Under-resourced studies of under-resourced languages: lemmatization and POS-tagging with LLM annotators for historical Armenian, Georgian, Greek and Syriac

Chahan Vidal-Gorène[1,2], Bastien Kindt[3], Florian Cafiero[2]

[1]LIPN, CNRS UMR 7030, France
[2]École nationale des chartes, Université Paris-Sciences-et-Lettres (PSL), France
[3]Université catholique de Louvain, Belgium
Correspondence: chahan.vidal-gorene@chartes.psl.eu

Abstract

Low-resource languages pose persistent challenges for Natural Language Processing tasks such as lemmatization and part-of-speech (POS) tagging. This paper investigates the capacity of recent large language models (LLMs), including GPT-4 variants and open-weight Mistral models, to address these tasks in few-shot and zero-shot settings for four historically and linguistically diverse under-resourced languages: Ancient Greek, Classical Armenian, Old Georgian, and Syriac. Using a novel benchmark comprising aligned training and out-of-domain test corpora, we evaluate the performance of foundation models across lemmatization and POS-tagging, and compare them with PIE, a task-specific RNN baseline. Our results demonstrate that LLMs, even without fine-tuning, achieve competitive or superior performance in POS-tagging and lemmatization across most languages in few-shot settings. Significant challenges persist for languages characterized by complex morphology and non-Latin scripts, but we demonstrate that LLMs are a credible and relevant option for initiating linguistic annotation tasks in the absence of data, serving as an effective aid for annotation.

1 Introduction

While not always surpassing state-of-the-art specialized methods, Large Language Models demonstrated remarkable zero-shot performance in tasks such as POS-tagging for languages such as Arabic (Alyafeai et al., 2023) and Armenian (Vidal-Gorène et al., 2024). In related tasks such as NER, recent advancements for low-resource languages have demonstrated promising results through knowledge transfer from large pre-trained language models (PLMs). In particular, a recent study (Tolegen et al., 2024) showcased significant improvements in Kazakh NER by leveraging multilingual models such as XLM-RoBERTa, achieving a 7% increase in F1-score compared to previous approaches. However, their findings also revealed that few-shot learning scenarios remain a challenge, with multilingual PLMs struggling to achieve high performance when trained on limited data. Interestingly, their comparison with GPT-4 indicated that such large-scale generative models exhibit superior adaptability, even in low-data scenarios, suggesting the potential of these models to generalize across diverse linguistic settings.

This paper aims to build upon these insights by exploring how GPT-4 and similar foundation models can offer a more agile and effective approach to low-resource NLP, demonstrating their ability to bridge the gap between high- and low-resource languages with minimal fine-tuning effort.

1.1 State of the Art

Morphological annotation tasks such as lemmatization and POS tagging in under-resourced and historically diverse languages have been approached using rule-based systems, sequence-based neural architectures, and more recently, large language models.

For Armenian, Vidal-Gorène et al. (2020) demonstrate that RNN-based models trained on Eastern Armenian can be reused to annotate dialectal and Western varieties, outperforming rule-based baselines and showing high transferability. Similarly, in the context of Ancient Greek, Kindt et al. (2022) evaluate an RNN trained on patristic Greek to annotate historiographical prose, achieving over 97% accuracy in both lemmatization and POS-tagging, with strong performance on ambiguous and unseen forms.

Manjavacas et al. (2019) introduce a joint-learning architecture for lemmatization that combines character-level transduction with a sentence-level language modeling objective. Applied to non-standard historical languages such as Latin, Old French, the model surpasses baselines, especially on ambiguous forms, without relying on gold mor-

Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026), pages 324–334

phological annotations. It has then been applied to other historical languages like Old French and Classical French (Camps et al., 2021; Cafiero and Camps, 2019; Clérice and Pinche, 2025) or 14th century Dutch (Creten et al., 2020), and integrated into specialized annotation infrastructures (Clérice et al., 2022).

The feasibility of extending such models to typologically diverse languages is addressed by Vidal-Gorène and Kindt (2020), who apply a character-level RNN (PIE) to Classical Armenian, Old Georgian, and Syriac. Their models reach over 91% accuracy in lemmatization and 92% in POS tagging, using curated corpora derived from the GRE*g*ORI project. Results confirm the adaptability of RNNs to languages with complex polylexical structures, although accuracy on ambiguous and unknown forms remains lower (e.g., 71.9% F1 on unknown tokens in Classical Armenian). The authors emphasize the limitations of rule-based pipelines and the potential of neural models to handle diachronic and morphologically rich corpora written in the different languages of the Christian East.

More recently, Vidal-Gorène et al. (2024) benchmark RNNs, mDeBERTa, and GPT-4 across four Armenian varieties, including Classical and a rare spoken dialect. RNNs remain competitive for POS-tagging (F1 $>$ 0.98), while GPT-4 demonstrates strong generalization in zero- and few-shot setups, achieving F1 = 0.83 in lemmatization on unseen dialect data. Transformer models underperform on morphologically rich and low-resource dialects, particularly in low-data conditions.

Taken together, these studies highlight the importance of context-aware neural architectures for low-resource NLP. They also point to persistent challenges: heterogeneous annotation schemes, limited corpus coverage, and the need for robust cross-dialectal generalization.

2 Datasets

Our data sets comprise texts in Greek, Armenian, Georgian, and Syriac, representing the linguistic and cultural diversity of the Christian East. Each language represents a sub-dataset, divided into two parts: a 'Training corpus' (5,000 words) for model training, and an out-of-domain 'Test corpus' (300 words) for evaluation purposes (see Table 1).

Each text underwent preliminary linguistic analysis, including lemmatization and POS-tagging, tagged with a hybrid approach of rule-based and RNN models (Vidal-Gorène and Kindt, 2020; Kindt et al., 2022).

2.1 Language families

The languages in this dataset belong to different language families: Greek (Indo-European) with a rich inflectional system; Armenian (Indo-European) with a partially inflectional and agglutinative system; Georgian (Kartvelian) with a more distinctly agglutinative system; and Syriac (Semitic), which presents a non-concatenative, templatic morphology. The texts were produced between the 4th and 15th centuries AD and cover a wide range of topics, including asceticism, epistolography, exegesis, hagiography, historiography, homiletics, patristics, pseudepigrapha, and theology. They are either original texts or ancient translations, all previously published in critical editions. This sample, though only partially representative, is highly varied and diversified.

2.2 Annotation guidelines and tagsets

The lemmatization and POS tagging follow the GRE*g*ORI annotation guidelines UCLouvain – CIOL (Centre d'études orientales – Institut Orientaliste de Louvain) (2022), which use the @ symbol to discriminate the lexical elements constituting a polylexical form (crases of Greek; agglutinated forms of Georgian; prefixed and suffixed forms of Syriac, etc.). For instance, a word composed of multiple lemmas is annotated as lemma1@lemma2 with corresponding POS tags as pos1@pos2 (see Figure 1). Unlike standard annotation frameworks such as Universal Dependencies, this method provides a more granular representation by explicitly marking all constituent lemmas and their POS tags. However, it lacks direct equivalents in publicly available datasets and increases task complexity, making model training more challenging. Additionally, the GRE*g*ORI tagset diverges from conventional tagsets in its categorization and structure, further complicating annotation and processing (Coulie et al., 2013, 2022; Atas, 2022).

The full tagset is provided in Table 7 in the appendix. These POS tags introduce additional complexity compared to standard tagsets, as some would typically be treated as morphological markers rather than POS categories (e.g., `PRO+Per1d`, `PRO+Per3s`, `PRO+Ref1s`). We retain the original annotations without normalization[1].

[1]For full annotation scheme guidelines, see Armenian (Coulie et al., 2022), Georgian (Coulie et al., 2013), Greek

	Training Corpus	Test Corpus
Greek	John Anagnostes, *De Thessalonica Capta* (15th c. AD - historiography)	Gregory of Nazianzus, *Homily 1. In Sanctum Pascha* (4th c. AD - patristics, homiletic)
Armenian	Evagrius, *Letters* (13th c. AD - epistolography, ascetism)	Step'anos of Siwnik' (Dub.), *The Genesis Commentary* (8th-9th c. AD - exegesis, theology)
Georgian	Anonymus, *Conversion of the Kartli* (5th-9th c. AD - hagiography, historiography)	*History of the Rechabites* by Zosimus (8th-10th c. AD - hagiography, historiography)
Syriac	Jacob of Serugh, *Homilies* (5th-6th c. AD - patristics, homiletic)	*History of the Rechabites* by Zosimus (Long Version) (4th c. AD - hagiography, pseudepigraphs)

Table 1: Overview of training and test corpora: chronological and genre diversity.

Token	Lemma	POS
ἀναστάσεως *anastáseōs*	ἀνάστασις *anástasis*	N+Com
զաստուածսն *zastwacsn*	զ@աստուած@ն *z@astwacs@n*	I+Prep@N+Com@PRO+Dem
ჩემგან *chemgan*	მე (ჩემ)@გან *me (chem)@gan*	PRO+Per1s@I+Prep
ܒܦܘܪܩܢܗ	ܒ-@ܦܘܪܩܢܐ@ܗܘ	PART@NOUN@PRO_pers

Figure 1: Annotation guidelines and tagset for Greek, Armenian, Georgian and Syriac, using @ to split agglutinated and polylexical forms

	Train		Test (In)		Test (Out)	
Lang	**Poly.**	**Simple**	**Poly.**	**Simple**	**Poly.**	**Simple**
GRC	13	4705	2	298	1	301
HYE	669	4043	56	244	55	259
KAT	156	4551	5	295	5	300
SYC	1696	2296	144	156	140	161

Table 2: Distribution of polylexical vs. simple forms across datasets.

2.3 Polylexicality and Segmentation

A critical challenge in our dataset, and particularly for Syriac, is the prevalence of polylexical forms, where a single graphical token corresponds to multiple syntactic units (e.g., a preposition attached to a noun). In the GRE*g*ORI schema, these are annotated by splitting the form with an `@` symbol (e.g., *lemma1@lemma2*).

As shown in Table 2, the density of these forms varies drastically across languages. Syriac exhibits an exceptionally high density, with nearly 42% of tokens in the training set being polylexical, reflecting its Semitic morphology where prepositions and conjunctions are prefixed. In contrast, Greek and Georgian (in this specific annotation schema) show significantly lower rates of explicit segmentation.

3 Tasks

3.1 Task definition: lemmatization and POS tagging

We consider two classic sequence-labeling tasks: lemmatization and part-of-speech (POS) tagging.

(Kindt, 2004), and Syriac (Kindt et al., 2018).

In each case, the problem can be formulated as follows. Given an input sequence of n tokens $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$, we wish to predict a sequence of labels $\hat{\mathbf{y}} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n\}$ from an appropriate label set. For *lemmatization*, each label y_i is the canonical lemma form of token x_i. In *POS tagging*, y_i is assigned one of a finite set of syntactic categories (e.g., noun, verb). A collection of input-label pairs forms our data set $\mathcal{D}$. We train a model with parameters $\boldsymbol{\theta}$ to maximize the conditional log-likelihood of the correct label sequences:

$$\max_{\boldsymbol{\theta}} \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}} \log P(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\theta}).$$

The *POS-tagging* label set includes a small number of NER-related categories (e.g., `N+Ant`, `N+Epi`, `N+Pat`, `N+Prop`, `N+Top`, `NAME`, `NAME_ant`, `NAME_top`, see Table 7 in the appendix), but these are too infrequent in the data to support separate evaluation.

3.2 Prompt Engineering

To adapt general-purpose LLMs to the specificities of the GRE*g*ORI annotation schema, we employed a structured prompting strategy based on the COSTAR framework (Teo, 2023). This approach

Context: You are an expert in computational linguistics with a specialization in [LANGUAGE]...
Objective: Examine the provided content and assign lemma and POS tags...
Format: A .tsv table with three columns...
Constraints:
1. All words must be annotated.
2. Only use tags from the provided list:
`[INSERT FULL TAGSET LIST HERE]`
Segmentation Rule:
A form can be a combination of multiple tokens. Identify combinations with '@'.
Example: [Language Specific Example of Split]

Few-Shot Examples:

form	lemma	pos
w_1	l_1	p_1
...		
w_k	l_k	p_k

Input text to process:
`[TEST INPUT]`

Figure 2: Schematic representation of the prompt structure used for all languages.

decomposes the prompt into six components: Context, Objective, Style, Tone, Audience, and Response.

Given the non-standard nature of our target tagset (which includes morphological markers fused with POS tags), standard instruction tuning was insufficient. We therefore implemented two critical constraints within the prompt design (see Figure 2):

1. **Tagset Injection:** We explicitly injected the full list of valid POS tags ($|\mathcal{T}| \approx 60$) into the prompt context. This acts as a strict in-context constraint, drastically reducing the likelihood of the model hallucinating invalid tags or reverting to Universal Dependencies tags.

2. **Segmentation Guidance:** To address the challenge of polylexicality (e.g., Syriac prefixed prepositions), the prompt includes a specific instruction and a concrete example demonstrating the use of the `@` delimiter to split lemmata and tags (e.g., mapping the Greek *tauton* to *ho@autos*).

For the **Few-Shot** settings ($k = 5, 50, 500$), we appended k aligned examples (token, lemma, POS) from the training set. These examples were selected sequentially from the start of the training corpus to maintain narrative coherence. Unlike random sampling, this sequential presentation allows the model to leverage sentence-level context for morphosyntactic disambiguation.

3.3 Decoding Strategy

We utilized model-specific decoding parameters to optimize for stability and adherence to the tagging schema.

- **Greedy Decoding ($\tau = 0.0$):** For *GPT-4o*, we employed strictly deterministic decoding. These models demonstrated high adherence to constraints without requiring stochastic noise to avoid degeneration.

- **Low-Temperature Sampling ($\tau = 0.2$):** For the *GPT-4o-mini*, *Mistral Nemo*, and *Mistral Large* experiments, we applied a slight temperature increase. We standardized the batch processing for these groups with $\tau = 0.2$. This setting served as a preventative measure against repetitive loops observed in preliminary experiments, ensuring robust generation across extensive test sets.

- **Reasoning Models:** For *o1-mini*, we utilized the model's default decoding parameters to preserve the integrity of the internal chain-of-thought process.

3.4 Baseline definition

While Transformer-based encoders (e.g., mDeBERTa) generally define the state-of-the-art for high-resource languages, recent work on historical and dialectal varieties suggests they are not always the optimal baseline. Vidal-Gorène et al. (2024) demonstrated that for character-level tasks in Classical and Modern Armenian, the RNN-based PIE architecture consistently outperformed mDeBERTa, particularly in low-data regimes where Transformers struggle to generalize without extensive fine-tuning. Furthermore, from a pragmatic perspective, lightweight RNN architectures remain significantly more accessible to the digital humanities community, especially in the context of low-resource studies. RNN offer lower computational overhead and greater ease of deployment for non-technical experts compared to large Transformer pipelines.

Consequently, we retain PIE as our primary supervised baseline. PIE is a multi-task character-level RNN architecture specifically designed for the joint annotation of morphologically rich languages. For all languages in this study, we utilize a standardized configuration, summarized in Table 3, to ensure comparability and force the model to

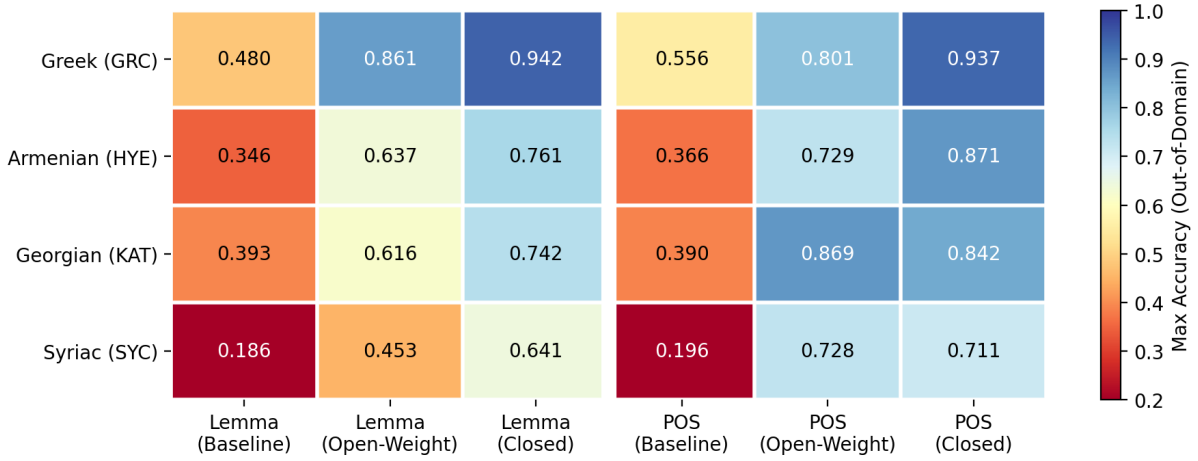

Figure 3: Out-of-domain accuracy for lemmatization and POS tagging across our four historical languages: comparing the supervised PIE baseline with best open-weight and closed LLM annotators; values report maximum accuracy by task and language.

Component / Parameter	Configuration
Encoder Architecture	2-layer Bi-directional GRU
Hidden Size	300
Char Embeddings	300 (2-layer RNN)
Word Embeddings	0 (Disabled)
Lemma Decoder	Attentional (Character-level)
POS Decoder	CRF (Conditional Random Field)
Dropout / Batch Size	0.25 / 50
Optimization / LR	Adam (0.001)
LR Scheduler	Factor: 0.75 (Patience: 2)
Early Stopping	Patience: 3 epochs (Max 100)

Table 3: Hyperparameter configuration for the PIE supervised baseline.

rely exclusively on character-level morphological patterns.

4 Results

Tables 5 and 6 summarize our complete results, evaluated in both in-domain and out-of-domain scenarios. A simplified comparison is provided in figure 3.

4.1 Lemmatization

Overall, large language models, particularly GPT-4o, but even open-weight models like mistral-large, significantly outperform the traditional RNN-based model PIE across most scenarios, especially in low-shot conditions. For Greek, GPT-4o attains the highest accuracy with **0.933** (in-domain) and **0.942** (out-of-domain) at 500 shots, though mistral-large also demonstrates robust zero-shot performance (0.863 in-domain, 0.861 out-of-domain). For Armenian, GPT-4o achieves notable accuracy (**0.796** in-domain, **0.761** out-of-domain at 500 shots), while mistral-large remains competitive in medium-shot scenarios. Georgian results indicate consistent difficulties across models, although GPT-4o still maintains leading performance at **0.760** (in-domain, 500 shots). Syriac poses the greatest challenge overall; GPT-4o reaches a peak accuracy of **0.656** (in-domain), and open-weight models generally demonstrate moderate performances.

The PIE baseline performs substantially lower, illustrating limitations of traditional sequence labeling architectures when data availability is limited.

4.2 POS Tagging

In POS-tagging, GPT-4o and mistral-large both show strong results, particularly in Greek, where GPT-4o reaches **0.963** (in-domain) and **0.937** (out-of-domain) at 500 shots. Armenian shows consistent robustness for GPT-4o (**0.863** in-domain, **0.871** out-of-domain) while mistral-large achieves competitive accuracies (0.740 in-domain, 0.729 out-of-domain). Georgian demonstrates variability; mistral-large achieves the highest out-of-domain accuracy (**0.869** at 500 shots), highlighting the capability of open-weight models. Syriac remains challenging; GPT-4o reaches **0.863** (in-domain), whereas mistral-large attains the best out-of-domain accuracy (**0.728**).

In general, large language models demonstrate clear advantages over the PIE baseline, especially in low-resource scenarios. Open-weight models sometimes closely match or even surpass GPT models in certain languages or data conditions.

Lang	Lemma Overlap (%)		Top 3 POS (Test Set)	Polylexical (@) %	
	Type	Token		Train	Test
GRC	47.6	62.9	V (17.9%), N+Com (17.5%), I+Part (16.2%)	0.3	0.3
HYE	52.7	67.2	V (18.8%), I+Conj (17.8%), N+Com (15.6%)	14.2	17.5
SYC	43.4	42.2	NOUN (15.6%), PART@NOUN (14.6%), ADJ (11.3%)	42.5	46.5
KAT	62.9	76.4	N+Com (22.3%), V+Mas (20.3%), I+Conj (17.4%)	3.3	1.6

Table 4: Lexical overlap, dominant POS distributions, and density of polylexical forms across languages.

5 Discussion

Our findings highlight both the capabilities and limitations of large language models in morphosyntactic annotation tasks for under-resourced and typologically diverse languages. While traditional methods such as the RNN-based PIE baseline require substantial annotated data to achieve competitive results, LLMs—both proprietary models like GPT-4o and open-weight alternatives like mistral-large—show significant promise even in few-shot and zero-shot contexts. This ability to perform reliably with minimal supervision makes them particularly valuable for languages lacking extensive annotated resources.

5.1 Lexical Overlap and the Generalization Gap

To rigorously assess whether these models rely on memorization or genuine linguistic reasoning, we analyzed the lexical overlap between training and test sets (see Table 4). A key finding is that *Lemma Type Overlap* remains consistently low across all languages ($< 63\%$), meaning the models must process a high volume of previously unseen vocabulary.

We observe a significant gap between *Type Overlap* and *Token Overlap*. In Greek, Armenian, and Georgian, token overlap is 13 to 24 percentage points higher than type overlap. This indicates that while models encounter many novel lemmas, they are primarily familiar with the most frequent lexical items. This familiarity with the "syntactic backbone" (Verbs, Nouns, and Particles) provides sufficient context to maintain robust POS-tagging accuracy (> 0.85 F1) even on Out-Of-Vocabulary (OOV) items.

However, the Syriac (SYC) case confirms that LLMs do more than simple pattern matching. Syriac exhibits the most challenging distribution, with both type and token overlap dropping to $\approx 42\%$. In this scenario, the majority of the text consists of novel lemmata. The fact that models maintain (relatively) competitive performance despite this low "seen-ness" is a measurable indicator of their capacity for morphological generalization.

5.2 The Impact of Polylexicality (@)

The density of polylexical forms (marked by `@`) is a major predictor of performance degradation. Syriac presents the highest complexity: 46.5% of its test tokens require predicting an internal segmentation boundary (e.g., *PART@NOUN*). For LLMs using sub-word tokenization, generating these non-standard delimiters within non-Latin scripts is highly error-prone, directly accounting for the performance gap compared to Greek or Armenian.

Interestingly, Georgian (KAT) shows that overlap is not the only factor. Despite having the highest lexical overlap and few polylexical forms ($< 2\%$), it remains more challenging than Greek. This suggests that the complexity of its agglutinative morphology—often not explicitly segmented in this schema—requires deeper linguistic pre-training that foundation models may still lack for kartvelian languages.

5.3 Conclusion on Model Robustness

In general, while acknowledging the limitations inherent to out-of-domain benchmarks, our results confirm that LLMs provide an efficient and cost-effective path to annotating under-resourced languages. They serve as robust "cold-start" annotators, capable of facilitating the incremental creation of gold-standard datasets for historically and linguistically diverse corpora.

Our analysis of lexical seen-ness reveals an important distinction: while high token overlap provides a performance "floor" for frequent items, the models' ability to maintain accuracy in low-overlap environments—most notably in Syriac, where the majority of the vocabulary is novel—demonstrates a capacity for genuine morphological reasoning over simple lexical memorization. Future work should focus on moving beyond random sampling towards strategic input selection (Bansal and

Sharma, 2023). Such strategies could further narrow the performance gap between seen and unseen tokens, significantly enhancing model generalization in the high-sparsity and domain-shift scenarios typical of under-resourced languages and non-Latin languages.

5.4 Error Typology: Formatting vs. Linguistic Competence

A manual review, but very limited at this stage, of the outputs reveals a fundamental distinction between structural failures and genuine morphosyntactic competence. In Syriac, the high density of polylexical forms frequently causes structural desynchronization in TSV outputs. Mismanagement of the @ delimiter often results in lower raw accuracy scores that may not fully reflect the model's underlying linguistic competence. Similarly, in Classical Armenian, models exhibit a bias toward modern training data by hallucinating reformed orthography instead of the required classical standard.

These observations suggest that the models' true competence exceeds raw metrics, leading us to recommend the use of more robust formats, such as JSON, for future workflows to decouple data structure from philological content.

Limitations

Our evaluation is constrained by the small size and genre specificity of the benchmark dataset. Although the corpora were selected to represent typological and historical diversity, each training and test set obviously remains limited in size.

Our annotation scheme, based on GRE*g*ORI guidelines, introduces structural complexity through polylexical and agglutinative segmentation marked by the @ delimiter. The tagset is also highly specific to GRE*g*ORI, with morphological tags within the POS list. While linguistically motivated, this representation lacks standardization, limiting cross-corpus comparability and hindering broader model development.

Model performance varies significantly across tasks and languages. LLMs showed strong results in POS tagging and lemmatization for some languages but consistently underperformed for some languages. These discrepancies suggest that few-shot generalization remains highly sensitive to script, morphology, and training distribution, with models struggling to identify entities in highly inflected, under-represented languages.

Although our experiments include a range of foundation models (GPT-4 variants and Mistral models), we do not systematically investigate the impact of prompt engineering, decoding strategies, or intermediate fine-tuning. Future work should explore how task- and language-specific prompts might further boost performance in few-shot settings.

Finally, our approach presumes the availability of token-level supervision for evaluation, which may not be feasible for other truly underdocumented languages. The reliance on pre-existing annotated corpora, even if limited, underscores a key challenge in extending this approach to languages without any available training data.

Data availability

Code, results and datasets are available at: https://github.com/CVidalG/EACL2026-historical-languages.

Acknowledgements

This research was supported by the French National Research Agency (ANR), grant ANR-21-CE38-0006 (DALiH project) and conducted as part of the PSL Research University's Major Research Program CultureLab, implemented by the ANR (reference ANR-10-IDEX-0001). We also thank the GRE*g*ORI lab for providing access to the data. The authors used generative AI to improve the linguistic clarity and assist with LaTeX formatting of the final document.

Ethics Statement

This research adheres to the ACL Ethics Policy. Our experiments use openly accessible linguistic datasets and publicly available LLMs, ensuring transparency and reproducibility. We acknowledge that working with historical and culturally significant texts requires sensitivity and care.

Although our datasets consist exclusively of texts published and available in critical scholarly editions, researchers employing our methods in other contexts must ensure respect for cultural and historical heritage, particularly when annotating or disseminating results related to under-resourced or minority languages. Furthermore, we encourage future work to consider data sovereignty and community involvement when extending these techniques to languages without established digital resources.

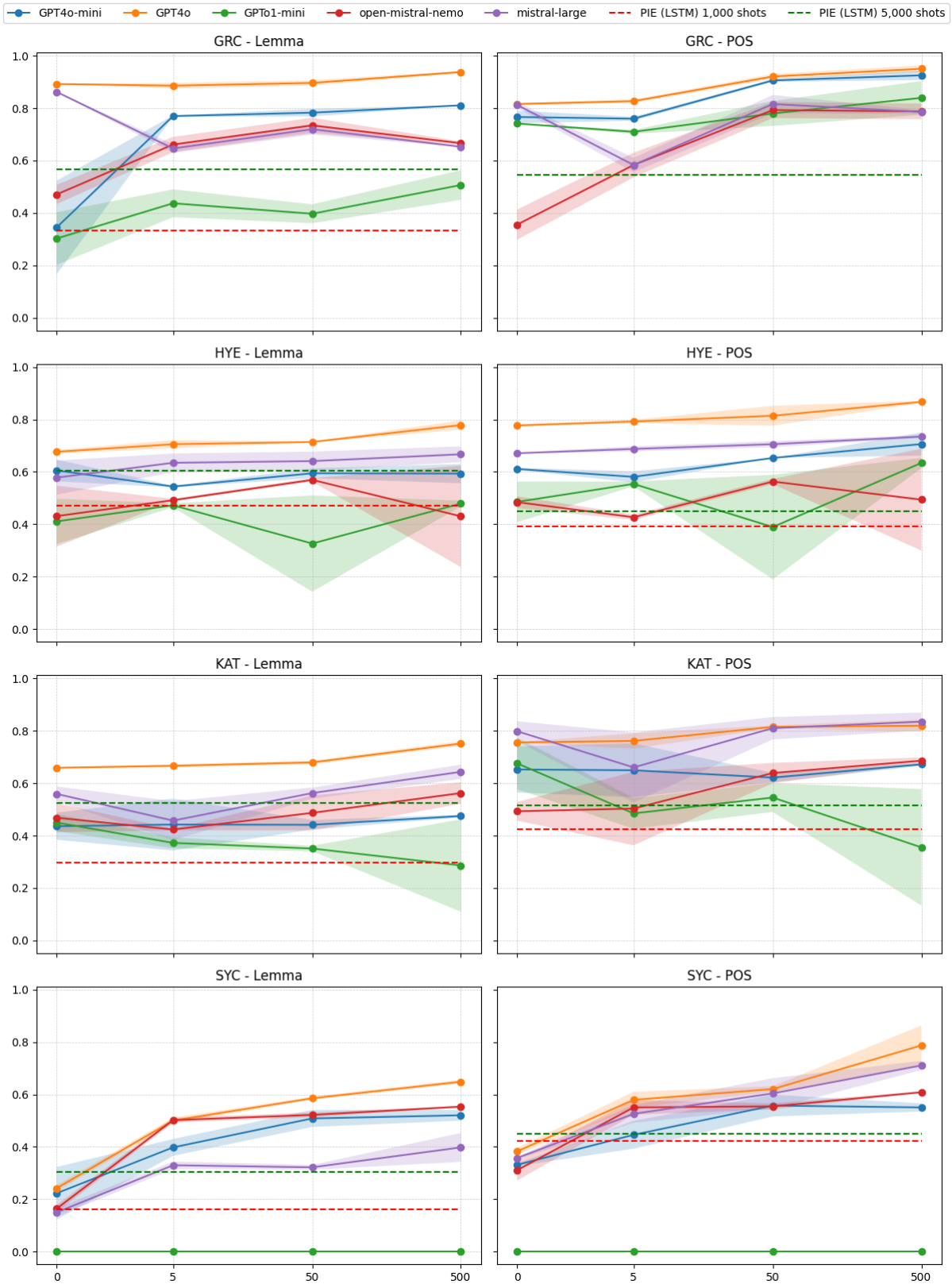

Figure 4: Figure 4. Accuracy as a function of the number of in-context examples (0, 5, 50, 500 shots) for lemmatization and POS tagging, reported separately for each language (GRC, HYE, KAT, SYC). Solid lines show the model accuracy at each shot count; shaded bands span the two evaluation conditions and thus indicate the range between in-domain and out-of-domain accuracies at the same shot count. Dashed horizontal lines mark the supervised PIE reference accuracies obtained with 1,000 and 5,000 training examples.

References

Zaid Alyafeai, Maged S. Alshaibani, Badr AlKhamissi, Hamzah Luqman, Ebrahim Alareqi, and Ali Fadel. 2023. Taqyim: Evaluating Arabic NLP tasks using ChatGPT models. *Computing Research Repository*, abs/2306.16322.

Nicolas Atas. 2022. Principles of Syriac lemmatisation: Summary version. UCLouvain – Institut orientaliste – GREgORI Project. Accessed: 2026-01-06.

Parikshit Bansal and Amit Sharma. 2023. Large language models as annotators: Enhancing generalization of NLP models at minimal cost. *Computing Research Repository*, abs/2306.15766.

Florian Cafiero and Jean-Baptiste Camps. 2019. Why molière most likely did write his plays. *Science advances*, 5(11):eaax5489.

Jean-Baptiste Camps, Simon Gabay, Paul Fièvre, Thibault Clérice, and Florian Cafiero. 2021. Corpus and models for lemmatisation and POS-tagging of classical French theatre. *Journal of Data Mining & Digital Humanities*, 2021.

Thibault Clérice, Vincent Jolivet, and Julien Pilla. 2022. Building infrastructure for annotating medieval, classical and pre-orthographic languages: the Pyrrha ecosystem. In *Digital Humanities 2022 (DH2022)*, Tokyo, Japan.

Thibault Clérice and Ariane Pinche. 2025. Wauchier, is that you? a multi-manuscript authorship analysis of saint lambert's life. *Anthology of Computers and the Humanities*, 3:149–165.

Bernard Coulie, Bastien Kindt, Gabriel Kepeklian, and Emmanuel Van Elverdinghe. 2022. Étiquettes morphosyntaxiques et flexionnelles pour le traitement automatique de l'arménien ancien. *Le Muséon*, 135(1-2):209–241.

Bernard Coulie, Bastien Kindt, and Tamara Pataridze. 2013. Lemmatisation automatique des sources en géorgien ancien. *Le Muséon*, 126(1-2):161–201.

Silke Creten, Peter Dekker, and Vincent Vandeghinste. 2020. Linguistic enrichment of historical Dutch using deep learning. *Computational Linguistics in the Netherlands Journal*, 10:57–72. Submitted 10/2020; Published 12/2020.

Bastien Kindt. 2004. La lemmatisation des sources patristiques et byzantines au service d'une description lexicale du grec ancien. les principes de formulation des lemmes du dictionnaire automatique grec (d.a.g.). *Byzantion : Revue internationale des études byzantines*, LXXIV:213–272. HAL Id: hal-01018202 (version 1).

Bastien Kindt, Jean-Claude Haelewyck, Andrea Schmidt, and Nicolas Atas. 2018. La concordance bilingue grecque-syriaque des discours de grégoire de nazianze. *Bulletin de l'Académie Belge pour l'Étude des Langues Anciennes et Orientales*, 7:51–80.

Bastien Kindt, Chahan Vidal-Gorène, and Saulo Delle Donne. 2022. Analyse automatique du grec ancien par réseau de neurones. évaluation sur le corpus de thessalonica capta. *Bulletin de l'Académie Belge pour l'Étude des Langues Anciennes et Orientales*, 10-11:537–562.

Enrique Manjavacas, Ákos Kádár, and Mike Kestemont. 2019. Improving lemmatization of non-standard languages with joint learning. 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 1493–1503. Association for Computational Linguistics.

Sheila Teo. 2023. How i won singapore's GPT-4 prompt engineering competition. Towards Data Science. Published: 2023-12-29. Accessed: 2026-01-06.

Gulmira Tolegen, Alymzhan Toleu, and Rustam Mussabayev. 2024. Enhancing low-resource NER via knowledge transfer from LLM. In *Computational Collective Intelligence: 16th International Conference, ICCCI 2024, Proceedings, Part I*, volume 14810 of *Lecture Notes in Computer Science*, pages 238–248. Springer.

UCLouvain – CIOL (Centre d'études orientales – Institut Orientaliste de Louvain). 2022. The online corpus of the GREgORI project. Online. Accessed: 2026-01-06.

Chahan Vidal-Gorène, Victoria Khurshudyan, and Anaïd Donabédian-Demopoulos. 2020. Recycling and comparing morphological annotation models for Armenian diachronic-variational corpus processing. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 90–101, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Chahan Vidal-Gorène and Bastien Kindt. 2020. Lemmatization and POS-tagging process by using joint learning approach: Experimental results on classical Armenian, old Georgian, and Syriac. In *Proceedings of LT4HALA 2020 – 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 22–27.

Chahan Vidal-Gorène, Nadi Tomeh, and Victoria Khurshudyan. 2024. Cross-dialectal transfer and zero-shot learning for Armenian varieties: A comparative analysis of RNNs, transformers and LLMs. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 438–449, Miami, USA. Association for Computational Linguistics.

A Appendix

	In-domain						Out-of-domain					
Model	0-shot	5-shots	50-shots	500-shots	1.000-shots	5.000-shots	0-shot	5-shots	50-shots	500-shots	1.000-shots	5.000-shots
GRC												
GPT4o-mini	0.523	0.767	0.797	0.81	-	-	0.166	0.772	0.768	0.811	-	-
GPT4o	**0.893**	**0.896**	**0.906**	**0.933**	-	-	**0.891**	**0.875**	**0.886**	**0.942**	-	-
GPTo1-mini	0.403	0.49	0.433	0.563	-	-	0.202	0.384	0.361	0.45	-	-
open-mistral-nemo	0.507	0.69	0.763	0.673	-	-	0.434	0.632	0.705	0.659	-	-
mistral-large	0.863	0.63	0.74	0.653	-	-	0.861	0.662	0.699	0.652	-	-
PIE	-	-	-	0.48	0.333	0.566	-	-	-	0.381	0.301	0.453
HYE												
GPT4o-mini	0.647	0.547	0.613	0.63	-	-	0.564	0.541	0.576	0.557	-	-
GPT4o	**0.683**	**0.69**	**0.716**	**0.796**	-	-	**0.67**	**0.721**	**0.712**	**0.761**	-	-
GPTo1-mini	0.497	0.48	0.143	0.49	-	-	0.325	0.465	0.51	0.468	-	-
open-mistral-nemo	0.547	0.497	0.563	0.237	-	-	0.315	0.487	0.576	0.624	-	-
mistral-large	0.643	0.67	0.677	0.697	-	-	0.513	0.599	0.605	0.637	-	-
PIE	-	-	-	0.346	0.47	0.603	-	-	-	0.359	0.468	0.598
KAT												
GPT4o-mini	0.487	0.343	0.423	0.48	-	-	0.384	0.541	0.459	0.469	-	-
GPT4o	**0.656**	**0.66**	**0.686**	**0.76**	-	-	**0.661**	**0.672**	**0.672**	**0.742**	-	-
GPTo1-mini	0.417	0.393	0.34	0.11	-	-	0.482	0.351	0.361	0.462	-	-
open-mistral-nemo	0.523	0.42	0.553	0.603	-	-	0.413	0.426	0.42	0.521	-	-
mistral-large	0.587	0.383	0.583	0.67	-	-	0.531	0.531	0.541	0.616	-	-
PIE	-	-	-	0.393	0.296	0.523	-	-	-	0.347	0.265	0.485
SYC												
GPT4o-mini	**0.323**	0.366	0.476	0.500	-	-	0.123	0.43	0.541	0.541	-	-
GPT4o	0.23	0.49	**0.58**	**0.656**	-	-	**0.254**	**0.511**	**0.591**	**0.641**	-	-
GPTo1-mini	-	-	-	-	-	-	-	-	-	-	-	-
open-mistral-nemo	0.15	**0.493**	0.513	0.556	-	-	0.178	**0.511**	0.531	0.551	-	-
mistral-large	0.166	0.316	0.313	0.343	-	-	0.134	0.344	0.331	0.453	-	-
PIE	-	-	-	0.186	0.162	0.305	-	-	-	0.20	0.093	0.25

Table 5: Lemmatization results (best performances highlighted in bold)

	In-domain						Out-of-domain					
Model	0-shot	5-shots	50-shots	500-shots	1.000-shots	5.000-shots	0-shot	5-shots	50-shots	500-shots	1.000-shots	5.000-shots
GRC												
GPT4o-mini	0.79	0.767	0.9	0.943	-	-	0.742	0.752	**0.911**	0.907	-	-
GPT4o	**0.82**	**0.82**	**0.93**	**0.963**	-	-	**0.811**	**0.833**	**0.911**	**0.937**	-	-
GPTo1-mini	0.74	0.717	0.827	0.903	-	-	0.742	0.702	0.732	0.775	-	-
open-mistral-nemo	0.413	0.63	0.823	0.817	-	-	0.298	0.536	0.762	0.758	-	-
mistral-large	0.823	0.61	0.85	0.793	-	-	0.801	0.556	0.781	0.778	-	-
PIE	-	-	-	0.556	0.546	0.546	-	-	-	0.46	0.45	0.476
HYE												
GPT4o-mini	0.607	0.56	0.653	0.75	-	-	0.615	0.602	0.653	0.662	-	-
GPT4o	**0.773**	**0.786**	**0.776**	**0.863**	-	-	**0.781**	**0.798**	**0.852**	**0.871**	-	-
GPTo1-mini	0.563	0.563	0.19	0.653	-	-	0.408	0.545	0.589	0.615	-	-
open-mistral-nemo	0.507	0.437	0.553	0.3	-	-	0.462	0.417	0.573	0.688	-	-
mistral-large	0.673	0.697	0.717	0.74	-	-	0.669	0.678	0.694	0.729	-	-
PIE	-	-	-	0.366	0.393	0.45	-	-	-	0.35	0.379	0.461
KAT												
GPT4o-mini	0.74	0.547	0.603	0.677	-	-	0.564	0.751	0.639	0.666	-	-
GPT4o	**0.756**	**0.79**	**0.82**	0.796	-	-	0.754	0.732	0.809	0.842	-	-
GPTo1-mini	0.577	0.543	0.49	0.133	-	-	0.774	0.426	0.6	0.577	-	-
open-mistral-nemo	0.457	0.363	0.677	0.673	-	-	0.528	0.643	0.6	0.698	-	-
mistral-large	0.76	0.527	0.767	**0.8**	-	-	**0.836**	**0.793**	**0.852**	**0.869**	-	-
PIE	-	-	-	0.39	0.423	0.516	-	-	-	0.403	0.406	0.561
SYC												
GPT4o-mini	0.327	0.393	0.516	0.566	-	-	0.336	0.499	0.599	0.535	-	-
GPT4o	0.37	**0.61**	**0.61**	**0.863**	-	-	**0.396**	0.549	**0.631**	0.711	-	-
GPTo1-mini	**0.513**	0.513	0.53	0.513	-	-	0.465	0.179	0.163	0.429	-	-
open-mistral-nemo	0.35	0.523	0.566	0.606	-	-	0.272	**0.578**	0.543	0.611	-	-
mistral-large	0.353	0.56	0.546	0.693	-	-	0.362	0.492	0.662	**0.728**	-	-
PIE	-	-	-	0.196	0.423	0.45	-	-	-	0.235	0.431	0.518

Table 6: POS-tagging results (best performances highlighted in bold)

Tag	Explanation	GRC	HYE	KAT	SYC
A	Adjective	✓	✓	✓	✓
ADV	Adverb				✓
AMORPH	Element of morphological analysis	✓			
CARD	Cardinal Number				✓
DET	Article	✓			
ETYM	Etymon	✓			
I+Adv	Adverb	✓	✓	✓	
I+AdvPr	Prepositional Adverb	✓	✓		
I+Conj	Conjunction	✓	✓	✓	
I+Intj	Interjection	✓	✓	✓	
I+Neg	Negation	✓	✓		
I+Part	Particle	✓	✓	✓	
I+Prep	Preposition	✓	✓	✓	
LF	Unanalyzable form chosen as lemma (lemma-form)	✓	✓	✓	
N+Ant	Anthroponymic Name	✓	✓	✓	
N+Com	Common Noun	✓	✓	✓	
N+Epi	Epiclesis (Nickname)	✓	✓		
N+Lettre	Name of a letter	✓			
N+Let	Name of a letter		✓		
N+Pat	Patronymic Name	✓	✓	✓	
N+Prop	Proper Noun	✓	✓	✓	
N+Top	Toponym (Place Name)	✓	✓	✓	
NAME	Proper Name				✓
NAME_ant	Anthroponymic Name				✓
NAME_top	Toponymic Name				✓
NUM+Car	Cardinal Number (word)	✓	✓	✓	
NUM+Ord	Ordinal Number (word)	✓	✓	✓	
NUMA+Car	Cardinal Number (alphanumeric system)	✓	✓	✓	
NUMA+Ord	Ordinal Number (alphanumeric system)	✓	✓	✓	
ORD	Ordinal Number				✓
PART	Particle				✓
PRO+Dem	Demonstrative Pronoun	✓	✓	✓	
PRO+Ind	Indefinite Pronoun	✓	✓	✓	
PRO+Int	Interrogative Pronoun	✓	✓	✓	
PRO+Per1d	Personal Pronoun 1st Person Dual	✓			
PRO+Per1p	Personal Pronoun 1st Person Plural	✓	✓	✓	
PRO+Per1s	Personal Pronoun 1st Person Singular	✓	✓	✓	
PRO+Per2p	Personal Pronoun 2nd Person Plural	✓	✓	✓	
PRO+Per2s	Personal Pronoun 2nd Person Singular	✓	✓	✓	
PRO+Per3p	Personal Pronoun 3rd Person Plural	✓			
PRO+Per3s	Personal Pronoun 3rd Person Singular	✓			
PRO+Pos1p	Possessive Pronoun 1st Person Plural	✓	✓	✓	
PRO+Pos1s	Possessive Pronoun 1st Person Singular	✓	✓	✓	
PRO+Pos2p	Possessive Pronoun 2nd Person Plural	✓	✓	✓	
PRO+Pos2s	Possessive Pronoun 2nd Person Singular	✓	✓	✓	
PRO+Pos3p	Possessive Pronoun 3rd Person Plural	✓		✓	
PRO+Pos3s	Possessive Pronoun 3rd Person Singular	✓		✓	
PRO+Rec	Reciprocal Pronoun	✓	✓	✓	
PRO+Ref	Reflexive Pronoun		✓		
PRO+Ref1s	Reflexive Pronoun 1st Person Singular	✓			
PRO+Ref2s	Reflexive Pronoun 2nd Person Singular	✓			
PRO+Ref3s	Reflexive Pronoun 3rd Person Singular	✓			
PRO+Rel	Relative Pronoun	✓	✓	✓	
PRO_dem	Demonstrative Pronoun				✓
PRO_ind	Indefinite Pronoun				✓
PRO_int	Interrogative Pronoun				✓
PRO_pers	Personal Pronoun				✓
V	Verb	✓	✓		
V+Mas	Mazdar Verb			✓	
V+Part	Participial Verb			✓	
V1–V33	Various Syriac verb forms (Pe'al, Pa"el, etc.)				✓

Table 7: List of POS-tags in GRE*g*ORI datasets