Unlocking Medieval Texts: How Large Language Models Transform POS Tagging for Historical Romance Languages

Anonymous authors
Paper under double-blind review

Abstract

Part-of-speech (POS) tagging for medieval romance languages presents unique challenges due to linguistic variation, historical orthography, and limited annotated resources. This study investigates the effectiveness of large language models (LLMs) in enhancing POS tagging accuracy for three medieval romance languages: Medieval Occitan, Medieval Catalan, and Medieval French. We compare traditional rule-based and statistical approaches (COLaF and UDPipe) with modern open-source LLMs (Gemma3-12B and Phi4-14B). Our methodology encompasses zero-shot and few-shot learning paradigms, fine-tuning experiments, and cross-lingual transfer learning. Using historically significant datasets including the *Nouvelle Acquisition Française 6195* manuscript, *Llibre dels Fets*, and Gui de Chauliac's *Anathomie*, we evaluate the performance gains achievable through neural approaches across different domains. The findings demonstrate that LLMs can significantly improve POS tagging accuracy for medieval texts, showing substantial improvements over traditional taggers. Cross-lingual transfer learning reveals shared linguistic features across medieval romance languages that can be leveraged for better performance on under-resourced historical varieties. These results have important implications for digital humanities research, enabling more accurate downstream tasks such as syntactic parsing, named entity recognition, and diachronic linguistic analysis. We make our codebase, datasets, and models publicly available [1].

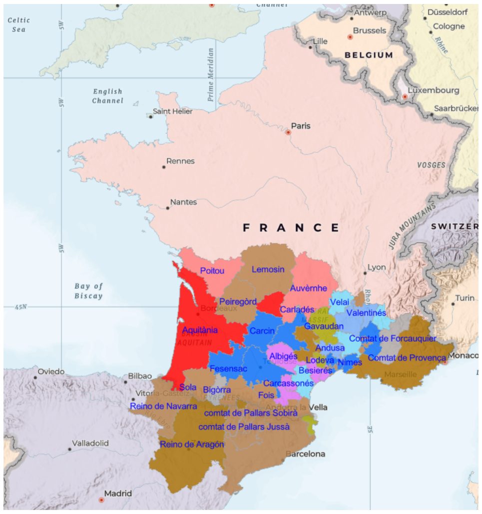
1 Introduction

The computational processing of historical texts represents a critical challenge in digital humanities, where accurate linguistic annotation enables sophisticated analyses of cultural, social, and linguistic evolution. Part-of-speech (POS) tagging, as a fundamental preprocessing step, underpins numerous downstream applications including syntactic parsing, semantic analysis, and diachronic linguistic studies (Piotrowski, 2012; Ehrmann et al., 2020). For medieval romance languages, this task is particularly challenging due to substantial orthographic variation, morphological complexity, and the scarcity of large-scale annotated corpora (Schöffel et al., 2025a).

Medieval romance languages—descendants of Latin that evolved between the 6th and 15th centuries—exhibit significant linguistic diversity (cf. Figure 1) and historical importance. Medieval Occitan served as the literary language of troubadour poetry across western Europe, Medieval Catalan documented the expansion of the Crown of Aragon, and Medieval French preserved extensive literary and administrative records. Despite their cultural significance, these languages remain under-resourced in terms of computational tools and annotated datasets, limiting scholarly research and accessibility of historical documents.

Traditional approaches to POS tagging for historical languages have relied primarily on rule-based systems and statistical models adapted from modern language resources. Tools such as COLaF for Medieval French and Occitan, and UDPipe (Straka et al., 2016) for various languages including Medieval Catalan, have provided baseline performance but face limitations when confronting the linguistic complexities of historical texts. These challenges include non-standardized spelling, dialectal variation, lexical gaps, and morphological ambiguity.

[1] https://anonymous.4open.science/r/medieval-romance-pos-4C8C/README.md



(a) Map of Medieval Occitan and Medieval Catalan variations (13th century)

Spelling variants	Modern/Known spelling
Medieval French	
deffendre	défendre (engl. 'to defend')
joinncture	jointure (engl. 'knuckle')
sun	son (engl. 'his')
pruz	preux (engl. 'brave')
Medieval Catalan	
ssaber	saber (engl. 'to know')
Ffrança	França (engl. 'France')
hòmens	homes (engl. 'men')
jóvens	joves (engl. 'youth')
Medieval Occitan	
deceplina	disciplina, disiplina, desiplina (engl. 'discipline')
falssa	fals (engl. 'false')
liech/lech	lloc (engl. 'place')
fuoc/foc	foc (engl. 'fire')

(b) Spelling characteristics across medieval language variations

Figure 1: Geographic distribution and spelling characteristics of medieval Romance languages (13th century). Left: Map showing Medieval Occitan and Medieval Catalan regional variations Cabré (2014). Right: Comparative analysis of spelling variants across Medieval French, Medieval Catalan, and Medieval Occitan sources.

The emergence of large language models (LLMs) presents new opportunities for improving historical language processing. Recent work has demonstrated the potential of neural approaches for various NLP tasks on historical texts (Bollmann et al., 2019; Manjavacas et al., 2019; Schöffel et al., 2025a). However, systematic evaluation of LLMs for medieval romance language POS tagging remains limited, particularly regarding the comparative effectiveness of different model architectures, prompting strategies, and training paradigms.

This study addresses these gaps through a comprehensive evaluation framework comparing traditional and neural approaches to POS tagging for medieval romance languages. We investigate three key research questions: (1) How do LLMs perform compared to existing tools for medieval romance language POS tagging? (2) What is the relative effectiveness of different prompting strategies and decoding parameters? (3) Can cross-lingual transfer learning improve performance across related medieval languages?

Our contributions include: (1) a systematic comparison of traditional and neural POS tagging approaches for three medieval romance languages, (2) comprehensive evaluation of prompting strategies and decoding parameters for historical language processing, (3) investigation of cross-lingual transfer learning potential across medieval romance varieties, and (4) practical recommendations for implementing LLM-based approaches in historical text processing workflows.

2 Related Work

Historical language processing has evolved from early rule-based approaches to sophisticated statistical and neural methods. Piotrowski (2012) provided foundational work on computational approaches to historical texts, highlighting the unique challenges posed by orthographic variation and linguistic change. Scheible et al. (2011) developed normalization approaches for Early Modern German, demonstrating the importance of preprocessing for historical text analysis. For romance languages specifically, recent work has focused on developing specialized tools and corpora. Camps et al. (2021) introduced methods for lemmatization and POS-tagging of Classical French theatre, demonstrating significant improvements over general-purpose tools when adapted for historical varieties. Their work established benchmarks for evaluation and showed the importance of domain-specific training data for historical language processing.

The application of neural methods to historical languages has gained momentum in recent years. Bollmann et al. (2019) demonstrated that neural networks could improve performance on historical text normalization tasks when applied at scale. Manjavacas et al. (2019) showed that joint learning approaches could enhance lemmatization for non-standard historical varieties, establishing that modern neural techniques could capture historical linguistic patterns effectively. Kestemont et al. (2016) investigated lemmatization for variation-rich languages using deep learning, showing that neural approaches could handle the morphological complexity typical of historical texts. Springmann & Lüdeling (2016) extended this work to OCR post-correction, demonstrating the broader applicability of neural methods to historical text processing pipelines. Garces Arias et al. (2023) proposed a Transformer-based pipeline for HTR to digitize Old Occitan pairs of graphical variants and lemmas, aiming at expanding the DOM dictionary[2]. Furthermore, Schöffel et al. (2025a); Schöffel et al. (2025b), who built upon the dataset released by Wiedner (2025), examined the impact of prompting LLMs on Medieval Romance Languages, highlighting the potential of LLMs for historical language processing. Recent work has explored the application of large language models to various linguistic annotation tasks. Brown et al. (2020) demonstrated the few-shot learning capabilities of large language models, showing promising results for various NLP tasks without task-specific training. Wei et al. (2022) investigated prompting strategies that could elicit reasoning in large language models, establishing best practices for few-shot learning scenarios. For multilingual applications, Müller et al. (2021) demonstrated that multilingual neural models could perform well on historical text translation, while Karthikeyan et al. (2020) showed that cross-lingual transfer learning could improve performance on individual varieties within language families.

3 Methodology

We analyze four distinct tasks: traditional POS tagging, LLM prompting, LLM fine-tuning, and LLM cross-lingual transfer learning (LLM-CLTL). The first task serves as a baseline to establish the capabilities of traditional models. The latter three tasks involve open-source LLMs through different methodologies: prompting with multiple decoding strategies, monolingual and multilingual fine-tuning, enabling us to investigate how exposure to both the target language and syntactically similar languages impacts model performance. Experimental details are presented in Table 2.

3.1 Datasets

We employ three historically relevant datasets representing different medieval romance varieties and textual genres. **Medieval Occitan**: The Nouvelle Acquisition Française 6195 (NAF6195), also known as manuscript M of the *Vida de Sant Honorat*, dating from the 14th century. This manuscript represents Provençal literary tradition and contains approximately 45,457 tokens with manual POS annotations (Wiedner, 2025). **Medieval Catalan**: The *Llibre dels Fets*, a historical chronicle documenting the reign of James I of Aragon, composed in the 13th century. This text represents early administrative Catalan and contains approximately 59,359 tokens with consistent morphological annotation (Pujol i Campeny & Meelen, 2021). **Medieval French**: *Anathomie* from Gui de Chauliac's *Grande Chirurgie*, a 15th-century medical treatise. This technical text provides examples of specialized medieval vocabulary and contains approximately 2,443 tokens with detailed linguistic annotation granted by Tittel (2004)[3].

3.2 Models

We analyze the performance of traditional POS tagger models as baselines for our analysis: COLaF, UDPipe (Straka et al., 2016). Further, we explore the potential of modern open-source LLMs: Gemma3-12B (Gemma-Team et al., 2024a) and Phi4-14B (Abdin et al., 2024a), when conducting prompting, fine-tuning, and multilingual fine-tuning. Excluding COLaF and UDPipe for Medieval French, none of the models has been previously exposed to Medieval Occitan, Catalan, or French. For an overview of representative languages supported by each model, we refer to Table 5 in Appendix A.

[2]https://dom-en-ligne.de/ is the reference dictionary for Medieval Occitan with 79,913 entries, 38,869 unique lemmas, and 41,044 graphical variants as of March 2025.

[3]Each dataset underwent preprocessing including tokenization, sentence segmentation, and manual verification of annotations. We standardized tagsets across languages using Universal Dependencies conventions.

3.3 Prompting Strategies

As detailed in Table 1, we investigate the effect of two prompting strategies: Zero-shot and Few-shot. In **zero-shot**, models receive task instructions and examples without target-domain training data. Prompts include clear task descriptions, tagset definitions, and output format specifications, while in **few-shot**, models receive target-domain examples within prompts. Examples are selected to represent diverse linguistic phenomena, including morphological variations for each language.

Prompting Strategy	Prompt
Zero-shot	*You are a linguistic expert in Medieval Romance languages.* *Analyze the given text and assign Universal Dependencies Part-of-Speech tags (UPOS) to each token.* *Available tags: “ADJ”, “ADP”, “ADV”, “AUX”, “CCONJ”, “DET”, “INTJ”, “NOUN”, “NUM”, “PART”, “PRON”, “PROPN”, “PUNCT”, “SCONJ”, “VERB”, “X”, “SYM”.* *Return a JSON array of objects, each with only “word” and “UPOS” keys.* *Output only the JSON array, properly formatted and closed, with no extra text or explanation.*
Few-shot	***Zero-shot prompt*** *+* *Consider syntactic and semantic relationships, including agreement, word order, and morphology. Medieval Romance languages often exhibit significant spelling variation; for example, Old Occitan: ‘ansy’, ‘eynsi’, or ‘anes’; Old Catalan: ‘fiyl’, or ‘conseyl’; Middle French: ‘norryr’ or ‘norrir’.* *Example format:* `{“word”: “bo”, “UPOS”: “ADJ”}, {“word”: “volch”, “UPOS”: “VERB”}, {“word”: “seyor”, “UPOS”: “NOUN”}, {“word”: “homps”, “UPOS”: “NOUN”}, {“word”: “sant”, “UPOS”: “ADJ”}, {“word”: “iorn”, “UPOS”: “NOUN”}, {“word”: “ilz”, “UPOS”: “PRON”}, {“word”: “addicions”, “UPOS”: “NOUN”}, {“word”: “deffendre”, “UPOS”: “VERB”}]`

Table 1: Comparison of different prompting strategies for UD POS tagging.

3.4 Decoding Strategies

We systematically evaluate the impact of different decoding strategies (Wiher et al., 2022; Garces Arias et al., 2025) on model performance. Specifically, we compare four widely-used approaches: beam search, temperature sampling Ackley et al. (1985), top-k sampling (Fan et al., 2018), and top-p sampling (Holtzman et al., 2019). The complete hyperparameter choices are detailed in Table 2.

3.5 Fine-tuning Experiments

We conduct fine-tuning experiments using two approaches. First, we fine-tune each LLM on individual target datasets using an 80%-20% train-test split. Second, we investigate cross-lingual transfer learning by training models on the combined data from all three datasets and evaluating performance on each target language separately. This cross-lingual approach tests whether shared linguistic features across medieval romance varieties (Blaschke et al., 2025) can improve performance on individual target languages, particularly for under-resourced varieties with limited training data. For the cross-lingual experiments, we maintain the same split ratio across the combined dataset. Detailed hyperparameters are provided in Table 7.

3.6 Evaluation Metrics

We employ standard metrics (cf. Appendix E). **Accuracy**: Percentage of correctly predicted tags across all tokens, providing overall performance assessment. **Macro-averaged F1**: Average F1 score across all POS categories, ensuring balanced evaluation across frequent and rare tags.

3.7 Experimental Setup

Models & Datasets	
Traditional	COLaF, UDPipe
LLMs	Gemma3-12B (Gemma-Team et al., 2024b), Phi4-14B (Abdin et al., 2024b) *Language support in Table 5, Appendix A*
Datasets	NAF (Medieval Occitan, 14th c.), CAT (Medieval Catalan, 13th c.), Chauliac (Medieval French, 15th c.)
Experimental Tasks	
Task 1: Traditional	Direct evaluation using COLaF and UDPipe on all datasets
Task 2: LLM Prompting	Zero-shot & few-shot prompting (Table 1) Decoding: beam search ($w \in \{1, 15\}$), temperature ($\tau \in \{0.6, 0.8, 0.9\}$), top-$k$ ($k \in \{5, 20, 50\}$), top-p ($p \in \{0.75, 0.85, 0.95\}$)
Task 3: LLM Fine-tuning	80/20 train/test split per dataset Each model fine-tuned and tested on same dataset (1-to-1 mapping)
Task 4: LLM CLTF	80% of all datasets for training, 20% per dataset for testing Multilingual training $\rightarrow$ monolingual testing (N-to-1 transfer)

Table 2: Experimental setup for POS tagging of medieval romance languages. Evaluation focused on accuracy with precision, recall, and F1-measures available (Appendix E). All experiments used NVIDIA H100-96GB GPU. Hyperparameters detailed in Appendices B and C.

4 Results

4.1 Overall Performance Comparison

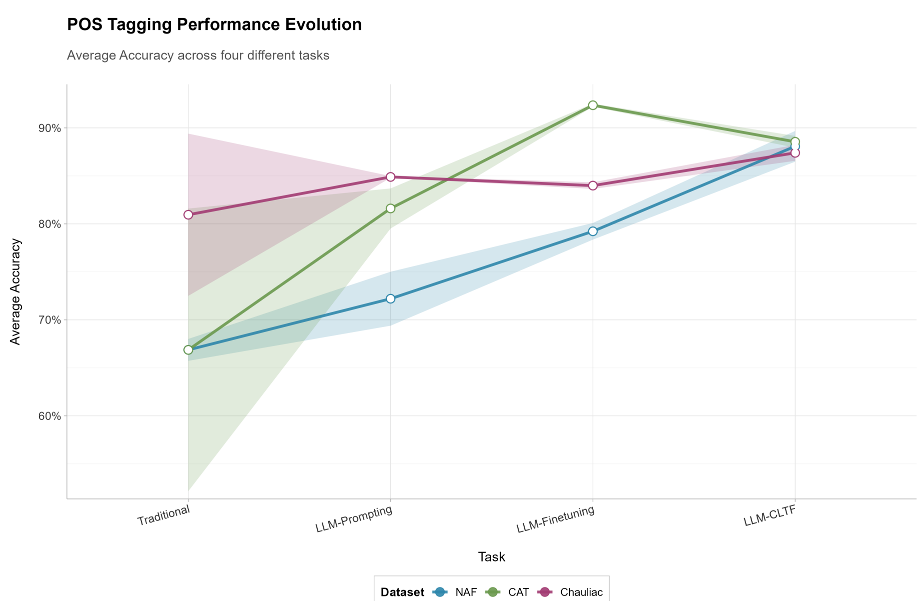


Figure 2: Performance evolution at a dataset level. From traditional POS taggers to multilingual fine-tuning with LLMs. Shaded areas represent variability.

The results demonstrate a clear performance evolution across four distinct tasks, as visualized in Figure 2. Traditional methods achieve 71.56% average accuracy with high variability [61.17%, 81.95%], reflecting inconsistent performance across datasets. LLM-based prompting improves performance to 77.35% [76.17%, 78.52%] with notably reduced variability, indicating more reliable baseline capabilities. The three datasets exhibit distinct performance profiles: CAT shows the largest performance gap between traditional methods and fine-tuned LLMs (10.93 percentage points), followed by NAF and Chauliac (with average performances of approximately 5 percentage points).

LLM fine-tuning represents a substantial advancement, reaching 85.19% average accuracy [80.41%, 89.97%]. This approach demonstrates particular strength on the CAT dataset, where both Gemma3 and Phi4 exceed 92% performance. However, the wider confidence interval suggests sensitivity to dataset characteristics.

The proposed LLM-CLTF task achieves the highest performance at 88.01% average accuracy with the tightest confidence interval [86.96%, 89.06%], indicating both superior effectiveness and remarkable consistency. Compared to traditional UDPipe baseline, CLTF shows substantial improvements on NAF (+21.67%) and CAT (+7.57%), while exhibiting marginal decreases on Chauliac (-1.17%), suggesting dataset-dependent optimization patterns. A detailed overview is presented in Table 3.

The systematic progression from 71.56% (Traditional) through 77.35% (Prompting) and 85.19% (Fine-tuning) to 88.01% (CLTF) illustrates clear methodological advancement, with each approach building upon previous strengths while addressing performance limitations.

Task	Model/Strategy	NAF		CAT		Chauliac	
		Acc.	F1	Acc.	F1	Acc.	F1
Traditional	UDPipe	**68.01**	**67.29**	**81.59**	**81.19**	**89.40**	**89.53**
	COLaF	65.73	65.47	52.15	51.50	72.50	67.43
Prompting	Gemma3 Zero-shot	62.53	61.81	72.54	74.03	82.49	82.58
	Gemma3 Few-shot	69.39	69.22	79.48	80.49	84.80	85.20
	Phi4 Zero-shot	72.78	71.94	80.84	81.01	84.45	84.61
	Phi4 Few-shot	**75.01**	**74.31**	**83.69**	**83.75**	**84.98**	**85.19**
Fine-tuning	Gemma3	**80.09**	**79.99**	**92.52**	**92.50**	83.64	83.74
	Phi4	78.36	78.35	92.20	92.13	**84.33**	**84.10**
CLTF	Gemma3	**89.68**	**89.66**	**89.16**	**89.11**	**88.23**	**88.09**
	Phi4	86.48	86.39	87.94	87.74	86.57	86.48
$\Delta_{CLTF,Traditional}$	Gemma3 vs UDPipe	+21.67	+22.37	+7.57	+7.92	-1.17	-1.44

Table 3: Overall Performance Comparison Across Methods and Datasets. Best result per method is highlighted in **bold**, while best overall results per column are highlighted in **green**.

4.2 Task-Specific Analysis

4.2.1 Traditional vs. LLM-based Approaches

The comparison between traditional POS taggers and LLM-based methods reveals substantial performance gains for LLM approaches. UDPipe, the superior traditional method, achieves competitive performance on Chauliac (89.40% accuracy) but significantly underperforms on NAF (68.01% accuracy), highlighting the challenges posed by the Medieval Occitan dataset.

4.2.2 Prompting Strategy Effectiveness

Few-shot prompting consistently outperforms zero-shot approaches across all datasets, models, and decoding strategies (cf. Fig 3). The performance gains are substantial, ranging from 2.94 percentage points on the Chauliac dataset with Gemma3 to 10.24 percentage points on NAF with Phi4. Among the models tested, Phi4 demonstrates superior prompting capabilities, achieving the best results across all datasets compared to Gemma3. For decoding strategies, deterministic methods proved more effective than sampling-based alternatives, with beam search using a beam width of 15 yielding optimal performance. Comprehensive performance metrics at the dataset level, including variation analysis, are detailed in Tables 8 and 9 (Appendix D.1).

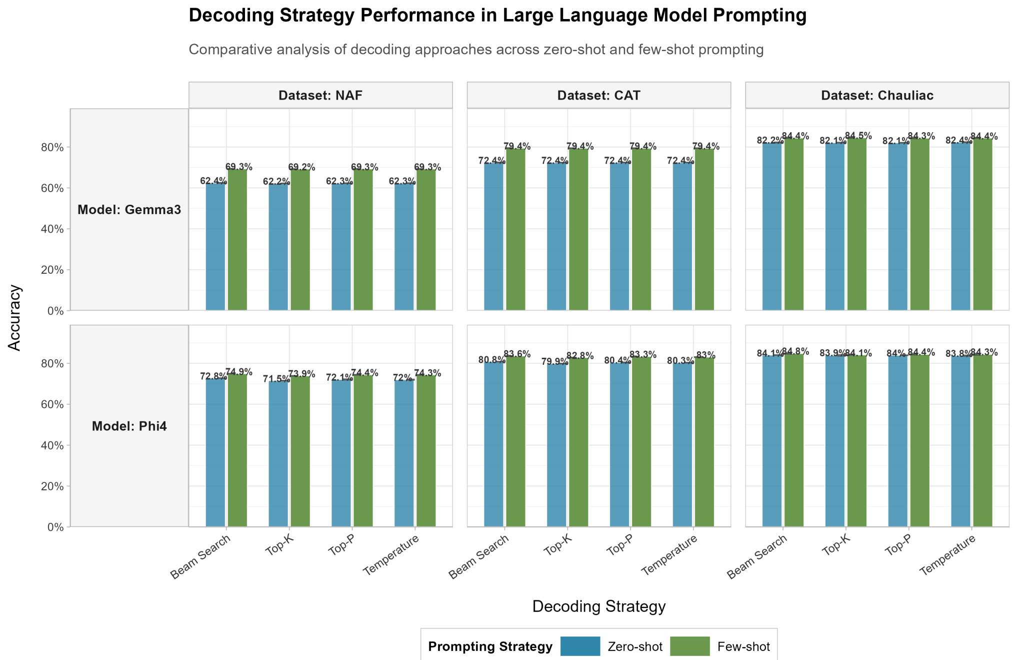


Figure 3: Decoding strategy performance across varying prompts, models and datasets.

4.2.3 *Fine-tuning vs. Cross-Lingual Transfer Learning*

Fine-tuning on individual datasets yields the highest performance for CAT (92.52% with Gemma3), while CLTF demonstrates remarkable effectiveness for NAF, improving accuracy by 9.59 percentage points over single-dataset fine-tuning with Gemma3. This suggests that cross-lingual transfer learning particularly benefits resource-scarce languages like Medieval Occitan. Figure 4 illustrates the effects of LLM-CLTF on a dataset-model level.

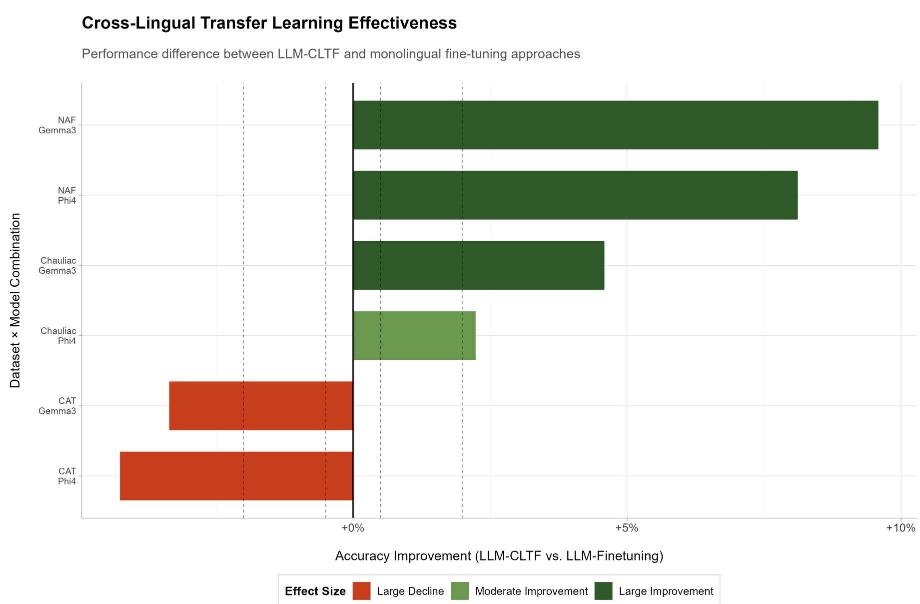


Figure 4: Effect of CLTF with respect to single-dataset LLM finetuning.

5 Error Analysis

5.1 Part-of-Speech Class Performance

Tables 10, 11, and 12 in Appendix D.2 present F1-scores for major POS classes across representative methods, revealing systematic patterns in method effectiveness. Adjectives (ADJ) and adverbs (ADV) present the greatest challenges for traditional methods, with UDPipe achieving F1-scores below 55%. These classes show substantial improvement with LLM-based approaches, particularly fine-tuning, which achieves improvements exceeding 25 percentage points. This pattern suggests that LLMs better capture the contextual nuances necessary for disambiguating these semantically complex categories. Pronouns (PRON) demonstrate consistent improvement across LLM methods, with fine-tuning achieving 84.47% F1-score compared to 62.19% for UDPipe. This improvement likely reflects LLMs' enhanced capacity for processing anaphoric relationships and contextual reference resolution. On the other hand, function words, particularly adpositions (ADP) and coordinating conjunctions (CCONJ), maintain high performance across all methods. UDPipe achieves 94.34% F1-score for ADP, demonstrating that traditional approaches effectively handle these syntactically predictable categories. Verbs show remarkable consistency across methods, with performance ranging from 91.31% (UDPipe) to 93.55% (Phi4 few-shot prompting). This stability suggests that verbal morphology provides sufficient surface-level cues for accurate classification across methodological approaches.

5.2 Cross-Lingual Transfer Effects

Analysis of CLTF results reveals differential benefits across POS classes. Content words (NOUN, ADJ, VERB) show greater improvement from cross-lingual exposure compared to function words, suggesting that semantic representations benefit more from multilingual training than syntactic patterns. The NAF dataset exhibits the most substantial CLTF gains, with accuracy improving from 80.09% (single-dataset fine-tuning) to 89.68% (CLTF). This improvement is particularly pronounced for low-frequency POS classes, indicating that cross-lingual transfer learning effectively addresses data sparsity issues in medieval language processing.

6 Practical Recommendations

6.1 Method Selection Framework

Performance analysis reveals distinct optimal strategies depending on computational resources and target language characteristics, as illustrated in Table 4.

Dataset	High Resources	Limited Resources
NAF (Medieval Occitan)	CLTF (89.68% acc.)	Few-shot Prompting (75.01% acc.)
CAT (Medieval Catalan)	Fine-tuning (92.52% acc.)	Few-shot Prompting (83.69% acc.)
Chauliac (Medieval French)	UDPipe or CLTF (88.23% acc.)	UDPipe (89.40% acc.)

Table 4: Method selection by dataset and computational constraints.

For resource-scarce languages like Medieval Occitan, CLTF provides substantial gains (+21.67 percentage points over traditional methods). Medieval Catalan benefits most from dedicated fine-tuning, while Medieval French technical texts show strong performance with existing traditional tools under resource constraints.

6.2 Implementation Guidelines

Prompting Configuration Few-shot prompting consistently outperforms zero-shot across all conditions, with improvements ranging from 2.94 to 10.24 percentage points. Phi4 demonstrates superior prompting capabilities, achieving 81.23% average accuracy compared to Gemma3's 77.81%. For decoding, beam search with width 15 provides optimal results across all datasets and models.

Cross-Lingual Transfer Learning CLTF shows particular effectiveness for under-resourced varieties. Medieval Occitan achieves the largest improvement (+9.59 percentage points over monolingual fine-tuning), while Medieval Catalan shows marginal gains (+3.36 percentage points). We recommend CLTF in the presence of languages with syntactic similarities and following resource availability.

Performance-Cost Trade-offs The progression from prompting (77.35% average accuracy) to fine-tuning (85.19%) to CLTF (88.01%) represents diminishing returns relative to computational investment. For production systems processing single languages, the 7.84 percentage point improvement from prompting to fine-tuning may justify computational costs. The additional 2.82 percentage point gain from CLTF requires multilingual training data and infrastructure.

6.3 Quality Assurance Considerations

Error analysis reveals systematic performance patterns across POS classes. Content words (ADJ, ADV, PRON) show the largest improvements with neural methods, with F1-score gains exceeding 25 percentage points for adjectives and adverbs. Function words (ADP, CCONJ) maintain consistently high performance (>90% F1) across all methods, suggesting reliable baseline capabilities. For production deployment, we recommend implementing class-specific validation protocols, particularly for content word categories where traditional methods show substantial limitations (ADJ: 54.12% F1 with UDPipe vs. 79.75% with fine-tuned models).

6.4 Resource Allocation Strategy

Based on performance variance analysis, CLTF provides the most stable results across datasets (coefficient of variation: 0.001), while traditional methods show high variability (standard deviation: 10.08 percentage points). For multi-language digital humanities projects, CLTF training followed by language-specific evaluation provides robust performance with predictable resource requirements.

7 Conclusion

This study systematically evaluates large language models for POS tagging across three medieval romance languages, comparing neural approaches with traditional tools through four distinct experimental tasks. Our results demonstrate measurable performance improvements: LLM-based approaches achieve 77.35% average accuracy through prompting, 85.19% through fine-tuning, and 88.01% through cross-lingual transfer learning, compared to 71.56% for traditional methods. Cross-lingual transfer learning shows particular effectiveness for resource-scarce varieties, with Medieval Occitan (NAF) exhibiting a 21.67 percentage point improvement over the traditional UDPipe baseline. Few-shot prompting consistently outperforms zero-shot approaches across all datasets, while beam search with width 15 emerges as the optimal decoding strategy. Our evaluation framework provides systematic guidance for implementing neural approaches to historical language processing. Performance gains vary substantially across POS classes, with content words (adjectives, adverbs, pronouns) showing greater improvements than function words. These findings suggest that LLMs can enhance accuracy for downstream tasks in digital humanities research, including syntactic parsing and diachronic analysis. Future work should examine additional historical language families, investigate the potential of syntactic similarities for optimized cross-lingual transfer learning.

Limitations This study focuses on three medieval romance varieties from specific periods (13th-15th centuries) and domains (literary, administrative, medical), which limits generalizability to other historical language families. Dataset sizes vary considerably (2,443 to 59,359 tokens), reflecting historical corpus constraints but potentially affecting cross-language performance comparisons. The cross-lingual transfer learning approach assumes sufficient linguistic similarity among medieval romance varieties to enable effective knowledge transfer—an assumption supported by historical linguistics but requiring validation for more distantly related languages. Computational requirements for fine-tuning (NVIDIA H100-96GB GPU) may limit accessibility, though our prompting results provide viable alternatives for resource-constrained environments. Our evaluation centers on POS tagging accuracy as a fundamental task, establishing baseline performance for historical language processing. Downstream task improvements remain to be validated in future work.

Acknowledgments

Acknowledgments have been omitted during the anonymization period.

Ethics Statement

This work involves the use of publicly available datasets and does not involve human subjects or any personally identifiable information. We declare that we have no conflicts of interest that could potentially influence the outcomes, interpretations, or conclusions of this research. All funding sources supporting this study are acknowledged in the acknowledgments section. We have made our best effort to document our methodology, experiments, and results accurately and are committed to sharing our code, data, and other relevant resources to foster reproducibility and further advancements in research.

References

Marah Abdin, Jyoti Aneja, Ahmed Awadallah, Abhishek Awasthi, Ankur Bapna, Aseem Bhargava, Sarah Benchekroun, Aaron Bingeman, Shuo Chen, Weizhu Cheng, et al. Phi-4 technical report, 2024a.

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 technical report, 2024b. URL https://arxiv.org/abs/2412.08905.

David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. A learning algorithm for boltzmann machines. *Cognitive science*, 9(1):147–169, 1985.

Verena Blaschke, Masha Fedzechkina, and Maartje ter Hoeve. Analyzing the effect of linguistic similarity on cross-lingual transfer: Tasks and experimental setups matter, 2025. URL https://arxiv.org/abs/2501.14491.

Marcel Bollmann, Florian Petran, and Stefanie Dipper. Large-scale historical text normalization with neural networks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3904–3914. Association for Computational Linguistics, 2019.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

Miriam Cabré. Trob-eu - troubadours and the european identity. https://www.trob-eu.net/, 2014. Accessed: 2025-06-30.

Jean-Baptiste Camps, Simon Gabay, Paul Fièvre, Thibault Clérice, and Florian Cafiero. Corpus and models for lemmatisation and pos-tagging of classical french theatre. *Journal of Data Mining and Digital Humanities*, 2021. doi: 10.46298/jdmdh.6485.

Maud Ehrmann, Matteo Romanello, Alex Flückiger, and Simon Clematide. Named entity recognition and classification in historical documents: a survey. *ACM Computing Surveys*, 53(4):1–47, 2020.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 889–898, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1082. URL https://aclanthology.org/P18-1082.

Esteban Garces Arias, Vallari Pai, Matthias Schöffel, Christian Heumann, and Matthias Aßenmacher. Automatic transcription of handwritten old Occitan language. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15416–15439, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.953. URL https://aclanthology.org/2023.emnlp-main.953/.

Esteban Garces Arias, Meimingwei Li, Christian Heumann, and Matthias Assenmacher. Decoding decoded: Understanding hyperparameter effects in open-ended text generation. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 9992–10020, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.668/.

Gemma-Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology, 2024a.

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. Gemma 2: Improving open language models at a practical size, 2024b. URL https://arxiv.org/abs/2408.00118.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

K Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth. Cross-lingual ability of multilingual bert: An empirical study. In *Proceedings of the 8th International Conference on Learning Representations*, 2020.

Mike Kestemont, Enrique Manjavacas, and Folgert Karsdorp. Lemmatization for variation-rich languages using deep learning. In *Digital Humanities 2016: Conference Abstracts*, pp. 193–195, 2016.

Enrique Manjavacas, Ákos Kádár, and Mike Kestemont. 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*, pp. 1493–1503. Association for Computational Linguistics, 2019.

Lukas Müller, Rico Sennrich, and Martin Volk. First experiments in neural translation of historical texts. In *Proceedings of the Workshop on Language Technologies for Historical and Ancient Languages*, pp. 23–32. Association for Computational Linguistics, 2021.

Michael Piotrowski. *Natural Language Processing for Historical Texts*, volume 5 of *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers, 2012.

Afra Pujol i Campeny and Marieke Meelen. Annotated corpora of historical catalan (hiscat) - llibre dels fets, October 2021. URL https://doi.org/10.5281/zenodo.5615759.

Silke Scheible, Sarah Schulz, and Michael Wojatzki. Token-based chunking of historical german. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pp. 84–89. Association for Computational Linguistics, 2011.

Matthias Schöffel, Esteban Garces Arias, Marinus Wiedner, Paula Ruppert, Meimingwei Li, Christian Heumann, and Matthias Aßenmacher. Unveiling factors for enhanced pos tagging: A study of low-resource medieval romance languages. Under review, 2025b. arXiv preprint, submitted.

Matthias Schöffel, Marinus Wiedner, Esteban Garces Arias, Paula Ruppert, Christian Heumann, and Matthias Aßenmacher. Modern models, medieval texts: A pos tagging study of old occitan, 2025a. URL https://arxiv.org/abs/2503.07827.

Uwe Springmann and Anke Lüdeling. Automatic quality evaluation and (semi-)automatic improvement of mixed models for ocr on historical documents. In *Proceedings of the 12th Conference on Natural Language Processing (KONVENS 2014)*, pp. 179–185, 2016.

Milan Straka, Jan Hajič, and Jana Straková. Udpipe: Trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 4290–4297, Portorož, Slovenia, 2016. European Language Resources Association (ELRA).

Sabine Tittel. Die "anathomie" in der "grande chirurgie" des gui de chauliac : wort- und sachgeschichtliche untersuchungen und edition, 2004.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

Marinus Wiedner. Cometa : Corpus de l'occitan médiéval comparatif et annoté: Provence et languedoc, May 2025. URL https://doi.org/10.5281/zenodo.15300719.

Gian Wiher, Clara Meister, and Ryan Cotterell. On decoding strategies for neural text generators. *Transactions of the Association for Computational Linguistics*, 10:997–1012, 2022. doi: 10.1162/tacl_a_00502. URL https://aclanthology.org/2022.tacl-1.58/.

Appendix

A Supported languages (pre-training)

Language	COLaF	UDPipe	Phi4-14B	Gemma3-12B
Occitan (modern)	✓		✓	
Medieval Occitan				
Catalan (modern)		✓	✓	
Medieval Catalan				
French (modern)	✓	✓	✓	✓
Medieval French	✓	✓		
Spanish (modern)		✓	✓	✓
Italian (modern)		✓	✓	✓
Portuguese (modern)		✓	✓	✓
Romanian (modern)		✓	✓	✓
Galician (modern)		✓	✓	
Asturian (modern)			✓	
Sardinian (modern)			✓	
Sicilian (modern)			✓	
Ligurian (modern)			✓	
Lombard (modern)			✓	
Venetian (modern)			✓	
Friulian (modern)			✓	
Arabic		✓	✓	✓
English		✓	✓	✓

Table 5: Language support (modern vs. medieval) across traditional POS taggers (COLaF, UDPipe) and LLMs (Phi4-14B and Gemma3-12B).

B Hyperparameters for LLM Prompting

Category	Hyperparameter	Value
Tokenizer	Max Length	8192
	Padding Side	left
	Data Type	torch.bfloat16 (Gemma) torch.float16 (Phi-4)
Model	Max New Tokens	300
	Batch Size	8
Processing	Chunk Size	20
	Window Length	5

Table 6: Hyperparameters used for LLM prompting experiments.

412 ## C Hyperparameters for LLM Fine-tuning

Category	Hyperparameter	Value
LoRA	LoRA Rank (r)	16
	LoRA Alpha (α)	32
	LoRA Dropout	0.1
	Target Modules	q_proj, v_proj, k_proj, o_proj
Training	Learning Rate	2×10^{-4}
	Batch Size	4
	Number of Epochs	10
	Optimizer	AdamW
	Weight Decay	0.01

Table 7: Hyperparameters used for LLM fine-tuning experiments with LoRA.

413 ## D Performance Analysis

414 ### D.1 Effect of Decoding Strategies

Model	Strategy	NAF	CAT	Chauliac	Average	Std Dev
Gemma3	Zero-shot + Beam-15	62.53	72.54	82.36	72.48	10.08
	Few-shot + Beam-1	69.24	79.37	84.27	77.63	7.51
	Few-shot + Beam-15	**69.39**	**79.52**	84.51	77.81	7.50
	Few-shot + Top-k-5	69.22	79.48	**84.80**	**77.83**	7.79
	Few-shot + Top-k-20	69.29	79.33	84.35	77.66	7.51
	Few-shot + Top-k-50	69.17	79.41	84.28	77.62	7.56
	Few-shot + Top-p-0.75	69.33	79.47	84.56	77.79	7.62
	Few-shot + Top-p-0.85	69.34	79.42	84.44	77.73	7.54
	Few-shot + Top-p-0.95	69.27	79.31	83.99	77.52	**7.34**
	Few-shot + Temp-0.6	69.23	79.46	84.49	77.73	7.63
	Few-shot + Temp-0.8	69.30	79.35	84.31	77.65	7.48
	Few-shot + Temp-0.9	69.35	79.43	84.39	77.72	7.52
Phi4	Zero-shot + Beam-15	72.77	80.84	84.09	79.23	6.67
	Few-shot + Beam-1	74.86	83.47	84.60	80.98	5.37
	Few-shot + Beam-15	**75.01**	**83.69**	**84.98**	**81.23**	5.32
	Few-shot + Top-k-5	74.02	82.88	84.31	80.40	5.15
	Few-shot + Top-k-20	73.76	82.85	83.98	80.20	5.55
	Few-shot + Top-k-50	73.80	82.81	84.11	80.24	5.21
	Few-shot + Top-p-0.75	74.50	83.46	84.51	80.82	5.53
	Few-shot + Top-p-0.85	74.49	83.34	84.00	80.61	5.43
	Few-shot + Top-p-0.95	74.19	83.16	84.56	80.64	5.70
	Few-shot + Temp-0.6	74.48	83.23	84.53	80.75	5.53
	Few-shot + Temp-0.8	74.27	82.94	83.78	80.33	**4.84**
	Few-shot + Temp-0.9	74.26	82.97	84.69	80.64	5.95

Table 8: Comprehensive decoding strategy performance analysis. Best results per model are highlighted in **bold**, while best overall results per column are highlighted in **green** .

Strategy Type	Mean Acc.	Std Dev.	CV	Range	Recommendation
Phi4 Few-shot					
Beam Search	81.23	0.12	0.001	0.5	Most reliable
Top-k Sampling	80.28	0.20	0.002	1.1	Good alternative
Top-p Sampling	80.69	0.18	0.002	0.8	Balanced performance
Temperature	80.57	0.21	0.003	0.9	Moderate variance
Gemma3 Few-shot					
Beam Search	77.81	0.14	0.002	0.3	Consistent but lower
Top-k Sampling	77.70	0.11	0.001	0.5	Very consistent
Top-p Sampling	77.68	0.14	0.002	0.4	Stable performance
Temperature	77.70	0.08	0.001	0.2	Most consistent

Table 9: Decoding Strategy Robustness and Variance Analysis. CV = Coefficient of Variation (Std Dev / Mean), Range = Max - Min across datasets. Highlighted cells indicate the best combination of performance and stability.

D.2 POS Class Performance

POS Class	UDPipe	Phi4 Few-shot	Gemma3 Fine-tuned	Gemma3 CLTF	Improvement
PROPN	25.85	72.34	78.31	**92.47**	+66.62
NUM	28.92	61.39	**91.89**	86.01	+62.97
AUX	38.39	45.08	53.58	**61.04**	+22.65
PRON	45.80	52.77	76.38	**81.51**	+35.71
ADV	50.61	54.19	66.92	**74.38**	+23.77
SCONJ	52.94	54.73	57.97	**94.62**	+41.68
ADJ	65.29	72.05	71.17	**73.58**	+8.29
VERB	67.77	79.91	75.79	**89.00**	+21.23
DET	73.81	73.48	**89.97**	87.99	+16.16
NOUN	76.45	83.65	82.81	**89.44**	+12.99
CCONJ	83.06	81.72	86.67	**96.34**	+13.28
ADP	85.51	89.93	88.59	**92.78**	+7.27

Table 10: POS Class Performance (F1-scores) on NAF, for low performing (upper section) and high performing (bottom section) classes. Best results per POS class are highlighted in **green**.

POS Class	UDPipe	Phi4 Few-shot	Gemma3 Fine-tuned	Gemma3 CLTF	Improvement
ADJ	54.12	58.11	**79.75**	71.94	+25.63
ADV	51.19	58.79	**77.30**	72.93	+21.74
PRON	62.19	68.66	**84.47**	81.10	+22.28
DET	71.69	74.40	87.32	**89.38**	+17.69
PROPN	79.90	76.26	**98.07**	91.22	+18.17
NOUN	86.14	85.34	**91.84**	88.72	+5.70
VERB	91.31	**93.55**	92.29	87.91	+2.24
ADP	**94.34**	93.09	94.16	92.65	-0.18
CCONJ	95.02	95.86	**98.89**	96.22	+3.87

Table 11: POS Class Performance (F1-scores) on CAT, for low performing (upper section) and high performing (bottom section) classes. Best results per POS class are highlighted in **green**.

POS Class	UDPipe	Phi4 Few-shot	Gemma3 Fine-tuned	Gemma3 CLTF	Improvement
NUM	48.78	60.87	83.33	**88.04**	+39.26
AUX	**56.45**	32.97	40.00	49.30	-7.15
ADJ	68.66	64.52	57.14	**70.27**	+1.61
ADV	75.59	**77.83**	64.15	74.02	+2.24
PROPN	76.19	66.67	75.00	**90.47**	+14.28
VERB	86.32	82.55	67.39	**86.71**	+0.39
PRON	**88.50**	76.66	83.72	79.90	-4.78
DET	**91.79**	82.72	76.47	89.25	-2.54
NOUN	**92.88**	90.87	89.51	88.29	-2.01
ADP	**93.14**	87.70	90.62	92.16	-0.98
CCONJ	93.99	89.42	91.30	**95.65**	+1.66

Table 12: POS Class Performance (F1-scores) on Chauliac, for low performing (upper section) and high performing (bottom section) classes. Best results per POS class are highlighted in **green**.

E Evaluation metrics

We assessed our model using several standard metrics, defined as follows.

Accuracy Accuracy quantifies the proportion of correctly predicted POS tags relative to the total number of tags:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{1}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

Precision Precision measures the fraction of correct POS tag predictions among all instances predicted as a given tag:

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{2}$$

Recall Recall determines the proportion of actual POS tag instances that were correctly predicted:

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{3}$$

F1-score The F1-score, representing the harmonic mean of precision and recall, is computed as:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{4}$$