.
    • rarely: Use this feature only a few times.
    • frequently: Use this feature regularly, but
      do not overuse it.

    • strong: Use this feature a lot.
   In your spoken statements, you are required to
use the following linguistic features with the
indicated frequency: In your spoken statements,
you are required to use the following linguistic
features with the indicated frequency:
    • Lexical Diversity to express nuanced
      viewpoints and considerations. Use a
      wide-ranging vocabulary by using synonyms
      and varied expressions throughout your
      statements. This corresponds to the usage
      category "strong".
    • Sentence complexity: Use a lot of complex
      and long sentences. This corresponds to the
      category "strong".
    • Discourse markers (e.g., "because", "however",
      "while") to indicate causal reasoning,
      contrasts, or transitions. Use these words
      rarely.
    • Modals: Words like "could", "should", "would",
      "may", "might" are modal verbs and are used
      to convey obligations or possibilities. Use
      these words rarely.
    • Nominalizations: Nominalizations are
      verbs that are turned into nouns, like e.g.
      "the denial", "the recommendation", "the
      rehabilitation". This is a strong feature.
      Use nominalizations very often.
    • Jargon: Strongly use legal terms legal terms
      like "recidivism", "suitability", "mitigating
      factors" and other technical terms that
      are typically used in the context of parole
      hearings.
    • First-Person Singular: Avoid using
      first-person singular pronouns in phrases
      like "I reviewed", "I find" to refer to the
      presiding commissioner's decision.
    • First-Person Plural: Strongly use phrases
      like "we reviewed", "we find" to refer to the
      panel's collective voice.
    • Second-Person Singular References: Avoid
      to directly address the inmate by using
      second-person singular pronouns.
    • Third-person singular when referencing the
      inmate in a detached or formal sense (e.g.,
      "he is not suitable for parole", "the inmate
      has demonstrated insight"). Frequently refer
      to the inmate by using third-person singular
      pronouns to address them in a more detached
      way.
   At the very end, include one of the following
lines:
    • Decision: granted
    • Decision: denied
   If you deny parole, also add:
    • Denial length: X years You can set 3, 5, 7,
      10 or 15 years of denial, depending on the
      severity of each case.
   Use the following details:
    • Inmate ID: inmate_id
    • Gender: gender
    • Current Age: current_age
    • Age of Imprisonment: age_of_imprisonment
   Main Part of the Hearing Transcript:
    • {transcript}
   Base your decision solely on the facts provided.
Write your response as one continuous speech,
providing detailed reasoning. Your statement must
be very long and detailed.
```

B Fingerprints primed-by-corpus condition

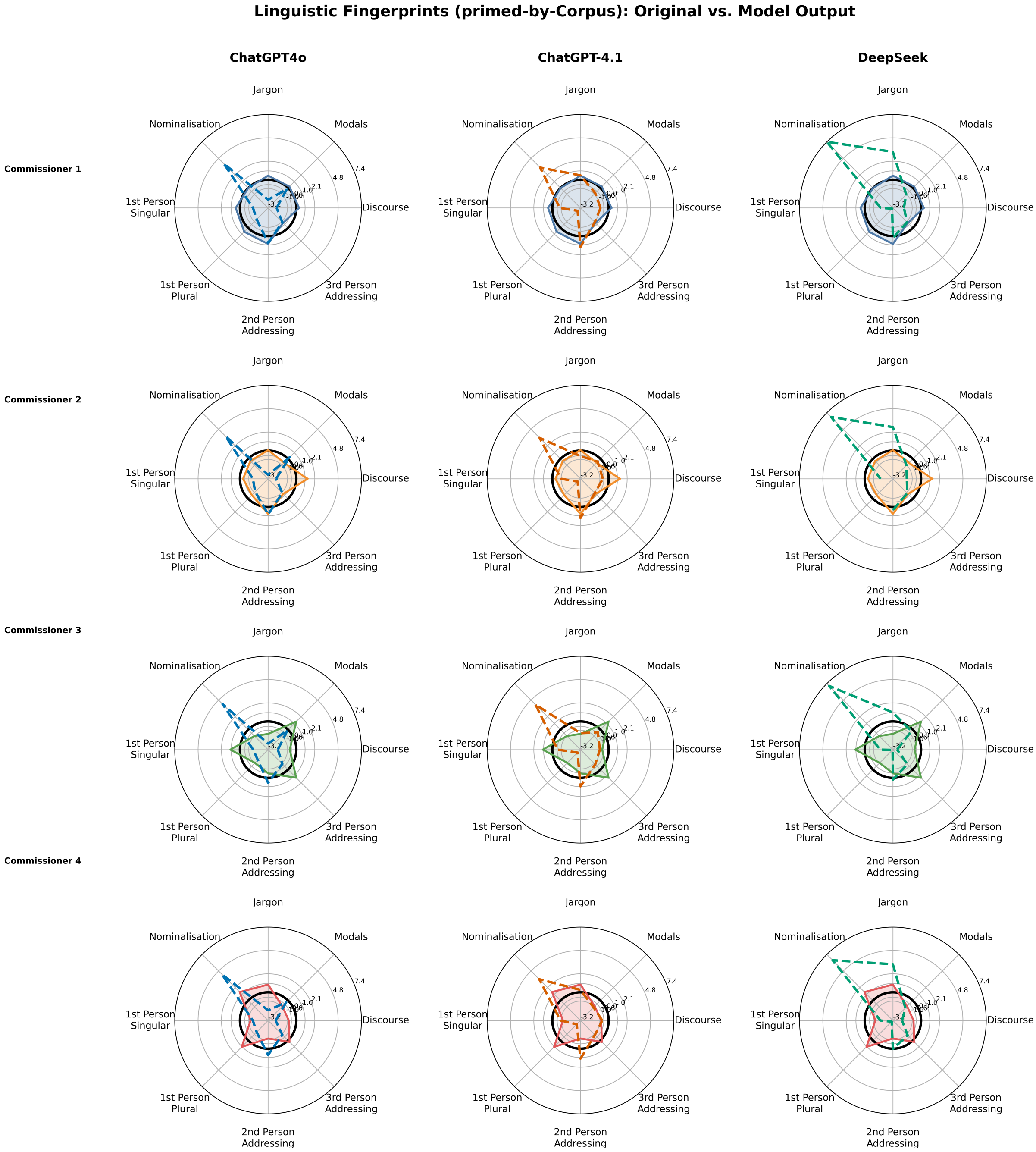


Figure 3: Comparison of linguistic fingerprints when primed-by-corpus. Original vs. Generated statements. Solid lines indicate original commissioner fingerprints. Dotted lines indicate the fingerprint of the respective models. The thick grid line indicates a z-score of 0.

Annotating Personal Information in Swedish Texts with SPARV

Maria Irena Szawerna[*], **David Alfter**[†‡], **Elena Volodina**[*]
[*]Språkbanken Text, SFS, University of Gothenburg, Sweden
[†]GRIDH, LIR, University of Gothenburg, Sweden
[‡]InfraVis, Sweden
`mormor.karl@svenska.gu.se`
`{maria.szawerna,david.alfter,elena.volodina}@gu.se`

Abstract

Digital Humanities (DH) research, among many others, relies on data, a subset of which comes in the form of language data that contains personal information (PI). Working with and sharing such data has ethical and legal implications. The process of removing (anonymization) or replacing (pseudonymization) of personal information in texts may be used to address these issues, and often begins with a PI detection and labeling stage. We present a new tool for personal information detection and labeling for Swedish, SBX-PI-DETECTION (henceforth SBX-PI), alongside a visualization interface, (IM)PERSONAL DATA, which allows for the comparison of outputs from different tools. A valuable feature of SBX-PI is that it enables the users to run the annotation locally. It is also integrated into the text annotation pipeline SPARV, allowing for other types of annotation to be performed simultaneously and contributing to the privacy by design requirement set by the GDPR. A novel feature of (IM)PERSONAL DATA is that it allows researchers to assess the extent of detected PI in a text and how much of it will be manipulated once anonymization or pseudonymization are applied. The tools are primarily aimed at researchers within Digital Humanities and Natural Language Processing and are linked to CLARIN's Virtual Language Observatory.[1]

1 Introduction and prior work

Personal information (PI) [2] is ubiquitous in many kinds of language data – data which oftentimes is the basis for research in fields such as Digital Humanities (DH), Natural Language Processing (NLP), or linguistics. When working with this kind of data, one may choose to employ privacy-protection measures in order to comply with appropriate legislation (e.g. GDPR (Official Journal of the European Union, 2016)) or out of ethically motivated concerns for data subject privacy. Commonly employed privacy-protection methods at the text level are the removal (*anonymization*) or replacement (*pseudonymization)* of PI. The difference is illustrated in (1), where *the original sentence* is anonymized using [redacted], pseudonymized using `tags` and pseudonymized using **replacement entities**.

(1) *Mitt namn är Sonja och jag är 29*
Mitt namn är [redacted] och jag är [redacted]
Mitt namn är `name` och jag är `age`
Mitt namn är **Anna** och jag är **31**

'My name is Sonja and I am 29'

While anonymization and pseudonymization can be carried out manually, it is very time-consuming. As with many other types of annotation, this can be sped up with the help of automated methods.

Both pseudonymization and anonymization can be construed of as two-stage processes, with the first one — PI detection and labeling, necessary in order to know what elements need to be handled — being shared by both, and succeeded by, respectively, the removal or replacement of detected spans.[3] Automatic PI detection and labeling is a task closely related to Named Entity Recognition and Classification (NER, NERC) (Lison et al., 2021). The key distinction is that while the overlap between PI and Named Entities (NEs) is large,

[1] `https://spraakbanken.gu.se/en/analyses/sbx-swe-pi_detection-sparv` (DOI) `https://vlo.clarin.eu/record/Other_47_doi_10_23695_6wp0_ds77.xml?1&count=3&index=0&q=sparv-plugin` (VLO)

[2] In this paper this term is used to refer to both information that on its own or in combination with other pieces of information can be used to re-identify a natural person, i.e. Personally Identifiable Information (PII), and sensitive information (e.g. sexuality, religious beliefs).

[3] There exist approaches where this distinction is not made, e.g. seq2seq or LLM prompting to directly return a sanitized text, cf. Yermilov et al. (2023).

Proceedings of the First Workshop NLP and Language Models for Digital Humanities associated with RANLP 2025, pages 155–163, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-106-6-015

not all PIs are NEs and not all NEs are personal in nature; whether a piece of information in a text is personal is highly dependent on the context.

The approaches to PI detection and labeling range from rule-based systems (Accorsi et al., 2012; Dalianis, 2019; Blokland et al., 2020; Volodina et al., 2020) through machine learning approaches using e.g. Conditional Random Fields (Berg and Dalianis, 2019, 2020; Adams et al., 2019; Eder et al., 2020), Recurrent Neural Networks (Adams et al., 2019; Eder et al., 2020; Jensen et al., 2021; López-García et al., 2023), or Transformer-based classifiers (Johnson et al., 2020; Jensen et al., 2021; Eder et al., 2022; Meaney et al., 2022; López-García et al., 2023; Ngo et al., 2024; Szawerna et al., 2024, 2025), to Large Language Model (LLM) prompting (Yang et al., 2023; Ilinykh and Szawerna, 2025) or knowledge distillation from LLMs (Deußer et al., 2025), and combined approaches (Jensen et al., 2021; Eder et al., 2022; Cabrera-Diego and Gheewala, 2024). Notably, the ability to conduct automated PI detection locally is advantageous, since sending potentially sensitive data to external tools or APIs increases the chance of a security breach and information leakage. Ease of use also plays a role, since the more complicated a tool is to set up or use, the smaller its userbase is likely to become.

In this paper we present a new, flexible plugin SBX-PI for Personal Information detection and labeling in Swedish for the SPARV text annotation pipeline, alongside a novel visualization tool, (IM)PERSONAL DATA, intended for analyzing and comparing the performance of such systems. The tools stand out in two ways: on the one hand, SBX-PI embeds ethics into SPARV through addressing the g the 'privacy by design' requirement imposed by the GDPR (Official Journal of the European Union, 2016); on the other hand, (IM)PERSONAL DATA is the only tool known to us allowing the users to assess how much of the research data will be altered once anonymization or pseudonymization are applied, making it possible to assess the value and validity of the data after the applied manipulations. We showcase both tools on a sample text[4] and compare the performance of the PI detection plugin to that of the commercially available tool MICROSOFT PRESIDIO,[5], the web-based tool for Swedish texts HB DEID (Berg and Dalianis, 2019, 2021),[6] and the results of prompting GEMMA 2 9B (Gemma Team et al., 2024),[7] with the help of the visualization tool. GEMMA 2 9B has previously been used by Ilinykh and Szawerna (2025) to detect and label personal information and performed best out of the tested models; due to its size it could potentially be run locally. We discuss the differences in performance and the advantages and disadvantages of the aforementioned approaches. Finally, we present plans and suggestions for further development of both of our tools.

2 SPARV plugin

SPARV (v5.3.0, Hammarstedt et al., 2022)[8] is a Python-based modular command line tool for text annotation designed primarily for Swedish, created and maintained by Språkbanken Text. It can be run locally and is designed to handle importing the data, annotating it, and exporting it. The choice of formats and annotations is controlled using a corpus configuration file. SPARV's design makes it also very easy to extend it with new modules or plugins, which has enabled the addition of Personal Information detection and labeling, meaning that this task can be performed together with other kinds of annotation, e.g. part-of-speech tagging.

2.1 System Design

Our plugin[9] makes use of six PI detection classifiers for Swedish (Szawerna et al., 2025) based on `KB/bert-base-swedish-cased` (Devlin et al., 2019; Malmsten et al., 2020), hosted on Språkbanken Text's HuggingFace page.[10] The models' performance reported in Szawerna et al. (2025) is shown in Table 1. The models differ in terms of the tags that they can assign to the detected spans, as outlined in Table 2. It is very important to highlight that, as per their HuggingFace model cards, these models "[...] perform best on [...] second-language learner essays," the type of texts that they were trained on. By not including the models in the plugin itself but accessing them

[4] While we acknowledge that evaluation measures would be valuable, there is no openly available dataset for PI detection in Swedish. The performance of the models used in our tool on their test set was reported in Szawerna et al. (2025).

[5] https://microsoft.github.io/presidio/

[6] https://hbdeid.dsv.su.se/

[7] https://huggingface.co/google/gemma-2-9b

[8] https://spraakbanken.gu.se/sparv/

[9] https://github.com/spraakbanken/sparv-sbx-pi-detection

[10] https://huggingface.co/sbx

Model	F2
detailed_iob	0.519 ± 0.085
detailed	0.558 ± 0.063
general_iob	0.720 ± 0.054
general	0.763 ± 0.059
basic_iob	0.800 ± 0.045
basic	0.824 ± 0.038

Table 1: Mean results ± standard deviation for each type of model, courtesy of the authors (Szawerna et al., 2025).

via HuggingFace, we make it possible to access newer versions of the same models, should they ever be released. It also makes it relatively simple to expand the plugin with additional models by modifying very little of the code.

We follow the general recommended structure for SPARV plugins.[11] The code is accompanied by a number of required or recommended files specifying the functionality or behavior of the plugin for both SPARV itself and the user. The plugin's requirements are `Sparv 5.0` or higher, `Transformers 4.51.3` or higher (Wolf et al., 2020), and `PyTorch` (Ansel et al., 2024).

The core of the plugin's functionality lies in the `pi_detection.py` file, which defines the functions called when the annotations provided by this plugin are requested by the user. In such a case the input is first tokenized at word level using a user-defined or SPARV's default tokenizer. The appropriate classifier model and corresponding tokenizer are loaded in according to the corpus configuration using `Transformers`. Since BERT-based models use sub-word tokenization and have a maximum input length, each input text is chunked if it were to exceed the length of 512 sub-word tokens, with the boundaries following the word-level tokenization. Next, predictions are obtained from the model for each chunk. Finally, these are mapped back to the word-level tokens. In cases where multiple sub-word tokens constituting one word have received different tags, the following heuristics are applied: i) if at least one sub-word token is tagged as personal information, the entire word is tagged as that and ii) if two different personal information tags were assigned to two sub-word tokens of one word, the one closest to the beginning of the word is selected, as we consider that to be more likely to be the meaning-bearing element of the word. These tags are then forwarded to the export method defined by the user in the corpus configuration.

We also provide a sample corpus with our plugin, which consists of two text files, one with an example essay and one intended for the user to edit, alongside a corpus configuration file.

2.2 Functionality

Once the plugin is installed following the instructions that come with it, the user can request the PI annotation in the `config.yaml` configuration file for their corpus. First of all, in `annotations`, one has to specify the annotation type as `<token>:sbx_pi_detection.pi`. Next, the specific tagset (and, consequently, classifier) has to be specified, e.g. `pi_detection: general`. The names of the available tagsets in the configuration are the same as in Table 2, and more detailed user instructions are included in the plugin's README file. Both the input and output format for the data are independent from our plugin and depend on the user choice defined in the corpus configuration.

A key advantage that our plugin has is its integration into SPARV, as that allows for other types of annotation to be carried out simultaneously, according to what is defined in the corpus configuration. This also makes it easy for current SPARV users to incorporate PI annotation in their workflow.

3 (Im)Personal Data visualization

In order to visualize the system output, a custom visualization[12] was commissioned with INFRAVIS[13], the Swedish National Research Infrastructure for Data Visualization.

3.1 System Design

The interface is realized as a Vue 3 frontend and builds on two modules. The first module uses Texty (Nualart and Pérez-Montoro, 2013), "an icon that represents the physical distribution of keywords of a text as a flat image," to give an overall impression of the distribution of labels in the text (see Figure 1). The second module aligns the input texts on the word level and allows for the comparison of labels across different methods (see Figure 2).

In order to allow for new text additions, the interface is written in such a way as to adapt to new

[11] `https://spraakbanken.gu.se/sparv/developers-guide/writing-sparv-plugins/`

[12] `https://github.com/spraakbanken/impersonaldata`

[13] `infravis.se`

Model	Tags
detailed	O, firstname_male, firstname_female, firstname_unknown, initials, middlename, surname, school, work, other_institution, area, city, geo, country, place, region, street_nr, zip_code, transport_name, transport_nr, age_digits, age_string, date_digits, day, month_digit, month_word, year, phone_nr, email, url, personid_nr, account_nr, license_nr, other_nr_seq, extra, prof, edu, fam, sensitive
detailed_iob	O, B-firstname_male, I-firstname_male, ...
general	O, personal_name, institution, geographic, transportation, age, date, other
general_iob	O, B-personal_name, I-personal_name, B-institution, ...
basic	O, S
basic_iob	O, B, I

Table 2: Tagsets in the models available in the plugin. O appears in all of them and marks the non-PI tokens. IOB models have the same semantic categories as their base versions, but with the addition of marking the beginning and inside of the span. See Megyesi et al. (2018) and Szawerna et al. (2025) for more details on the tagsets and models.

data automatically. New data is added by adding the annotated texts in a specific folder, running the Texty Python script, and, finally, running a custom script that calculates word alignment and copies all the relevant data to the frontend folder.

3.2 Functionality

In the interface, the user can choose one of the preselected texts, which is then displayed. The user can then visualize the high-level label distribution of different methods with Textys, and inspect the labels more closely in detail view.

The current interface loads static pre-computed files, but this behavior may be changed in the future — adding a proper backend could allow users to test their own texts dynamically.

4 Case study

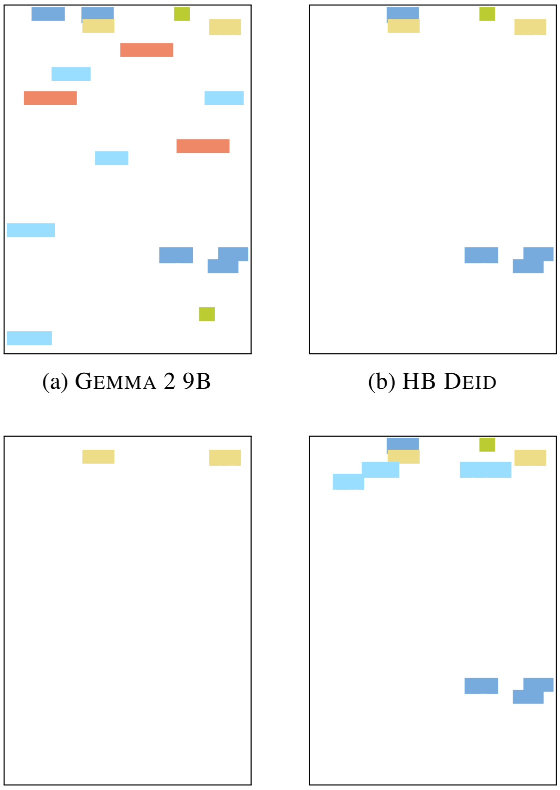
We use a sample text in order to better illustrate the performance of our plugin and the visualization tool. The text is a fictive personal story in Swedish, i.e. it contains information that would be personal if it referred to any natural person, and is structured like a personal story.[14]

We obtained PI annotations from four different tools: (a) GEMMA 2 9B (Gemma Team et al., 2024), (b) HB DEID (Berg and Dalianis, 2019, 2021), (c) MICROSOFT PRESIDIO, (d) our plugin. In the case of MICROSOFT PRESIDIO and HB DEID we mapped these tools' tagsets to the one used by the general model in our plugin, and we instructed GEMMA 2 9B to follow the same type of annotation (one-shot prompting adapted from Ilinykh and Szawerna (2025) with alterations for a different tagset and enforcing a JSON output). Importantly, using MICROSOFT PRESIDIO for a language other than English requires additional coding to enable the use of NER models for the language in question. While MICROSOFT PRESIDIO can be further customized (e.g. by adding rules), we opted for trying to use it as "out of the box" as possible; the same is true in the case of prompts for GEMMA 2 9B, which we did not engineer beyond including our tagset in the aforementioned prompt structure. We unified the output formats to follow the requirements for inputs to the visualization tool.

Figure 1 shows the generated Texty images for the annotated text. It is immediately visible that GEMMA 2 9B predicts the most diverse categories, followed by our plugin; upon closer inspection, though, it can be noted that the models disagree on which entities should be marked as other. HB DEID only detects three categories, and MICROSOFT PRESIDIO just one; in the web interface of the visualization tool these are identified as personal_name, age and geographic for the former and only geographic for the latter.

The detailed view — shown in Figure 2 — is required to properly assess the performance of the tools against each other, as that is where the anno-

[14] The text can be found here: https://github.com/mormor-karl/annotating-PI-with-SPARV



(a) GEMMA 2 9B (b) HB DEID

(c) MICROSOFT PRESIDIO (d) SPARV plugin

Figure 1: Texty visualizations for the sample essay for each annotation tool. Colors represent PI of different categories across the running text.

Text	gemma_label	hbdeid_label	presidio_label	sparv_label
namn	personal_name			
Sonja	personal_name	personal_name		personal_name
29	age	age		age
Polen	geographic	geographic	geographic	geographic
Visby	geographic	geographic	geographic	geographic
polska				other
engelska				other
tyska				other
förskolan	institution			
kl.6.00	other			
förskolan	institution			
kl.7.00	other			
förskolan	institution			
11.30	other			
16-tiden	other			
Kathy	personal_name	personal_name		personal_name
Anna	personal_name	personal_name		personal_name
Måns	personal_name	personal_name		personal_name
23	age			
midnatt	other			

Figure 2: Detailed view of annotation differences from the visualization tool.

tated tokens are displayed. All of the tools agree on the annotation of the `geographic` entities, and all but MICROSOFT PRESIDIO correctly identify the ages and personal names present in the text; here GEMMA 2 9B returns two false positives, marking *namn* 'name' and *23* (which in the context clearly refers to a time). Further differences between our Sparv plugin and GEMMA 2 9B concern the `institution` and `other` categories. The Sparv plugin is the only one to mark the foreign languages *polska* 'Polish', *engelska* 'English', and *tyska* 'German' as `other`, which is an expected behavior. GEMMA 2 9B instead assigns this tag to four time points (including *midnatt* 'midnight') marking the daily routine of the person in the text; while this was not overtly stated in the prompt, it is interesting to see the LLM make this generalization, as this type of information could in some cases lead to re-identification. Finally, GEMMA 2 9B also tags the three mentions of *förskolan* 'the kindergarten' as `institution`. This is another justified generalization on the LLM's part, as the Sparv plugin tends to only mark specific institutions with that tag due to the way it was used in the training data (i.e. these would likely have been tagged by the plugin if they mentioned the name of the kindergarten). It is, however, worth pointing out that a big part of the text describes the activities at the kindergarten, meaning that if the type of workplace were to lead to reidentification, more than just 'kindergarten' would have to be handled. Interestingly, none of the models marked *förskollärare* 'kindergarten teacher' as `other` (which is meant to include professions).

Overall, the Sparv plugin and GEMMA 2 9B are the clear forerunners for this text. Sparv is somewhat more conservative and does not make the same kinds of generalizations as GEMMA 2 9B, but it does not return false positives either. Both of these tools potentially miss some additional personal information, which highlights the importance of using these to assist de-identification, but not completely automatize it, as it is a high-stakes task and no existing tool can guarantee 100% accuracy.

Another relevant point for comparison is the time it takes the tools to annotate the data. Table 3 shows the times we have measured, although they are not fully comparable. HB DEID is a web-based demonstrator and while the annotation seemed instantaneous, it then had to be manually transfered into a machine-readable format. While Sparv and MI-

Tool	Time
Sparv plugin	15s
MICROSOFT PRESIDIO	10s
HB DEID	-
GEMMA 2 9B	59s

Table 3: A comparison of time it takes to run the different annotation tools.

CROSOFT PRESIDIO were run on one of our local machines, GEMMA 2 9B was run on a server with two GeForce RTX 2080 Ti GPUs. With that in mind, our plugin seems to strike a good balance between performance, speed, and hardware requirements, and therefore a clear winner when it comes to sustainability and eco-friendliness, as the other well-performing tool takes nearly four times longer on better hardware.

5 Discussion and conclusions

We have presented SBX-PI, a new tool for personal information detection and labeling for Swedish texts which empowers researchers in Digital Humanities, linguistics, Natural Language Processing and other research domains dependent on access to language data. SBX-PI functions as a plugin for the text annotation tool SPARV. As such, its functionality can be combined with a range of other annotations. Our plugin is relatively lightweight and fast for its performance. It does not require extensive programming knowledge to use, but with some knowledge of Python it can be easily modified. Such modifications may include allowing the model to use different PI classifier models, potentially extending its use beyond Swedish. Additionally, framing this tool as a SPARV plugin makes it easy to combine it in the future with plugins that would carry out the second stage of anonymization or pseudonymization (i.e. remove or replace the personal information in the text), effectively completing the de-identification pipeline, which we consider as our future goals.

We have also introduced the (IM)PERSONAL DATA visualization tool which can be used to illustrate and qualitatively analyze the output of our plugin and other PI detection and labeling models, as well as to visualize the extent of research data that needs to be manipulated before it is shared with other researchers. This interface, which currently operates by displaying static files, has the potential to be expanded with a backend to display the performance of specific models on the go.

We have also performed a case study to demonstrate our plugin's performance and the usefulness of the (Im)Personal Data visualization tool. We have shown that our plugin performs noticeably better on the text we tested than one of the most popular openly available tools, MICROSOFT PRESIDIO. It can detect a wider range of personal information than the HB DEID tool, and seems to be less prone to false positives than the LLM GEMMA 2 9B, in comparison to which it is also much faster and less resource-intensive.

We believe that our plugin and the visualization tool can contribute to language resource construction by facilitating the de-identification procedures, indirectly contributing to research in a variety of fields which rely on such data, including but not limited to linguistics, Digital Humanities, or Natural Language Processing. At the same time, our tools will hopefully enable further research on NLP methods for anonymization and pseudonymization.

The future directions include, among others, (1) developing models for pseudonymization step (i.e. for the generation of replacement equivalents for the detected spans), and (2) developing solutions that would offer an option for customizing which subset of PI categories to replace, thus protecting research data from being over-manipulated. The two future steps address the duality of the problem: ethical requirement to protect people in the data versus the need of valid research data that is as close to the original as it is legally and ethically allowed.

Acknowledgments

This work was possible thanks to the funding of several grants from the Swedish Research Council. It has been funded by Språkbanken – jointly funded by its 10 partner institutions and the Swedish Research Council (2025–2028; project id 2023-00161). It has also been funded by the research environment project *Grandma Karl is 27 years old: Automatic pseudonymization of research data* (2023-2029, funding number 2022-02311). The authors acknowledge support from InfraVis for providing application expertise for visualization through the Swedish Research Council grant 2021-00181.

References

Pierre Accorsi, Namrata Patel, Cédric Lopez, Rachel Panckhurst, and Mathieu Roche. 2012. Seek&Hide: Anonymising a French SMS corpus using natural language processing techniques. *Linguisticae Investigationes*, 35:163–180.

Allison Adams, Eric Aili, Daniel Aioanei, Rebecca Jonsson, Lina Mickelsson, Dagmar Mikmekova, Fred Roberts, Javier Fernandez Valencia, and Roger Wechsler. 2019. AnonyMate: A toolkit for anonymizing unstructured chat data. In *Proceedings of the Workshop on NLP and Pseudonymisation*, pages 1–7, Turku, Finland. Linköping Electronic Press.

Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalambarkar, Laurent Kirsch, Michael Lazos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, CK Luk, Bert Maher, Yunjie Pan, Christian Puhrsch, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Michael Suo, Phil Tillet, Eikan Wang, Xiaodong Wang, William Wen, Shunting Zhang, Xu Zhao, Keren Zhou, Richard Zou, Ajit Mathews, Gregory Chanan, Peng Wu, and Soumith Chintala. 2024. PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. In *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*. ACM.

Hanna Berg and Hercules Dalianis. 2019. Augmenting a de-identification system for Swedish clinical text using open resources and deep learning. In *Proceedings of the Workshop on NLP and Pseudonymisation*, pages 8–15, Turku, Finland. Linköping Electronic Press.

Hanna Berg and Hercules Dalianis. 2020. A semi-supervised approach for de-identification of Swedish clinical text. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4444–4450, Marseille, France. European Language Resources Association.

Hanna Berg and Hercules Dalianis. 2021. HB Deid - HB de-identification tool demonstrator. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 467–471, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Rogier Blokland, Niko Partanen, and Michael Rießler. 2020. A pseudonymisation method for language documentation corpora: An experiment with spoken Komi. In *Proceedings of the Sixth International Workshop on Computational Linguistics of Uralic Languages*, pages 1–8, Wien, Austria. Association for Computational Linguistics.

Luis Adrián Cabrera-Diego and Akshita Gheewala. 2024. PSILENCE: A pseudonymization tool for international law. In *Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024)*, pages 25–36, St. Julian's, Malta. Association for Computational Linguistics.

Hercules Dalianis. 2019. Pseudonymisation of Swedish electronic patient records using a rule-based approach. In *Proceedings of the Workshop on NLP and Pseudonymisation*, pages 16–23, Turku, Finland. Linköping Electronic Press.

Tobias Deußer, Max Hahnbück, Tobias Uelwer, Cong Zhao, Christian Bauckhage, and Rafet Sifa. 2025. Resource-efficient anonymization of textual data via knowledge distillation from large language models. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 243–250, Abu Dhabi, UAE. 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.

Elisabeth Eder, Ulrike Krieg-Holz, and Udo Hahn. 2020. CodE alltag 2.0 — a pseudonymized German-language email corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4466–4477, Marseille, France. European Language Resources Association.

Elisabeth Eder, Michael Wiegand, Ulrike Krieg-Holz, and Udo Hahn. 2022. "beste grüße, maria meyer" — pseudonymization of privacy-sensitive information in emails. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 741–752, Marseille, France. European Language Resources Association.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A.

Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size.

Martin Hammarstedt, Anne Schumacher, Lars Borin, and Markus Forsberg. 2022. Sparv 5 user manual. Technical report, Institutionen för svenska, flerspråkighet och språkteknologi, Göteborgs universitet, Göteborg.

Nikolai Ilinykh and Maria Irena Szawerna. 2025. "I need more context and an English translation": Analysing how LLMs identify personal information in Komi, Polish, and English. In *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025)*, pages 165–178, Tallinn, Estonia. University of Tartu Library, Estonia.

Kristian Nørgaard Jensen, Mike Zhang, and Barbara Plank. 2021. De-identification of privacy-related entities in job postings. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 210–221, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Alistair E. W. Johnson, Lucas Bulgarelli, and Tom J. Pollard. 2020. Deidentification of free-text medical records using pre-trained bidirectional transformers. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, CHIL '20, pages 214–221, New York, NY, USA. Association for Computing Machinery.

Pierre Lison, Ildikó Pilán, David Sanchez, Montserrat Batet, and Lilja Øvrelid. 2021. Anonymisation models for text data: State of the art, challenges and future directions. 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 4188–4203, Online. Association for Computational Linguistics.

Guillermo López-García, Francisco J. Moreno-Barea, Héctor Mesa, José M. Jerez, Nuria Ribelles, Emilio Alba, and Francisco J. Veredas. 2023. Named Entity Recognition for De-identifying Real-World Health Records in Spanish. In *Computational Science – ICCS 2023*, pages 228–242, Cham. Springer Nature Switzerland.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with Words at the National Library of Sweden – Making a Swedish BERT.

Christopher Meaney, Wali Hakimpour, Sumeet Kalia, and Rahim Moineddin. 2022. A Comparative Evaluation Of Transformer Models For De-Identification Of Clinical Text Data. ArXiv:2204.07056 [cs, stat].

Beáta Megyesi, Lena Granstedt, Sofia Johansson, Julia Prentice, Dan Rosén, Carl-Johan Schenström, Gunlög Sundberg, Mats Wirén, and Elena Volodina. 2018. Learner corpus anonymization in the age of GDPR: Insights from the creation of a learner corpus of Swedish. In *Proceedings of the 7th workshop on NLP for Computer Assisted Language Learning*, pages 47–56, Stockholm, Sweden. LiU Electronic Press.

Phuong Ngo, Miguel Tejedor, Therese Olsen Svenning, Taridzo Chomutare, Andrius Budrionis, and Hercules Dalianis. 2024. Deidentifying a Norwegian clinical corpus - an effort to create a privacy-preserving Norwegian large clinical language model. In *Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024)*, pages 37–43, St. Julian's, Malta. Association for Computational Linguistics.

Jaume Nualart and Mario Pérez-Montoro. 2013. Texty, a visualization tool to aid selection of texts from search outputs. *Information Research*, 18(2).

Official Journal of the European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). *Official Journal*, (Document 02016R0679-20160504).

Maria Irena Szawerna, Simon Dobnik, Ricardo Muñoz Sánchez, Therese Lindström Tiedemann, and Elena Volodina. 2024. Detecting personal identifiable information in Swedish learner essays. In *Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024)*, pages 54–63, St. Julian's, Malta. Association for Computational Linguistics.

Maria Irena Szawerna, Simon Dobnik, Ricardo Muñoz Sánchez, and Elena Volodina. 2025. The devil's in the details: the detailedness of classes influences personal information detection and labeling. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 697–708, Tallinn, Estonia. University of Tartu Library.

Elena Volodina, Yousuf Ali Mohammed, Sandra Derbring, Arild Matsson, and Beata Megyesi. 2020. Towards privacy by design in learner corpora research: A case of on-the-fly pseudonymization of Swedish learner essays. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 357–369, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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, Mariama Drame, Quentin Lhoest, and Alexander Rush. 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.

Jianliang Yang, Xiya Zhang, Kai Liang, and Yuenan Liu. 2023. Exploring the application of large language models in detecting and protecting personally identifiable information in archival data: A comprehensive study*. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2116–2123.

Oleksandr Yermilov, Vipul Raheja, and Artem Chernodub. 2023. Privacy- and utility-preserving NLP with anonymized data: A case study of pseudonymization. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 232–241, Toronto, Canada. Association for Computational Linguistics.

Can LLMs Help Sun Wukong in his Journey to the West? A Case Study of Language Models in Video Game Localization

Xiaojing Zhao Han Xu Huacheng Song Emmanuele Chersoni Chu-Ren Huang
Department of Language Science and Technology, The Hong Kong Polytechnic University
{xiaojing.zhao, huacheng.song}@connect.polyu.hk
{han12.xu, emmanuele.chersoni, churen.huang}@polyu.edu.hk

Abstract

Large language models (LLMs) have demonstrated increasing proficiency in general-purpose translation, yet their effectiveness in creative domains such as game localization remains underexplored. This study focuses on the role of LLMs in game localization from both linguistic quality and sociocultural adequacy through a case study of the video game *Black Myth: Wukong*.

Results indicate that LLMs demonstrate adequate competence in accuracy and fluency, achieving performance comparable to human translators. However, limitations remain in the literal translation of culture-specific terms and offensive language. Human oversight is required to ensure nuanced cultural authenticity and sensitivity. Insights from human evaluations also suggest that current automatic metrics and the Multidimensional Quality Metrics framework may be inadequate for evaluating creative translation. Finally, varying human preferences in localization pose a learning ambiguity for LLMs to perform optimal translation strategies. The findings highlight the potential and shortcomings of LLMs to serve as collaborative tools in game localization workflows. Data are available at `https://github.com/zcocozz/wukong-localization`.

1 Introduction

Recent advances in large language models (LLMs) have significantly expanded the frontiers of machine translation (MT), achieving state-of-the-art performance across technical and literary domains (Hendy et al., 2023; Jiao et al., 2023; Wang et al., 2023). Unlike conventional MT systems, which struggle with idiomatic expressions and context-dependent scenarios, LLMs demonstrate potential in handling complex linguistic phenomena, including metaphor and idioms (Stowe et al., 2022; Tang et al., 2024; Yue et al., 2024). Moreover, recent studies suggest that LLMs can match the performance of junior human translators, signaling progress toward human parity in MT (Yan et al., 2024).

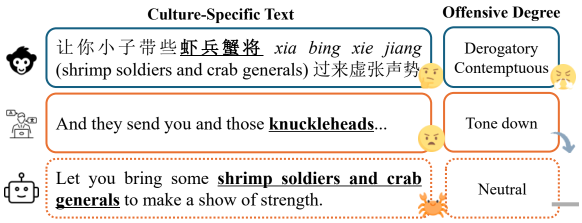


Figure 1: Human and LLM translations for cultural and offensive language in *Black Myth: Wukong*.

Despite these advancements, LLMs' potential for **video game localization** remains underexplored. The massive AAA game industry requires rapid translation of high-budget games from publishers into multiple languages for simultaneous global releases. While game localization teams face constraints including time and resources, the multilingual capabilities of LLMs offer a promising solution. Moreover, initiatives like Sony's China Hero Project aim to introduce more Chinese games into the global market[1], intensifying the need for culture-adapted multilingual translations. These games, rich in culture- and history-specific imagery, present a unique challenge of balancing preserving original cultural nuances with reshaping content to resonate with Western audiences.

This situation underscores a core dilemma in game localization—maintaining fidelity to the source material while adapting it appropriately for the target market. On one hand, successful localization requires preserving the "look and feel" of the original version by retaining key elements that define the game (O'Hagan and Mangiron, 2006). On the other hand, it demands sociocultural adaptation

[1] `https://www.playstation.com/en-us/china-hero-project/`

Proceedings of the First Workshop NLP and Language Models for Digital Humanities associated with RANLP 2025, pages 164–173, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-106-6-016

to meet target market expectations and avoid cultural sensitivities, with offensive language emerging as a prominent concern (Al-Batineh, 2021). This tension makes ensuring authentic gaming experiences across diverse cultural contexts inherently challenging. As illustrated in Figure 1, game localization needs to ensure sociocultural adequacy, including culture-specific terms and offensive language. Games embed cultural references, slang, and humor that demand context-aware translation. For instance, human translators adapt the mythological term "虾兵蟹将 *xia bing xie jiang*" into a colloquial equivalent "knuckleheads", whereas LLMs produce literal translation like "shrimp soldiers and crab generals", confusing audiences unfamiliar with the source culture. Such equivalents may also diminish original offensiveness by stripping culturally charged connotations.

To investigate the potential of LLMs for video game localization, this work presents a systematic evaluation combining automatic metrics with human assessments, using the recent *Black Myth: Wukong* game as a case study. We examine both the linguistic quality (accuracy, fluency) and sociocultural adequacy (cultural appropriateness, offensiveness rating) of LLM translations.

Our findings reveal mixed capabilities of LLMs in game localization. LLMs excel in linguistic quality, delivering satisfactory accuracy and fluency, yet they struggle with the cultural adaptation of culture-specific terms and offensive language. Human evaluations further suggest that automatic metrics and Multidimensional Quality Metrics (MQM) standards may not be appropriate for evaluating creative translation, while diverse human preferences in localization also pose learning ambiguity for LLMs to identify optimal translation strategies. The evolving capabilities of LLMs suggest their potential as collaborative partners in game localization workflows, though human post-editing remains essential for maintaining cultural authenticity and addressing cultural sensitivity. To our knowledge, this study represents the first systematic evaluation of LLMs in video game localization.

2 Related Work

2.1 LLMs for Translation

Recent studies demonstrate that LLMs can rival or even surpass traditional MT systems, achieving performance that nears human parity in basic tasks, such as GPT-4 showing ability competitive with commercial MT systems (Jiao et al., 2023). Furthermore, evaluations across 102 languages reveal steady improvements in high-resource languages, although challenges still remain in low-resource contexts (Hendy et al., 2023; Zhu et al., 2024).

Beyond general-purpose translation, LLMs have achieved notable advances in specialized professional and literary domains where traditional MT systems typically struggle. In legal translation, GPT-4 produces contextually accurate outputs comparable to human performance (Briva-Iglesias et al., 2024). This capability extends to culturally complex tasks, with LLMs successfully navigating context-dependent challenges including idiomatic expressions and poetry translation (Chen et al., 2024; Tang et al., 2024; Yao et al., 2024).

Building on this progress, LLMs present promising applications for video game localization. However, despite their great potential, research investigating LLM performance in video gaming localization remains limited. While Moreno García and Mangiron (2024) examined GPT-4's translation of *Pokémon* terminology and found that the model could successfully implement creative translations, the scope and scale need to be enlarged to encompass a broader range of linguistic contexts. Meanwhile, human evaluation is essential to assess the fine-grained quality of LLM translations.

2.2 Game Localization

The emergence of LLMs has unveiled novel opportunities in video game localization. Localization projects typically operate under severe time and budget constraints that can undermine creative adaptation and player experience (O'Hagan and Chandler, 2016). These limitations create a pressing need for cost-effective and efficient translation solutions. The ongoing progress of LLMs' contextual reasoning and cultural adaptation capabilities presents promising opportunities for improving efficiency and translation quality in localization.

The core of successful game localization is to deliver authentic player experiences. It involves adapting in-game texts, audio, and visual elements to match the target language and cultural context while preserving the narrative intent. Research reveals a strong player preference for localization that preserves original cultural elements rather than adapting them to local norms, as cultural authenticity is essential to player engagement and gaming experience (Costales, 2016; Ellefsen and Bernal-Merino, 2018; Khoshsaligheh et al., 2020; Wu and Chen, 2020). However, retaining culture- and history-specific elements, such as idioms, metaphors, and slang, remains a persistent challenge due to the limited translatability across sociolinguistic contexts.

Game localization also faces significant challenges in delivering culturally sensitive content, particularly in regions with stringent regulatory or sociocultural norms. For instance, Arabic-localized games frequently undergo systematic sanitization of profanity, nudity, and alcohol through omission, substitution, or euphemistic translation (Mahasneh and Abu Kishek, 2018; Al-Batineh, 2021). While these practices comply with censorship requirements and cultural expectations, they frequently lead to a loss of semantic or pragmatic nuances.

3 Methodology

Data We select the blockbuster video game *Black Myth: Wukong* as our data source due to its unique cultural and linguistic representativeness. The game is adapted from the 16th-century Chinese classic *Journey to the West*, blending poetic allusions, religious themes, and vernacular dialogue. Its dual mission—promoting traditional Chinese culture while advancing global gaming experience—creates salient localization challenges: preserving culturally embedded idioms and folklore while ensuring accessibility for international players. Given the limited familiarity of Western audiences with ancient Chinese cultures, the localization of such a product poses significant complexity.

Although this game supports official subtitles in 12 languages on the interfaces, its voice dialogues are available only in Chinese and English. Accordingly, we focus on Chinese-to-English translations. We transcribe subtitles from official cutscenes and publicly available videos, followed by manual proofreading of all content. The resulting parallel corpus comprises 2,259 sentence-pairs, capturing diverse narrative elements: main and side request dialogues, chapter-ending narratives rich in cultural metaphors, and song lyrics. Although the corpus of the in-game subtitles may be modest in size, *Black Myth: Wukong* is still unique as a Chinese AAA game with a special historical background, and its stylistic spectrum, spanning from story-based in-game banter to literary prose, makes it an ideal test case for investigating the culture-aware translation capabilities of LLMs.

Pipeline To benchmark LLM performance from coarse-grained to fine-grained perspectives, we regard the collected official human translations in English from the game developer as gold references. By comparing LLM outputs with human references, we conduct a two-stage evaluation on mainstream LLMs: 1) first, we employ multiple automatic metrics to evaluate diverse LLMs across strategically varied prompts using a randomly sampled subset from our *Black Myth: Wukong* parallel corpus. The top-performing LLM-prompt configurations identified in this phase serve as the basis for subsequent human evaluation, where 2) we compare the translations generated by the optimally prompted LLM against human gold references across multidimensional linguistic quality (accuracy, fluency) and sociocultural adequacy (audience appropriateness, offensiveness handling) based on subsets of sampled cases as well as manually extracted offensive instances. Further details for each evaluation stage are presented in the following.

3.1 Automatic Evaluation

At this stage, we generate trial translations for 101 randomly sampled sentences from the complete corpus first using four open-source LLMs with four zero-shot prompting strategies, and then evaluate them with six automatic metrics, aiming to build a global view of LLMs' performance in localization and identify the optimal model-prompt pairing for subsequent human evaluation.

Models Our model selection comprises Llama3(-8B) (Grattafiori et al., 2024), TowerBase(-7B) (Alves et al., 2024), Qwen2(-7B) (Yang et al., 2024), and DeepSeek-LLM(-7B) (Bi et al., 2024), chosen for their known abilities in multilingual processing tasks. We choose all LLMs with around 7/8 billion parameters to compare the performance of different architectures at a similar parameter size.

Prompts Prior work in translation prompt engineering demonstrated that direct and minimalist prompts outperform complex formulations and achieve competitive performance (Jiao et al., 2023; Yan et al., 2024). In light of this, we develop four tailored and straightforward prompts addressing main localization requirements through different strategies: concise, role-playing, adaptive, and authentic, as detailed in Figure 2.

Metrics Using official English subtitles as gold-standard references, we automatically evaluate

Concise
Translate the following sentence from Chinese into English.
Chinese: {text}
English:

Role-playing
As a professional game localization expert for *Black Myth: Wukong*, translate the following sentence from Chinese into English.
Chinese: {text}
English:

Adaptive
Translate the following sentence from Chinese to natural, fluent and engaging English. Adapt cultural references and idioms for international audiences while retaining the original essence.
Chinese: {text}
English:

Authentic
Translate the following sentence from Chinese into English. Retain the original mythological, Buddhist, and cultural elements.
Chinese: {text}
English:

Figure 2: Different prompt types and specifications, where {text} represents the input source sentence.

models with six metrics. BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), BERTScore (Zhang et al., 2020), COMET (Rei et al., 2022), XCOMET (Guerreiro et al., 2024), and XCOMET-QE (Guerreiro et al., 2024). They capture distinct performance aspects based on different principles: 1) string-overlap-based metrics (BLEU and ROUGE) emphasize surface-level equivalence; 2) among neural-network-based metrics (BERTScore, COMET, XCOMET, and XCOMET-QE), BERTScore uses the cosine similarity between token embeddings of candidate and reference texts, while the COMET-series metrics evaluate both meaning and form, as they are fine-tuned to predict human quality scores for translations. Among them, XCOMET-QE is an exception, as it assesses quality without any gold references by cross-lingually comparing source and target texts.

3.2 Human Evaluation

To further explore LLMs' culture-related capacity in game localization, we then conduct a human evaluation comparing the top-performing prompted LLM with gold human translations from aspects of linguistic quality in line with the MQM framework and sociocultural adequacy in terms of offensive language handling.

Multidimensional Translation Quality We evaluate translation quality using the MQM framework (Lommel et al., 2014), a comprehensive error typology enabling granular analysis of translation errors. Adopting the MQM core typology, we prioritize three key dimensions for localization: **accuracy** (semantic completeness and faithfulness), **fluency** (syntactic and grammatical correctness), and **audience appropriateness** (cross-cultural validity).

Offensive Language Annotation Beyond the MQM framework, we specifically assess sociocultural adequacy through offensive language handling. Offensive content was categorized along eight dimensions, spanning from explicit insult to culture-specific connotations. Two native Chinese speakers (C1 English-proficient video game players) independently labeled offensive expressions, achieving moderate agreement (Cohen's $\kappa = 0.48$). Discrepancies were resolved through consensus discussions with a third annotator, and 430 offensive cases were identified in the whole corpus.

Evaluation Protocol For both human evaluation tasks, ten postgraduate students specializing in Translation Studies were recruited as raters. They were paired into five groups, with each pair evaluating the same translations to ensure double annotation. Prior to the assessment, all these annotators received comprehensive training covering guidelines for the MQM translation error typology and offensive language classification, as well as gaining contextual familiarity with gameplay narratives through story walkthrough videos and detailed character biographies. Target human and machine translations were assessed on:

- **Translation Quality Scoring:** A five-point Likert scale for measuring accuracy, fluency, and audience appropriateness (from 1 = Poor to 5 = Excellent);
- **Offensive Language Rating:** A three-degree classification comparing the target translations to the source texts (less offensive, neutral, or more offensive).

To avoid the preference for human translations and biases to machine outputs, each rater conducted blind evaluations on a balanced mix of human and machine translations for sentences randomly selected from our original corpus. The evaluation yielded 1,972 ratings through a dual-rater process where each translation received independent ratings from two annotators. This included linguistic quality assessments of 900 translations across fluency, accuracy, and audience appropriateness, alongside offensiveness evaluations on 86 translations.

Although prior work has shown that MQM annotations typically achieve low inter-annotator agree-

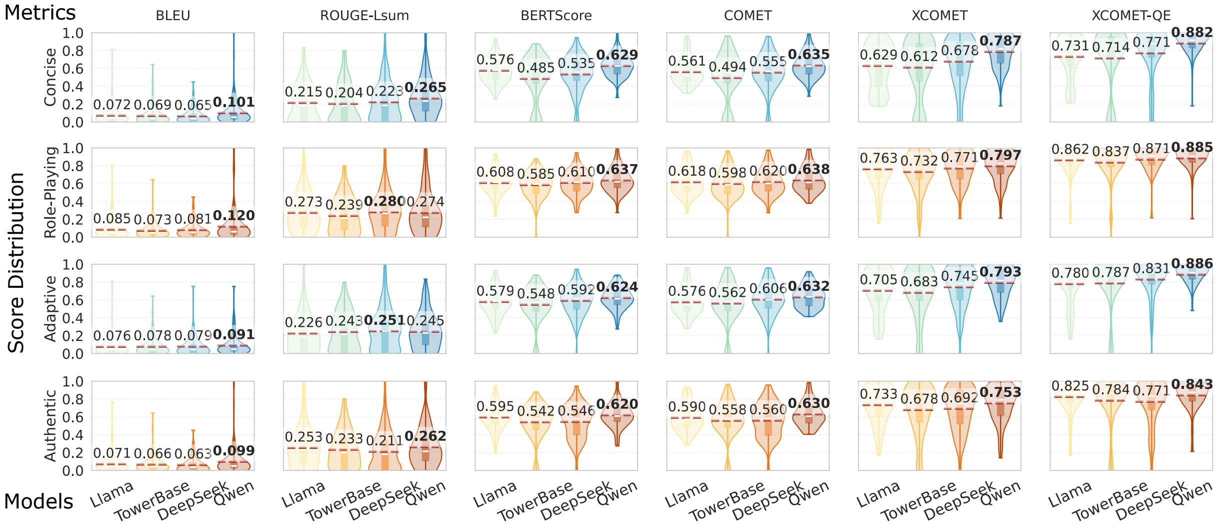


Figure 3: Model performance across different prompts at the sentence level. Each value represents the average score per LLM for the respective metric, with bold numbers highlighting the best performance within each prompt.

Task	Linguistics Quality		Sociocultural Adequacy	
Metric	*MQM Framework*			*Offensive Language*
Group	Accuracy	Fluency	Audience Appropriateness	Degree Rating
1	0.31	0.25	0.34	0.69
2	0.24	0.20	0.22	0.66
3	0.16	0.19	0.25	0.55
4	0.23	0.25	0.28	0.60
5	0.26	0.16	0.22	0.65

Table 1: Cohen's κ values for inter-rater agreement.

ment (Freitag et al., 2021), the framework remains valuable for identifying translation errors. As Table 1 shows, the Cohen's κ metric indicates a fair level of inter-annotator agreement overall for human evaluations in the present study. Notably, offensiveness ratings achieved higher consensus than translation quality assessment, as evaluating linguistic offensiveness is more straightforward than assessing holistic translation quality.

4 Results and Analysis

4.1 Automatic Evaluation

Results in Figure 3 show that LLMs achieve relatively low scores on n-gram overlap metrics, such as BLEU and ROUGE. In contrast, they achieve high scores on embedding-based metrics such as BERTScore and COMET-series, suggesting that LLMs make different lexical choices from reference translations, but still preserve source meaning.

In general, Qwen comes out on top among the four candidate models. When paired with role-playing prompts, it consistently delivers the best performance, only slightly trailing DeepSeek on the ROUGE metric. Therefore, we adopt Qwen with role-playing prompts as the exemplar LLM to generate the remaining translations for subsequent human evaluation.

4.2 Human Evaluation

This section presents human evaluation results from the MQM and offensive language ratings. Results indicate that while LLM translations achieve an adequate level of competence in linguistic quality comparable to human translations, they still fall short of sociocultural adequacy.

4.2.1 Linguistic Quality

Accuracy As shown in Table 2, LLM translations achieve scores comparable to human translations, with only marginal differences in accuracy. However, a closer analysis reveals that these similar scores hide different translation strategies. LLMs generally adopt a literal translation approach that prioritizes close alignment with the source text, yielding high accuracy scores when evaluations focus on textual correspondence. In contrast, human translations often incorporate creative adaptations to more effectively convey the underlying intent, which may result in deviations from the source text. For instance, while LLM translates "正好饿着 *zheng hao e zhe*" (happen to be hungry) as "I'm hungry", human translators render it as "Perfect timing", capturing the implied meaning that hunger coincides favorably with a meal opportunity. As a result, although human translations convey the communicative purposes, they may receive lower

Dimension	Group	Annotator	Human (M±SD)	LLM (M±SD)	t	p	Sig.
Accuracy	1	1	4.62±0.71	4.56±0.86	0.57	0.573	
		2	4.59±0.73	4.51±0.92	0.63	0.530	
	2	3	4.26±0.80	4.19±1.07	0.47	0.637	
		4	4.68±0.62	4.59±0.81	0.83	0.407	
	3	5	4.14±0.82	3.48±1.57	3.58	**< 0.001**	***
		6	4.32±0.78	4.62±0.88	-2.42	**0.016**	*
	4	7	4.73±0.65	4.70±0.76	0.32	0.752	
		8	4.11±1.08	4.66±0.81	-3.84	**< 0.001**	***
	5	9	4.16±1.03	4.01±1.30	0.83	0.410	
		10	3.82±1.41	3.49±1.67	1.45	0.150	

Table 2: Annotator's average accuracy ratings (M) with standard deviations (SD). Statistically significant differences (t, p, and Sig.) in scores between human and machine translations (Welch's t-test) are highlighted in bold.

scores under accuracy metrics emphasizing fidelity.

This observation suggests a potential limitation in applying MQM standards to creative texts. Unlike conventional translation tasks that prioritize accuracy in conveying propositional content, game localization often requires deliberate departures from the source text to achieve cultural adaptation, emotional resonance, and player engagement. Consequently, the emphasis on source-target fidelity may inadvertently penalize the creative translation that enhances quality in gaming contexts, indicating that specialized evaluation frameworks may be needed for assessing game localization.

Fluency Table 3 illustrates the model's capacity to generate fluent translations. Overall, the model performs strongly in terms of grammatical accuracy and punctuation. However, the lower fluency scores primarily stem from register inconsistencies and unnatural sentence flow, including awkward phrasing and redundant constructions that disrupt reading comprehension. These issues indicate persistent challenges in achieving stylistic precision for LLMs and directly affect player immersion in narrative-driven games. For instance, the LLM generates, "The evil monk *who incited* Jinchi Elder to *set fire to burn down* Tang Seng and his disciples *many years ago*". This translation suffers from redundant phrasing (*set fire to burn down*), and a passive, less engaging sentence structure caused by the use of a relative clause (*who*). In contrast, the human translation reads, "The evil monks *abetted* Elder Jinchi to *burn* the Great Sage and Tang Monk *alive*". This version is more concise and provides more vivid details (*burn ... alive*).

The difficulty of achieving natural fluency can be attributed to LLM's tendency toward literal translation, which results in rigid and lengthy sentences that disrupt the natural flow of dialogue. In contrast, human translators would restructure sentences to enhance readability. This observation aligns with previous findings that LLM outputs are generally more unnatural-sounding (Yan et al., 2024; Li et al., 2025). The linguistic complexity of our dataset also complicates the task, as it encompasses modern, classical, and vernacular Chinese variants. The diverse language styles require the model to balance formal linguistic structures with colloquial expressions, which creates a tension between maintaining structural fidelity and ensuring contextual fluency.

4.2.2 Sociocultural Adequacy

Audience Appropriateness Table 4 reveals a significant performance gap between LLM and human translations in tackling culture-specific terms, with LLM outputs consistently receiving statistically lower ratings. These relatively low ratings indicate that LLMs struggle with interpreting cultural nuances and appropriately conveying culture-specific terms, primarily due to their tendency toward cultural generalization and simplification. Table 5 illustrates culture-specific terms that require socioculturally aware translations. The Chinese expression "能耐 *neng nai*" (ability) carries nuanced connotations from genuine capability to sarcastic mockery, which require translators to interpret contextual cues to determine appropriate rendering. While human translators adapt their word choice to these varying contexts, LLM flattens this term to the emotionally neutral "ability" across all contexts and compromises the pragmatic information.

Beyond the issue of cultural neutralization that diminishes semantic connotations, the complexity of culturally appropriate translation is further compounded by diverse human preferences for localization strategies. The ratings for human translations exhibit notable variance and reflect disagreements about optimal cultural adaptation approaches. This

Dimension	Group	Annotator	Human (M±SD)	LLM (M±SD)	t	p	Sig.
Fluency	1	1	4.67±0.62	4.53±0.64	1.42	0.157	
		2	4.58±0.70	4.48±0.67	0.97	0.331	
	2	3	4.41±0.72	4.52±0.84	-0.96	0.341	
		4	4.84±0.47	4.74±0.73	1.09	0.276	
	3	5	4.78±0.68	3.23±1.79	7.64	**< 0.001**	***
		6	4.81±0.45	4.54±0.74	2.94	**0.004**	**
	4	7	4.79±0.59	4.62±0.73	1.69	0.093	
		8	4.59±0.78	4.51±0.82	0.65	0.516	
	5	9	4.57±0.72	4.42±1.02	1.10	0.273	
		10	4.20±1.32	3.83±1.44	1.78	0.076	

Table 3: Annotator's average fluency ratings (M) with standard deviations (SD).

Dimension	Group	Annotator	Human (M±SD)	LLM (M±SD)	t	p	Sig.
Audience Appropriateness	1	1	4.78±0.54	4.39±0.98	3.30	**0.001**	**
		2	4.60±0.73	4.31±0.99	2.23	**0.027**	*
	2	3	4.41±0.73	4.36±1.01	0.42	0.673	
		4	4.82±0.41	4.48±0.85	2.46	**0.015**	*
	3	5	4.39±0.83	4.20±1.30	1.16	0.247	
		6	4.94±0.23	4.71±0.71	2.98	**0.004**	**
	4	7	4.61±0.86	4.08±1.25	3.34	**< 0.001**	***
		8	4.51±0.97	4.27±1.26	1.46	0.148	
	5	9	4.32±0.89	4.31±1.07	0.08	0.940	
		10	4.52±0.97	3.68±1.34	4.84	**< 0.001**	***

Table 4: Annotator's average audience appropriateness ratings (M) with standard deviations (SD).

Source	这般能耐……正好正好
Human	A **strong foe**... Just what I need.
LLM	This kind of **ability**... just right, just right.
Source	有能耐，就在此间报仇罢！
Human	Then **if you can**, avenge him here and now!
LLM	If you have the **ability**, come and take revenge here!

Table 5: Contextual variations in translating the culture-specific term "能耐".

complexity is exemplified in the translation of "妖怪 *yaoguai*" (monster), where annotators disagreed on the optimal strategy. While some favor direct transliteration as "yaoguai" to preserve cultural authenticity and introduce players to Chinese mythological concepts, others advocate for the culturally adapted equivalent "monster" to enhance immediate comprehension and gameplay accessibility.

The divergent perspectives highlight a fundamental tension in game localization: the competing demands of cultural preservation versus target audience accessibility in creating engaging player experiences. As human translators show divided preferences for translation strategies, LLMs face an ambiguous learning environment that lacks clear optimization targets. Consequently, game localization may necessitate continued human oversight for cultural appropriateness. A potential solution is to adopt a perspectivist approach to dataset creation, which involves capturing multiple preferences from different annotators instead of enforcing a single ground truth (Cabitza et al., 2023). This would enable LLMs to generate a spectrum of choices to better serve the nuanced demands of localization.

Offensiveness Handling Figure 4 illustrates the shifts of offensive intensity in human and LLM translations compared to the source texts. Human translations exhibit a broader range of variation, with more instances of increased and decreased offensiveness, which reflects strategic adjustments on a case-by-case basis. For mitigation, translators often substitute or omit highly offensive language, particularly expressions targeting religion or gender, to avoid offending and alienating the target audience. Occasionally, they amplify provocative elements for better characterization, plot development, and emotional resonance to strengthen narrative engagement and player immersion.

In contrast, LLM translations consistently maintain a neutral stance due to their reliance on literal translation that fails to account for cultural nuances. Since offensive language is deeply embedded in cultural context and often carries implicit meaning, LLMs struggle to capture the subtleties required for effective cross-cultural adaptation. This limitation is more pronounced with culture-specific

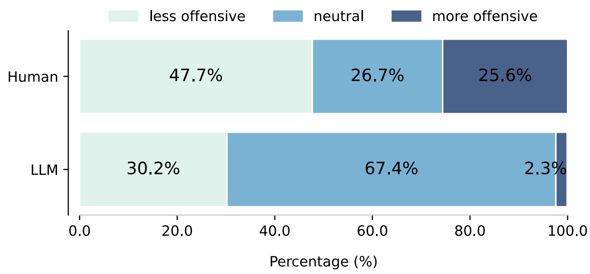


Figure 4: Distribution of offensive language intensity in human and LLM translations relative to source text.

expressions, such as idioms that demand cultural interpretation over direct linguistic conversion.

As exemplified in Figure 5, LLMs tend to preserve the original tone through literal translation, while human translators strategically adjust offensive language intensity to enhance immersive gameplay or align with target cultural norms. Interestingly, LLM's literal preservation can occasionally prove advantageous in gaming contexts where maintaining authentic character voices is crucial for narrative integrity, as such direct translation effectively conveys character traits like arrogance and rudeness. However, given localization requirements, human post-editing is needed to ensure cultural sensitivity in a global context.

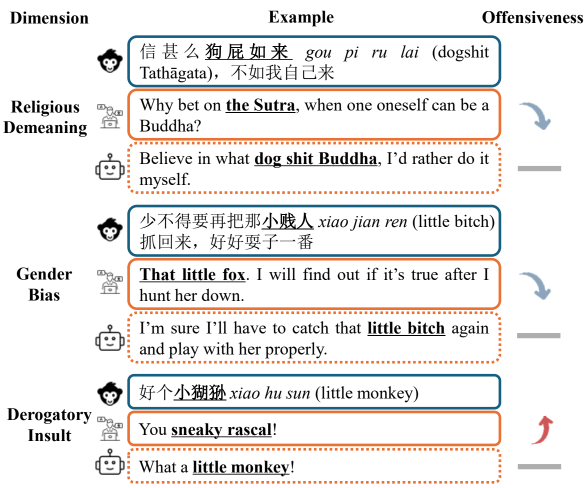


Figure 5: Examples of offensive language handling.

5 Conclusion

In this paper, we evaluate the performance of LLMs in game localization by examining linguistic quality and sociocultural adequacy. Our findings indicate that while LLMs demonstrate sufficient competence in general accuracy and fluency, they encounter challenges in contextual adaptation of culture-specific terms and offensive language. Evidence from human evaluations suggests that most automatic metrics and MQM standards may not be appropriate for evaluating creative translation. In addition, diverse human preferences in localization create learning ambiguity for LLMs to identify optimal translation strategies.

We consider our study as the first step to deepen our understanding of the strengths and limitations of LLMs in the translation of creative textual domains. As LLMs continue to evolve, our results highlight their promising potential as collaborative tools in the professional game localization workflow. Particularly in the context of the translation from Chinese to Western languages, where the biggest challenge is conveying meanings from an ancient culture with which those audiences are not familiar, we sincerely hope that LLMs can help Sun Wukong and the other heroes from future Chinese releases in their own Journey to the West.

Limitations

Our experiment exclusively employs open-source LLMs for reproducibility, thereby excluding proprietary systems such as the GPT series. Future studies could incorporate these models with larger sizes to enable comprehensive benchmarking, and clarify whether the limitations described above can be overcome through the use of larger-scale, more powerful architecture or they are inherent of the LLM paradigm. While our corpus provides creative contexts, its single-game focus limits the exposure to a wider range of linguistic patterns, potentially constraining the generalizability of our conclusions. This constraint highlights the pioneering and scarce availability of such culturally dense corpora. Additionally, although our Chinese-to-English focus aligns with the commercial demand for localizing Chinese games into global markets, game localization pipelines should support a larger number of languages and ensure fair treatment of all the cultures. Finally, due to the high cost of human annotation, we could evaluate only a randomly sampled subset of data. Our future work will address this limitation through more efficient annotation.

Acknowledgments

We express our sincere gratitude to the anonymous reviewers for their invaluable time and constructive suggestions. We also thank the annotators for their efforts and thorough work.

References

Mohammed Al-Batineh. 2021. Issues in arabic video game localization: A descriptive study. *Translation & Interpreting: The International Journal of Translation and Interpreting Research*, 13(2):45–64.

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Vicent Briva-Iglesias, Joao Lucas Cavalheiro Camargo, and Gokhan Dogru. 2024. Large language models" ad referendum": How good are they at machine translation in the legal domain? *arXiv preprint arXiv:2402.07681*.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6860–6868.

Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. 2024. Benchmarking llms for translating classical chinese poetry: Evaluating adequacy, fluency, and elegance. *arXiv preprint arXiv:2408.09945*.

Alberto Fernández Costales. 2016. Analyzing players' perceptions on the translation of video games: Assessing the tension between the local and the global concerning language use. In *Media Across Borders*, pages 183–201. Routledge.

Ugo Ellefsen and Miguel Á Bernal-Merino. 2018. Harnessing the roar of the crowd: A quantitative study of language preferences in video games of french players of the northern hemisphere. *The Journal of Internationalization and Localization*, 5(1):21–48.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Masood Khoshsaligheh, Saeed Ameri, Farzaneh Shokoohmand, and Mehdi Mehdizadkhani. 2020. Subtitling in the iranian mediascape: Towards a culture-specific typology. *International Journal of Society, Culture & Language*, 8(2):55–74.

Yafu Li, Ronghao Zhang, Zhilin Wang, Huajian Zhang, Leyang Cui, Yongjing Yin, Tong Xiao, and Yue Zhang. 2025. Lost in literalism: How supervised training shapes translationese in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12875–12894, Vienna, Austria. 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.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12):0455–463.

Anjad A Mahasneh and Maysa'Taher Abu Kishek. 2018. Arabic localization of video games "tomb raider™(2013)": A start or a failure. *Lebende Sprachen*, 63(1):47–62.

Luis Damián Moreno García and Carme Mangiron. 2024. Exploring the potential of gpt-4 as an interactive transcreation assistant in game localisation: A case study on the translation of pokémon names. *Perspectives*, pages 1–18.

Minako O'Hagan and Carmen Mangiron. 2006. Game localisation: Unleashing imagination with" restricted" translation. *The Journal of Specialised Translation*, (6):10–21.

Minako O'Hagan and Heather Chandler. 2016. Game localization research and translation studies: Loss and gain under an interdisciplinary lens. In *Border Crossings*, pages 309–330. John Benjamins Publishing Company.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Kevin Stowe, Prasetya Utama, and Iryna Gurevych. 2022. IMPLI: Investigating NLI models' performance on figurative language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5375–5388, Dublin, Ireland. Association for Computational Linguistics.

Kenan Tang, Peiyang Song, Yao Qin, and Xifeng Yan. 2024. Creative and context-aware translation of East Asian idioms with GPT-4. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9285–9305, Miami, Florida, USA. Association for Computational Linguistics.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.

Zhiwei Wu and Zhuojia Chen. 2020. Localizing chinese games for southeast asian markets: A multidimensional perspective. *The Journal of Internationalization and Localization*, 7(1-2):49–68.

Jianhao Yan, Pingchuan Yan, Yulong Chen, Judy Li, Xianchao Zhu, and Yue Zhang. 2024. Gpt-4 vs. human translators: A comprehensive evaluation of translation quality across languages, domains, and expertise levels. *arXiv preprint arXiv:2407.03658*.

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, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report.

Binwei Yao, Ming Jiang, Tara Bobinac, Diyi Yang, and Junjie Hu. 2024. Benchmarking machine translation with cultural awareness. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13078–13096, Miami, Florida, USA. Association for Computational Linguistics.

Shisen Yue, Siyuan Song, Xinyuan Cheng, and Hai Hu. 2024. Do large language models understand conversational implicature- a case study with a Chinese sitcom. In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 1: Main Conference)*, pages 1270–1285, Taiyuan, China. Chinese Information Processing Society of China.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

Author Index