Modern Models, Medieval Texts: A POS Tagging Study of Old Occitan

Matthias Schöffel[1,2], Marinus Wiedner[3], Esteban Garces Arias[2,4], Paula Ruppert[2], Christian Heumann[2], Matthias Aßenmacher[2,4]

[1]Bavarian Academy of Sciences, [2]LMU Munich, [3]University of Freiburg,
[4]Munich Center for Machine Learning (MCML)

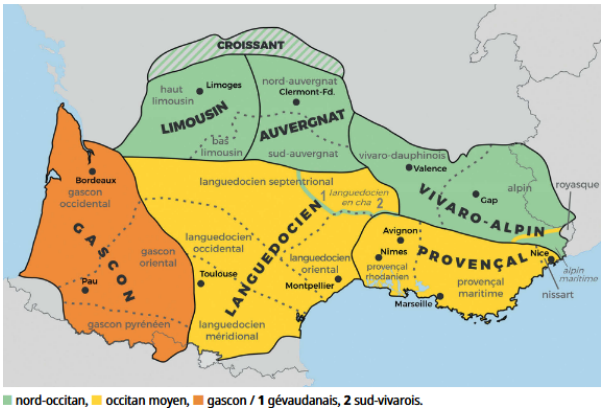
Correspondence: matthias.schoeffel@badw.de

Abstract

Large language models (LLMs) have demonstrated remarkable capabilities in natural language processing, yet their effectiveness in handling historical languages remains largely unexplored. This study examines the performance of open-source LLMs in part-of-speech (POS) tagging for Old Occitan, a historical language characterized by non-standardized orthography and significant diachronic variation. Through comparative analysis of two distinct corpora—hagiographical and medical texts—we evaluate how current models handle the inherent challenges of processing a low-resource historical language. Our findings demonstrate critical limitations in LLM performance when confronted with extreme orthographic and syntactic variability. We provide detailed error analysis and specific recommendations for improving model performance in historical language processing. This research advances our understanding of LLM capabilities in challenging linguistic contexts while offering practical insights for both computational linguistics and historical language studies.

1 Introduction

Old Occitan, also known as Old Provençal, was widely spoken from the 11th to the 16th century across southern France, northeastern Spain, and northwestern Italy (cf. Fig. 1(a)). This language played a pivotal role in shaping both Romance linguistics and medieval European literature, particularly through its renowned troubadour tradition. However, computational analysis and digital preservation of Old Occitan face significant challenges, primarily due to the limited availability of digitized manuscripts and annotated corpora compared to contemporary medieval languages such as Old French (Scrivner and Kübler, 2012). A key obstacle in processing Old Occitan texts is their pronounced orthographic variation, as illustrated in Figure 1(b) through the term *abeurador*



(a) Map of the Occitan-speaking region in southern France, northeastern Spain, and northwestern Italy.

abeurador *Citations*

variante(s): *abeirador*, *abeorador*, *abeorour*, *abeouradour*, *aberadour*, *abeuradé*, *abeurader*, *abeuratorium*, *abeuredee*, *aveurador*

n. m.

'abreuvoir, lieu où l'on mène boire les bestiaux'

(b) Graphical variations in spelling, exemplified by the term *abeurador*, highlighting the challenges posed by non-standardized orthography.

Figure 1: (a) Geographic distribution of Old Occitan with its principal dialect zones (Sibille, 2024). (b) Orthographic diversity in Old Occitan texts, as evidenced by multiple graphical variants of the same term, illustrating inherent challenges for modern LLMs.

('watering place'), which exhibits substantial regional and textual variations in spelling. These variations, while historically significant, present particular challenges for automated text processing tasks such as Part-of-Speech (POS) tagging, which is the focus of the present work.

The imperative for accurate POS tagging in low-resource languages like Old Occitan extends beyond mere technical curiosity. POS tagging is a foundational step in numerous natural language processing (NLP) applications, from syntactic parsing and information extraction to more advanced

Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities, pages 334–349
May 3-4, 2025

tasks in digital humanities. For historical languages, reliable tagging is critical not only for linguistic analysis but also for reconstructing the evolution of language, understanding regional variation, and supporting interdisciplinary research that bridges history and computational methods. Moreover, the performance of large language models (LLM) on such texts offers insights into the adaptability of modern models when confronted with non-standardized data – a challenge that remains largely unaddressed in contemporary NLP research.

In this study, we systematically evaluate a range of LLMs using various prompting strategies – (a) zero-shot, (b) few-shot, and (c) leveraging elaborate instructions – on a corpus comprising 91,953 tokens. Beyond a mere exploration of current capabilities, our work elucidates key factors influencing model performance and offers a rigorous error analysis and practical recommendations to mitigate the effects of input modifications and enhance POS tagging accuracy.

Research Questions: Our study addresses the following research questions: **RQ1:** How effectively can current LLMs perform POS tagging on Old Occitan texts, given the challenges posed by non-standardized orthography and sparse annotated resources? (§5.1) **RQ2:** Which prompting strategy – (a) zero-shot, (b) few-shot, and (c) leveraging elaborate instructions – yields the most robust performance on this low-resource, historical language? (§5.2) **RQ3:** Which specific error patterns and model biases emerge during POS tagging, and how can these insights inform practical improvements? (§6 and §7). By answering these questions, we aim to bridge the gap between modern NLP techniques and the nuanced demands of historical linguistics.

Contributions: We summarize our contributions as follows:

1. We provide the first comprehensive evaluation of multiple LLMs for POS tagging on Old Occitan texts, establishing a robust baseline for historical Romance languages.
2. We systematically compare concrete prompting strategies, including (a) zero-shot, (b) few-shot, and (c) leveraging elaborate instructions, to adapt LLMs to the irregularities of non-standardized historical data.
3. We perform a detailed error analysis to uncover model-specific biases and limitations, offering targeted recommendations to improve POS tagging performance on low-resource texts.
4. We release a novel POS Tagging dataset for Old Occitan, along with our code and experimental results, to facilitate future research in historical NLP.[1]

2 Related work

POS tagging for low-resource languages presents unique challenges that have gained increasing attention in computational linguistics. Several approaches have emerged to address data scarcity in these settings, with varying degrees of success. Cardenas et al. (2019) proposed a grounded unsupervised universal POS tagger for low-resource languages, framing tagging as a clustering problem followed by decipherment-based grounding. This approach requires no labeled training data and demonstrates reasonable performance across diverse languages. Building on this work, Plank et al. (2018) demonstrated that integrating conventional lexical information can significantly improve neural cross-lingual POS tagging, suggesting that even small amounts of symbolic lexical resources can be valuable when gold-standard corpora are unavailable. However, Kann et al. (2020) challenged the effectiveness of weakly supervised approaches for truly low-resource languages. Their evaluation across 15 typologically diverse languages revealed that state-of-the-art weakly supervised POS taggers perform significantly worse under realistic resource constraints than previously reported, with accuracy below 50% for most languages. This skepticism is further supported by Moeller et al. (2021), who found that the presence or absence of POS tags does not significantly impact performance in morphological learning tasks, with some cases showing improved performance when POS tags were removed. For endangered languages specifically, Anastasopoulos et al. (2018) evaluated POS tagging techniques on Griko, achieving 72.9% accuracy through combined semi-supervised methods and cross-lingual transfer. Similarly, Gore and Khatavkar (2022) demonstrated success with the endangered Indian tribal language Katkari, achieving 86.84% accuracy using Hidden Markov Models and the Viterbi algorithm, suggesting that traditional statistical approaches remain viable for low-resource scenarios. Recent work has focused particularly on languages with dialectal variation. The creation of CorpusAr-

[1] https://github.com/msch38/occ_pos_tagging

ièja by Poujade et al. (2024) provides a valuable resource for Occitan, containing 41,000 tokens with POS tags and handling both dialectal and spelling variations. Building on this, Hopton and Aepli (2024) demonstrated that large multilingual models can effectively handle dialectal variation in Occitan without requiring spelling normalization, particularly when fine-tuned for POS tagging. More recently, there have been efforts to ramp up the availability of resources for Old Occitan, including the creation of a digital version of the Old Occitan dictionary [2]at the Bavarian Academy of Sciences. Building on handwritten resources, Garces Arias et al. (2023) tackled automatic transcription, combining a custom-trained Swin image encoder with a BERT-based text decoder to enhance digitization of Old Occitan spelling variations.

3 Data

Our benchmark comprises two corpora drawn from distinct domains: a hagiographical text and a medical treatise. The former is represented by the *Vida de Sant Honorat*, while the latter is embodied by *On surgery and instruments* by Abū l-Qāsim al-Ḥalaf al-Zahrāwī (Albucasis).

For the hagiographical corpus, the primary source is the manuscript Nouvelle Acquisition Française 6195 (NAF6195, also known as manuscript M of the *Vida de Sant Honorat*), preserved at the Bibliothèque Nationale de France. Dated to the 14th century and originating from Provence, this manuscript was first digitised following an archival visit. Its contents were then semi-automatically transcribed using a handwritten text recognition model specifically developed for Old Occitan scripts (Wiedner, 2023) and subsequently subjected to rigorous manual revision. A pre-annotation step was performed with a modern Occitan part-of-speech tagger (Poujade, In progress), after which manual corrections were again applied. The final corpus comprises 44,044 tokens and, to our knowledge, has not previously underpinned any extant editions of the *Vida de Sant Honorat*. A notable linguistic feature of this text is the presence of graphical variants that markedly diverge from those catalogued in the DOM (79,840 entries, 38,861 unique lemmas, and 40,979 graphical variants as of February 2025), as detailed in Table 1.

In contrast, the medical corpus is derived from *On surgery and instruments* by Albucasis. Originally composed in Arabic as one volume of the thirty-volume medical encyclopedia commonly known as al-Tasrif and dating from the late 10th century, the text encompasses nearly 57 chapters and 42,099 word tokens. It was later translated into Latin by Gerard of Cremona at the Toledo School of Translators (circa 1180 AD) and subsequently into vernaculars, including Old French (mid-13th century) and Old Occitan (second quarter of the 14th century). For our purposes, we employed an existing electronic version of the Old Occitan edition (Elsheikh, 1992), originally compiled by P.T. Ricketts, converted to TEI format by Dominique Billy, and released in 2015 under a Creative Commons licence (CC BY-NC-SA 4.0). This edition is based on the manuscript preserved in the Bibliothèque de l'Université (Montpellier), Faculté de médecine, 95. The treatise is distinguished by its specialised technical vocabulary spanning surgery, anatomy, pharmacy, botany, and zoology, and it integrates a mosaic of linguistic influences, including Arabic, Latin, Greek, and vernacular elements. For instance, the Arabic term *taxmir* (connoting 'blepharoplasty')—derived from *tašmir*—is attested in several graphical variants (e.g. atactini, ataxmir, tactimi, tactinir, taxanir).

Both texts were manually annotated following the Universal Dependencies framework[3]. The annotation scheme was constrained to 15 part-of-speech categories (ADJ, ADP, ADV, AUX, CCONJ, DET, INTJ, NOUN, NUM, PRON, PROPN, PUNCT, SCONJ, VERB, and X) owing to the absence of the PART and SYM classes in both corpora. Figure 2 illustrates the part-of-speech distributions across the two texts.

New NAF6195 entry	Available DOM entries
homps (engl. 'man')	ome, om, omen, omne, hom, home
primpce (engl. 'prince')	prince, princep, princip, princer
penedensia (engl. 'penitence')	penedensa, pendensa, pentensa
ompnipotent (engl. 'allmighty')	omnipotent, omnipoten

Table 1: Graphical variants vs. known (DOM) entries.

3.1 Models and Hardware

In this study, we evaluated eight distinct models. Our set comprises the COLaF model (Clérice, 2020; Manjavacas et al., 2019; Nédey et al., 2024; Miletic et al., 2019) – a dedicated POS tagger trained on modern Occitan – alongside seven open-

[2]*DOM: Dictionnaire de l'occitan médiéval* http://www.dom-en-ligne.de/

[3]https://universaldependencies.org/u/pos/

Model	Old Occitan	Occitan	French	Spanish	Italian	Portuguese	Romanian	Arabic	English
COLaF		✓							
Phi4-14B		✓	✓	✓	✓	✓	✓	✓	✓
Mistral-7B									✓
Mistral-Nemo-12B			✓	✓	✓	✓			✓
Gemma2-9B									✓
Mixtral-8x7B			✓	✓	✓				✓
Aya-8B			✓	✓	✓	✓	✓	✓	✓
Qwen2.5-14B			✓	✓	✓	✓		✓	✓

Table 2: Language support across seven open-source instruction-tuned models and COLaF, a dedicated model for POS tagging of modern Occitan.

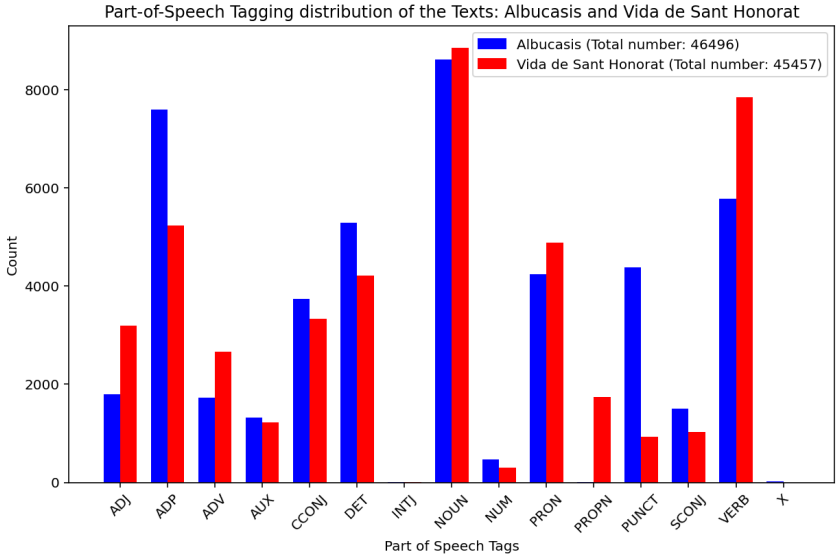


Figure 2: Distribution of Part-of-Speech (POS) tags for Albucasis (blue) and Vida de Sant Honorat (red).

source instruct models that exhibit varying levels of support for Romance languages (Tab. 2). Specifically, the instruct models include Phi4-14B (Abdin et al., 2024), Mistral-7B-Instruct-v0.2, Mistral-Nemo-12B, Mixtral-8x7B (Jiang et al., 2023), Gemma2-9B (Gemma-Team et al., 2024), Aya-8B (Aryabumi et al., 2024), and Qwen2.5-14B (Qwen-Team et al., 2025). Our experiments were conducted employing an NVIDIA Tesla V100-16 GB.

4 Experimental setup

4.1 Prompting strategies

We explore three prompting strategies, each increasing in contextual detail and specificity. The simplest approach, *Zero-shot*, directly instructs the model to assign Universal Dependencies Part-of-Speech tags to each word—without any additional context or expert framing. In *Prompt A*, the instructions are enhanced by explicitly positioning the model as a Medieval Occitan language expert. This prompt emphasizes strict token-by-token processing, ensuring that punctuation is preserved and that the order of words remains unchanged. Finally, *Prompt B* builds upon the previous strategies by incorporating rich linguistic context. It provides explicit examples of spelling variations characteristic of Medieval Occitan (such as variations in the spelling of common words), guiding the model to account for these variations during analysis. Table 5 in Appendix B provides a detailed description.

4.2 Metrics

To evaluate the performance of LLMs in POS tagging for Old Occitan, we focus on widely-used metrics: Accuracy, Precision, Recall and F1-score. Further, we measure the ratio of correctly POS-tagged phrases. A detailed overview on the metrics is provided in Appendix A.

5 Results

Our extensive evaluation of POS tagging in Old Occitan was performed using two datasets with distinct characteristics. The NAF6195 dataset is annotated from a challenging, non-standardized script with 28% unknown vocabulary, whereas Albucasis,

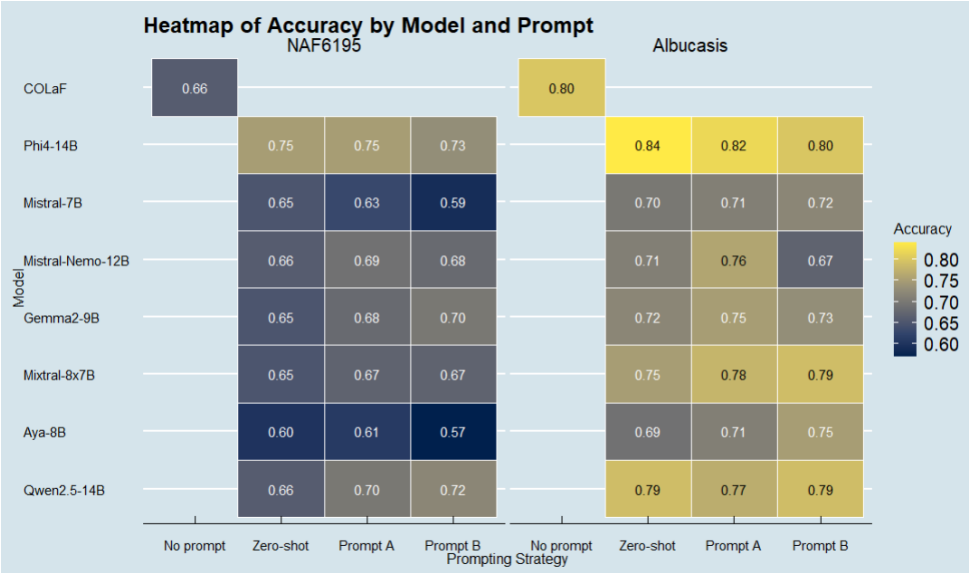


Figure 3: Accuracy heatmap for models and prompting strategies. Results on the left correspond to the NAF6195 dataset and on the right to *Albucasis*.

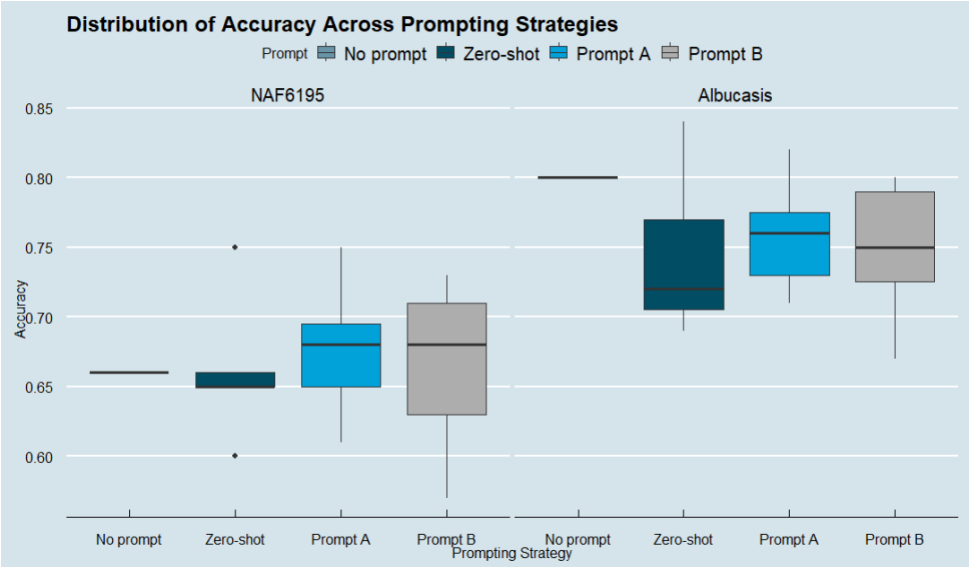


Figure 4: Accuracy distribution across different prompting strategies and datasets. Results on the left correspond to the NAF6195 dataset and on the right to *Albucasis*.

a publicly available resource, exhibits a slightly lower rate of unknown tokens (25%). Tables 6 and 7 (Appendix C) provide a comprehensive summary of POS tagging performance for a diverse set of models and prompting strategies.

5.1 Comparative Performance Across Datasets

Overall, the models achieve higher absolute performance on Albucasis compared to NAF6195. For example, the COLaF baseline, which does not utilize prompting, registers an accuracy of 0.80 on Albucasis compared to 0.66 on NAF6195. Similar trends are observed across micro-averaged Precision, Recall, and F1-score. This divergence is likely attributable to the increased orthographic variability and a larger proportion of unknown vocabulary in NAF6195. Figure 3 further highlights this discrepancy by visualizing the distribution of accuracy scores, revealing a broader spread and lower central tendency for NAF6195.

5.2 Influence of Prompting Strategies

Three prompting configurations were examined: Zero-shot, Prompt A, and Prompt B. In the NAF6195 dataset, a progressive increase in me-

POS Class	Accuracy	Precision		Recall		F1-score	
		NAF6195	Albucasis	NAF6195	Albucasis	NAF6195	Albucasis
ADJ	0.60	0.60	0.49	0.58	0.53	0.59	0.50
ADP	0.79	0.86	0.95	0.79	0.74	**0.81**	0.83
ADV	0.51	0.53	0.51	0.38	0.53	0.42	0.51
AUX	0.58	0.41	0.49	**0.91**	0.71	0.39	0.55
CCONJ	0.77	**0.94**	0.95	0.62	0.79	0.74	**0.85**
DET	0.78	0.59	0.72	0.71	0.79	0.63	0.75
INTJ	0.11	**0.00**	0.11	**0.06**	0.27	**0.00**	0.13
NOUN	0.83	0.77	0.84	0.76	**0.80**	0.76	0.81
NUM	0.69	0.47	0.61	0.39	0.75	0.39	0.65
PRON	0.47	0.57	0.71	0.40	0.46	0.46	0.53
PROPN	0.48	0.42	0.12	0.45	0.59	0.42	**0.10**
PUNCT	**0.99**	0.72	**0.99**	0.59	0.58	0.56	0.70
SCONJ	0.64	0.37	0.60	0.68	0.61	0.43	0.57
VERB	0.65	0.81	0.75	0.68	0.57	0.71	0.64
X	**0.03**	–	**0.01**	–	**0.02**	–	**0.01**

Table 3: Aggregated performance on UD POS tagging classes across datasets, models, and prompting strategies. The highest scores are highlighted in **green**, while lowest scores are highlighted in **red**.

dian accuracy is evident from Zero-shot (0.65) to Prompt B (up to 0.68 for some models), yet the associated variance also increases markedly (cf. Figure 4). This suggests that while Prompt B can boost performance, it does so at the cost of reliability. Conversely, in the Albucasis dataset, despite an overall high variability across prompting configurations, Prompt A emerges as the more balanced strategy. The data in Figure 5 indicate that competitive results are attained by combinations such as Phi4-14B in both Zero-shot and Prompt A modes, COLaF's baseline performance, as well as Qwen2.5-14B and Gemma2 when used with Prompt B. These observations underscore that the optimal prompting strategy is highly contingent on dataset-specific properties.

5.3 POS Class-Level Insights

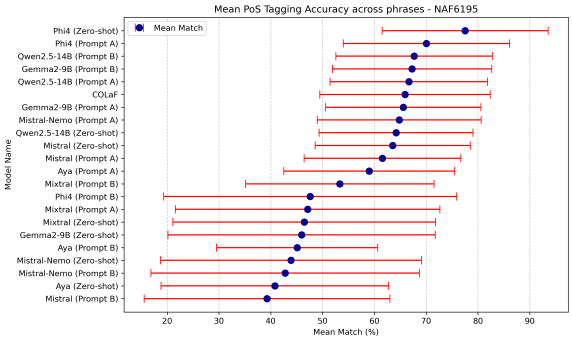
A more granular analysis is provided by the performance metrics on the POS-Tagging class level (cf. Table 3). High-frequency tags such as NOUN and VERB are consistently identified with accuracies of 0.83 and 0.65, respectively, and benefit from robust micro-averaged scores. In contrast, low-frequency tags such as INTJ yield extremely low accuracies (0.11 on NAF6195) and F1-scores that frequently approach zero, indicating a systemic difficulty in recognizing these classes. Moreover, classes like AUX and PROPN exhibit considerable discrepancies between macro- and micro-averaged metrics, hinting at a performance imbalance where errors in infrequent classes are overshadowed by successes in common ones.

5.4 Model Size and Sensitivity Effects

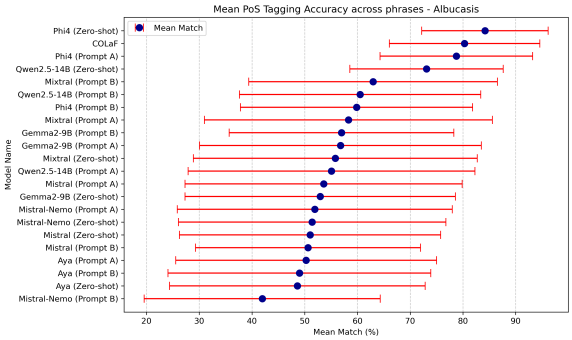
Our study also examines the effect of model scale on tagging performance. Models with larger parameter counts, such as Phi4-14B (14 billion parameters), generally outperform smaller counterparts like Aya-8B (8 billion parameters) across several metrics. Nonetheless, this relationship is moderated by the choice of prompting strategy as well as the supported languages (as illustrated in Table 2). Sensitivity analyses (cf. Figures 6 and 7 in Appendix D) reveal that models including Mistral-7B, Mistral-Nemo-12B, and Aya-8B display heightened responsiveness to the selected prompting configuration, leading to pronounced fluctuations in accuracy and F1-score. Finally, the fact that the mixture-of-experts model, Mixtral-8x7B, does not outperform other competing architectures is an indicator that size alone is not a determinant for enhanced POS tagging accuracy.

5.5 Interplay Between Model Architecture and Data Characteristics

A deeper dive into the inter-model performance comparison reveals that models pre-trained on related high-resource languages (e.g., French, Spanish) exhibit improved robustness when applied to Old Occitan. This is particularly evident in the performance of Phi4-14B and COLaF, which not only deliver competitive results in the Zero-shot setup but also maintain stability when prompted. The variability seen in models like Mistral-7B, especially with Prompt B in the NAF6195 dataset, suggests that the underlying architecture and pre-



(a) Accuracy vs. Prompting for the NAF6195 dataset.



(b) Accuracy vs. Prompting for the Albucasis dataset.

Figure 5: Accuracy across phrases vs. choice of prompting strategies for the NAF6195 and Albucasis datasets.

training corpus substantially influence model behavior in low-resource settings. Trends depicted in Figures 8 and 9 (Appendix E) further corroborate that both model and dataset characteristics jointly determine performance.

6 Error Analysis

In this section, we elucidate underlying causes of misclassifications and identify trends that could inform future improvements.

6.1 POS Class-Specific Error Dynamics

The analysis of Table 3 reveals a marked disparity in performance across different POS classes. High-frequency classes such as NOUN and ADP generally yield high precision and recall; however, classes like INTJ and AUX exhibit critical shortcomings. For instance, the INTJ category in NAF6195 shows an accuracy of merely 0.11, with Precision and Recall values that fail to reach operational thresholds. Such underperformance is indicative of the insufficient representation of these classes during training, compounded by their inherent linguistic ambiguity. Additionally, classes like PROPN display a stark contrast between the two datasets—where Albucasis records a precision as low as 0.12 compared to a higher value in NAF6195—suggesting that contextual or corpus-specific factors play a predominant role in POS class classification.

6.2 Dataset-Specific Error Patterns

The divergence in error profiles between NAF6195 and Albucasis is noteworthy. The NAF6195 dataset's challenging orthographic variations lead to lower overall scores, particularly affecting tags that rely on morphological subtleties (e.g., ADJ, ADV). The higher proportion of unknown vocabulary in NAF6195 exacerbates misclassification rates, as evidenced by lower Recall and F1-scores across multiple classes. Conversely, while Albucasis exhibits a generally higher baseline performance, its variability remains high; this is particularly evident when contrasting the more stable outcomes from Prompt A with the erratic performance of Prompt B. Such dataset-specific discrepancies might indicate the necessity for tailored preprocessing and normalization strategies.

6.3 Cross-lingual transfer and input modifications

A striking outcome of our analysis is that the best-performing model, Phi4, achieves superior POS tagging accuracy despite modifying the input text more frequently and occasionally omitting certain words. In contrast, Mistral—although it tends to preserve the input text more faithfully—consistently exhibits lower accuracy. Phi4 has been trained on multilingual corpora, and our results (Table 4) suggest that it leverages its exposure to Romance languages (including modern Occitan) more effectively, indicating a case of Cross-lingual Transfer Learning (CLTL). Intuitively, one might expect that higher rates of textual modification or omission would yield poor performance; however, the behavior of Phi4 indicates that strategic alterations, informed by multilingual training data, can result in accurate classifications. An illustrative example is the term *ancian* (English: elderly), which Mistral retains in its original form but misclassifies, whereas Phi4 transforms it into *ancià* (from Catalan) and correctly classifies it. This underscores the potential of CLTL, together with prompt engineering strategies that minimize omissions, such as Zero-shot and Prompt A.

Dataset	Model	Prompt	Average Levenshtein	Proportion Changed	Proportion Missing	Average Accuracy
NAF6195	Mistral-7B	Zero-shot	0,97	0,06	0,02	0,65
		Prompt A	0,97	0,05	0,02	0,63
		Prompt B	0,96	0,07	0,03	0,59
	Phi4-14B	Zero-shot	0,91	0,15	0,07	0,75
		Prompt A	0,84	0,23	0,13	0,75
		Prompt B	0,87	0,20	0,11	0,73
Albucasis	Mistral-7B	Zero-shot	0,94	0,10	0,05	0,70
		Prompt A	0,94	0,11	0,05	0,71
		Prompt B	0,91	0,13	0,08	0,72
	Phi4-14B	Zero-shot	0,90	0,15	0,08	0,84
		Prompt A	0,87	0,19	0,11	0,82
		Prompt B	0,86	0,20	0,12	0,80

Table 4: Comparison of Phi4-14B and Mistral-7B in terms of the ratio of changes of original input text, the ratio of omissions, and average accuracy, across the NAF6195 and *Albucasis* datasets.

6.4 Impact of Prompting Variability on Errors

The choice of prompting strategy considerably affects error propagation. In the NAF6195 dataset, while Prompt B occasionally produces higher median accuracies, it also results in a larger spread of errors, as seen in the increased variance of accuracy (Figure 8). This instability is less pronounced in Zero-shot and Prompt A configurations, which consistently produce more reliable outputs. In models with higher sensitivity—specifically Mistral-7B, Mistral-Nemo-12B, and Aya-8B—errors are further magnified when suboptimal prompting is employed. The analysis thus suggests that a careful balance must be struck between leveraging the potential gains of a targeted prompt and maintaining overall model robustness.

6.5 Error Propagation Across Model Architectures

Our sensitivity analysis, as depicted in Figures 6 and 7, indicates that the propagation of errors is not uniformly distributed across model architectures. Larger models such as Phi4-14B tend to contain errors within lower-frequency POS classes, whereas smaller or more sensitive models show a broader dispersion of misclassifications. The inherent variability in performance, particularly under Prompt B conditions, suggests that model architecture and pre-training corpus composition are critical determinants of error propagation in low-resource language processing.

7 Practical Recommendations

Drawing on the detailed results and error analyses, we propose several recommendations to optimize POS tagging in Old Occitan. Our suggestions address model selection, pre-processing strategies, and the tuning of prompting configurations.

7.1 Pre-processing and CLTL

To address the challenges posed by non-standard orthography and high rates of unknown vocabulary, solutions such as integrating pre-processing pipelines might be considered. Techniques such as orthographic normalization, vocabulary expansion using external resources like the DOM (Dictionnaire de l'Occitan Mediéval), and context-aware tokenization are recommended. Further, we observe that models that are exposed to languages of the same family tend to exhibit higher robustness toward spelling and prompting variations. These steps might reduce error rates in classes that require subtle morphological distinctions and improve overall tagging performance.

7.2 Optimizing Prompting Strategies

The data clearly indicate that the choice of prompting strategy influences model outcomes substantially. For datasets with high orthographic variability, such as NAF6195, while Prompt B can offer higher median accuracy, its increased variance necessitates cautious deployment. In contrast, Prompt A has demonstrated a better balance between performance and stability in Albucasis. Practitioners are advised to experiment with multiple prompting configurations during development and to select the one that offers the best trade-off between accuracy and consistency. Furthermore, automated prompt tuning and cross-validation across multiple runs can help in identifying the most robust configuration for a given dataset.

7.3 Model Selection and Configuration

For practitioners aiming to deploy robust POS tagging systems, our findings recommend prioritizing models that demonstrate consistent performance across both Zero-shot and prompted configurations. Models like Phi4-14B and COLaF exhibit superior performance and stability, making them prime candidates for further refinement. Given that larger models tend to perform better but may incur higher computational costs, the choice should balance resource availability with performance needs. Sensitivity analyses further suggest avoiding overly sensitive models, such as Mistral-7B and Aya-8B, unless ensemble methods or targeted fine-tuning strategies are employed to mitigate their variability.

8 Conclusion

This study provides the first systematic evaluation of LLMs for POS tagging in Old Occitan, a highly non-standardized and low-resource historical language. Our findings reveal that while larger models demonstrate some ability to generalize, all tested LLMs struggle with morphological and syntactic inconsistencies due to the lack of training data in similar linguistic contexts. Prompting strategies such as few-shot learning show potential for improving tagging accuracy, yet challenges remain in fine-tuning models for historical text understanding. Furthermore, our error analysis highlights specific areas where LLMs perform poorly, such as handling orthographic variation and a low degree of cross-lingual transfer learning. The insights gained from this work pave the way for further research in historical NLP, emphasizing the need for better-prepared training datasets and refined evaluation methodologies tailored to non-standardized languages. In future work, we plan to extend our analysis to other low-resource languages, including Old French and Medieval Latin, and evaluate the effect of fine-tuning and choice of decoding strategies over the POS tagging quality.

Limitations

While this study offers valuable insights into the application of modern natural language processing techniques to historical, low-resource languages, several limitations must be acknowledged.

Firstly, the analysis is based on a dataset comprised solely of archival Old Occitan texts. Despite considerable efforts to expand the corpus of Old Occitan material (Garces Arias et al., 2025), the inherent scarcity of such sources inevitably constrains the generalisability of our findings.

Secondly, our evaluation was restricted to eight open-source models. Consequently, the performance and potential of additional architectures—and notably, proprietary models—remain to be assessed.

Thirdly, our choice of open-source models was additionally limited due to the hardware requirements. Larger models like Llama 3.3 could therefore not be investigated.

Fourthly, although three prompting strategies of progressively increasing complexity were explored, alternative prompting designs merit further investigation. In particular, the impacts of varying tokenization procedures and the potential benefits of fine-tuning with dedicated Old Occitan corpora are avenues for future research.

Finally, the influence of decoding strategies on the quality of part-of-speech tagging predictions was not fully explored, representing an additional dimension for subsequent studies.

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.

Acknowledgments

We would like to express our sincere gratitude to Viola Baltzer and Verena Harrer for their valuable assistance in preparing and annotating our datasets. Matthias Aßenmacher was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under the National Research Data Infrastructure – NFDI 27/1 - 460037581. Additionally, we thank the Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften (LRZ) for providing computational resources essential for this research.

References

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. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Antonios Anastasopoulos, Maria B. Lekakou, Josep Quer, Eleni Zimianiti, Justin DeBenedetto, and David Chiang. 2018. Part-of-speech tagging on an endangered language: a parallel griko-italian resource. In *International Conference on Computational Linguistics*.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress. *Preprint*, arXiv:2405.15032.

Ronald Cardenas, Ying Lin, Heng Ji, and Jonathan May. 2019. A grounded unsupervised universal part-of-speech tagger for low-resource languages. 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 2428–2439, Minneapolis, Minnesota. Association for Computational Linguistics.

Thibault Clérice. 2020. Pie extended, an extension for pie with pre-processing and post-processing.

Mahmoud Salem Elsheikh, editor. 1992. *Abū'l Qāsim Halaf Ibn 'Abbās az-Zahrāwī, La Chirurgia. Versione occitanica della prima metà del Trecento*. nill, Firenze.

Esteban Garces Arias, Meimingwei Li, Christian Heumann, and Matthias Assenmacher. 2025. Decoding decoded: Understanding hyperparameter effects in open-ended text generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9992–10020, Abu Dhabi, UAE. Association for Computational Linguistics.

Esteban Garces Arias, Vallari Pai, Matthias Schöffel, Christian Heumann, and Matthias Aßenmacher. 2023. Automatic transcription of handwritten old Occitan language. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15416–15439, Singapore. Association for Computational Linguistics.

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. *Preprint*, arXiv:2408.00118.

Toshal Gore and Vaibhav Khatavkar. 2022. Development of part-of-speech tagger for a low-resource endangered language. In *2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, pages 1531–1535.

Zachary Hopton and Noëmi Aepli. 2024. Modeling orthographic variation in Occitan's dialects. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 78–88, Mexico City, Mexico. Association for Computational Linguistics.

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. *Preprint*, arXiv:2310.06825.

Katharina Kann, Ophélie Lacroix, and Anders Søgaard. 2020. Weakly supervised pos taggers perform poorly on truly low-resource languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8066–8073.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Jan Mieszkowski. 2019. *Crises of the Sentence*. University of Chicago Press.

Aleksandra Miletic, Delphine Bernhard, Myriam Bras, Anne-Laure Ligozat, and Marianne Vergez-Couret. 2019. Transformation d'annotations en parties du discours et lemmes vers le format Universal Dependencies : étude de cas pour l'alsacien et l'occitan. In *26e conférence sur le Traitement Automatique des Langues Naturelles (TALN-2019) et 21e édition la conférence jeunes chercheur×euse×s RECITAL*, volume 2 of *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles*, pages 427–435, Toulouse, France. ATALA.

Sarah Moeller, Ling Liu, and Mans Hulden. 2021. To POS tag or not to POS tag: The impact of POS tags on morphological learning in low-resource settings. 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 966–978, Online. Association for Computational Linguistics.

Oriane Nédey, Juliette Janès, Benoît Sagot, Rachel Bawden, and Thibault Clérice. 2024. Modèle occitan (0.0.1).

Barbara Plank, Sigrid Klerke, and Zeljko Agic. 2018. The best of both worlds: Lexical resources to improve low-resource part-of-speech tagging. *Preprint*, arXiv:1811.08757.

Clamenca Poujade. In progress. *La linguistique outillée à l'épreuve de la variation : Ressources pour l'analyse de parlers occitans de l'Ariège*. Ph.D. thesis, Université de Toulouse.

Clamenca Poujade, Myriam Bras, and Assaf Urieli. 2024. CorpusArièja: Building an annotated corpus with variation in Occitan. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 66–71, Torino, Italia. ELRA and ICCL.

Qwen-Team, :, 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. *Preprint*, arXiv:2412.15115.

Olga Scrivner and Sandra Kübler. 2012. Building an old Occitan corpus via cross-Language transfer. In *Proceedings of KONVENS 2012*, pages 392–400. ÖGAI. LThist 2012 workshop.

Jean Sibille. 2024. Les dialectes occitans. In Louise Esher and Jean Sibille, editors, *Manuel de linguistique occitane*, chapter 16, pages 423–471. De Gruyter, Berlin, Boston.

Marinus Wiedner. 2023. Old Occitan handwriting. (modell-nr. 52822, CER=3,51%), PyLaia-Modell for handwritten Occitan from the 13th and 14th century.

Appendix

A Metrics

Accuracy Accuracy measures the proportion of correctly predicted POS tags over the total number of tags:

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

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

Precision Precision evaluates the proportion of correctly predicted POS tags among all predicted instances of a given tag:

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

Recall Recall measures the proportion of correctly predicted POS tags out of all actual instances of that tag:

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

F1-score The F1-score provides a balance between precision and recall and is defined as:

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

Averaging in a Multiclass Setting Given that POS tagging is a multiclass task, the evaluation metrics are computed using different averaging strategies:

- **Micro Averaging:** This method aggregates the contributions of all classes by summing the individual true positives, false positives, and false negatives across all classes. The metrics are then computed from these global counts. As a result, micro averaging is particularly sensitive to the performance on frequent classes.
- **Macro Averaging:** In this approach, the metric is computed independently for each class, and the final score is obtained by taking the arithmetic mean of these per-class metrics. This gives equal weight to each class, thus emphasizing performance on both common and rare classes.
- **Weighted Averaging:** Here, each class's metric is weighted by its support (i.e., the number of true instances). The overall metric is computed as a weighted average of the individual class scores, thereby reflecting the class distribution in the dataset.

RCPTP: Ratio of Correctly POS-Tagged Phrases This metric measures the proportion of phrases without POS Tagging errors:

$$\text{RCPTP} = \frac{\text{Number of correct phrases}}{\text{Total number of phrases}} \tag{5}$$

This metric provides insights into how well LLMs refine and improve initial POS tagging predictions. Note that the term *sentence* or *phrase* is highly ambiguous; we find many different definitions ranging from purely pragmatical or semantical approaches to graphical or intonational definitions (Mieszkowski, 2019). For the purpose of this paper, we employed a syntactical definition based on punctuation: all words between two periods are seen as belonging to one phrase.

B Prompting Strategies

Prompting Strategy	Prompt
Zero-shot	Analyze the provided text and assign to each word Universal Dependencies Part-of-Speech tags: “ADJ", “ADP", “ADV", “AUX", “CCONJ", “DET", “INTJ", “NOUN", “NUM", “PRON", “PROPN", “PUNCT", “SCONJ", “VERB", “X". Return the results as a JSON array of objects, each containing only the ’word’ and ’upos’ keys. The output must be only the JSON array without any additional text, explanations, or formatting.
Prompt A	*You are a Medieval Occitan language expert. Analyze the provided text and assign to each word Universal Dependencies Part-of-Speech tags: “ADJ", “ADP", “ADV", “AUX", “CCONJ", “DET", “INTJ", “NOUN", “NUM", “PRON", “PROPN", “PUNCT", “SCONJ", “VERB", “X". Do not add or remove punctuation or tokens. Ensure to process token by token. Ensure that the order of words in the text is kept for the output. Return the results as a JSON array of objects, each containing only the ’word’ and ’upos’ keys. The output must be only the JSON array without any additional text, explanations, or formatting. Ensure that the JSON array is properly closed.*
Prompt B	*You are a medieval Occitan language expert specializing in linguistic analysis.* *This language is related to Catalan and Latin. In this text there is a high variety of spelling variations having the same meaning* *This is an example for spelling variation:* *homps, ome, om, omen, omne, hom, home.* *Another example is:* *acayson, achaison, acheison, acheson, aqueison, caiso, caison, cason, cayson, chaizo, queison or gaug, gauc, gautz, jau, jauvi.* *Your task is to analyze the given text and assign Universal Dependencies Part-of-Speech (UD POS) tags to each word.* *Return the results as a JSON array of objects, each containing only the ’word’ and ’upos’ keys.* *Ensure that the JSON array is properly formatted and closed.* *The output must be only the JSON array without any additional text, explanations, or formatting*

Table 5: Comparison of different prompting strategies for UD POS tagging.

C Dataset POS Tagging performance

Model	Accuracy	Precision			Recall			F1-score		
		micro	macro	wavg	micro	macro	wavg	micro	macro	wavg
COLaF	0.66	0.66	0.60	0.67	0.66	0.61	0.66	0.66	0.58	0.65
Phi4-14B (Zero-shot)	**0.75**	**0.75**	**0.65**	**0.77**	**0.75**	0.68	**0.75**	**0.75**	**0.66**	**0.75**
Phi4-14B (Prompt A)	**0.75**	**0.75**	0.64	0.76	**0.75**	0.67	**0.75**	**0.75**	0.64	0.74
Phi4-14B (Prompt B)	0.73	0.73	0.63	0.75	0.73	0.62	0.61	0.73	0.61	0.73
Mistral-7B (Zero-shot)	0.65	0.65	0.55	0.67	0.65	0.58	0.65	0.65	0.56	0.65
Mistral-7B (Prompt A)	0.63	0.63	0.55	0.67	0.63	0.56	0.63	0.63	0.54	0.64
Mistral-7B (Prompt B)	0.59	0.59	**0.48**	**0.62**	0.59	0.47	0.59	0.59	**0.41**	0.59
Mistral-Nemo-12B (Zero-shot)	0.66	0.66	0.53	0.71	0.66	0.59	0.66	0.66	0.51	0.67
Mistral-Nemo-12B (Prompt A)	0.69	0.69	0.60	0.73	0.69	**0.69**	0.68	0.69	0.58	0.69
Mistral-Nemo-12B (Prompt B)	0.68	0.68	0.54	0.71	0.68	0.60	0.68	0.68	0.51	0.68
Gemma2-9B (Zero-shot)	0.65	0.65	0.50	0.68	0.65	0.55	0.65	0.65	0.48	0.65
Gemma2-9B (Prompt A)	0.68	0.68	0.55	0.70	0.68	0.58	0.67	0.68	0.55	0.68
Gemma2-9B (Prompt B)	0.70	0.70	**0.65**	0.72	0.70	0.60	0.70	0.70	0.60	0.69
Mixtral-8x7B (Zero-shot)	0.65	0.65	0.60	0.69	0.65	0.56	0.65	0.65	0.56	0.66
Mixtral-8x7B (Prompt A)	0.67	0.67	0.56	0.70	0.67	0.57	0.67	0.67	0.55	0.68
Mixtral-8x7B (Prompt B)	0.67	0.67	0.59	0.70	0.67	0.57	0.67	0.67	0.57	0.68
Aya-8B (Zero-shot)	0.60	0.60	0.50	0.67	0.60	**0.46**	0.60	0.60	0.44	0.62
Aya-8B (Prompt A)	0.61	0.61	0.53	0.66	0.61	0.56	0.61	0.61	0.52	0.62
Aya-8B (Prompt B)	**0.57**	**0.57**	0.52	0.65	**0.57**	0.52	**0.57**	**0.57**	0.49	**0.58**
Qwen2.5-14B (Zero-shot)	0.66	0.66	0.60	0.72	0.66	0.59	0.66	0.66	0.56	0.67
Qwen2.5-14B (Prompt A)	0.70	0.70	0.63	0.75	0.70	0.64	0.70	0.70	0.61	0.71
Qwen2.5-14B (Prompt B)	0.72	0.72	**0.65**	0.75	0.72	0.61	0.72	0.72	0.61	0.71

Table 6: Average scores across all models for the NAF6195 dataset. The highest scores are highlighted in **green**, while lowest scores are highlighted in **red**.

Model	Accuracy	Precision			Recall			F1-score		
		micro	macro	wavg	micro	macro	wavg	micro	macro	wavg
COLaF	0.80	0.80	0.61	0.81	0.80	0.65	0.80	0.80	0.61	0.80
Phi4-14B (Zero-shot)	**0.84**	**0.84**	0.67	**0.87**	**0.84**	**0.77**	**0.84**	0.84	0.69	**0.85**
Phi4-14B (Prompt A)	0.82	0.82	0.67	0.85	0.82	0.74	0.82	0.82	0.67	0.83
Phi4-14B (Prompt B)	0.80	0.80	0.65	0.82	0.80	0.73	0.80	0.80	0.66	0.80
Mistral-7B (Zero-shot)	0.70	0.70	0.57	0.75	0.70	0.63	0.70	0.70	0.55	0.70
Mistral-7B (Prompt A)	0.71	0.71	0.55	0.76	0.71	0.64	0.71	0.71	0.54	0.72
Mistral-7B (Prompt B)	0.72	0.72	0.63	0.77	0.72	0.58	0.72	0.72	0.56	0.73
Mistral-Nemo-12B (Zero-shot)	0.71	0.71	0.57	0.75	0.71	0.68	0.71	0.71	0.58	0.72
Mistral-Nemo-12B (Prompt A)	0.76	0.76	0.62	0.82	0.76	0.66	0.76	0.76	0.57	0.76
Mistral-Nemo-12B (Prompt B)	**0.67**	**0.67**	0.54	**0.74**	**0.67**	0.65	**0.67**	**0.66**	0.56	**0.68**
Gemma2-9B (Zero-shot)	0.72	0.72	0.55	0.75	0.72	**0.56**	0.72	0.72	0.51	0.71
Gemma2-9B (Prompt A)	0.75	0.75	0.57	0.78	0.75	0.62	0.75	0.75	0.55	0.74
Gemma2-9B (Prompt B)	0.73	0.73	0.66	0.77	0.73	0.60	0.73	0.73	0.59	0.71
Mixtral-8x7B (Zero-shot)	0.75	0.75	0.59	0.77	0.74	0.63	0.75	0.75	0.58	0.75
Mixtral-8x7B (Prompt A)	0.78	0.78	0.60	0.79	0.78	0.65	0.78	0.78	0.60	0.78
Mixtral-8x7B (Prompt B)	0.79	0.79	0.66	0.80	0.79	0.68	0.79	0.79	0.66	0.79
Aya-8B (Zero-shot)	0.69	0.69	**0.49**	0.76	0.69	0.57	0.69	0.69	**0.48**	0.71
Aya-8B (Prompt A)	0.71	0.71	0.57	0.79	0.71	0.67	0.71	0.71	0.56	0.73
Aya-8B (Prompt B)	0.75	0.75	0.60	0.81	0.75	0.66	0.75	0.75	0.57	0.75
Qwen2.5-14B (Zero-shot)	0.79	0.79	0.64	0.86	0.79	0.75	0.79	**0.86**	**0.79**	0.82
Qwen2.5-14B (Prompt A)	0.77	0.77	0.60	0.84	0.77	0.73	0.77	0.77	0.59	0.79
Qwen2.5-14B (Prompt B)	0.79	0.79	**0.68**	0.81	0.79	0.75	0.79	0.79	0.68	0.79

Table 7: Average scores across all models for the *Albucasis* dataset. The highest scores are highlighted in **green**, while lowest scores are highlighted in **red**.

D Model sensitivity

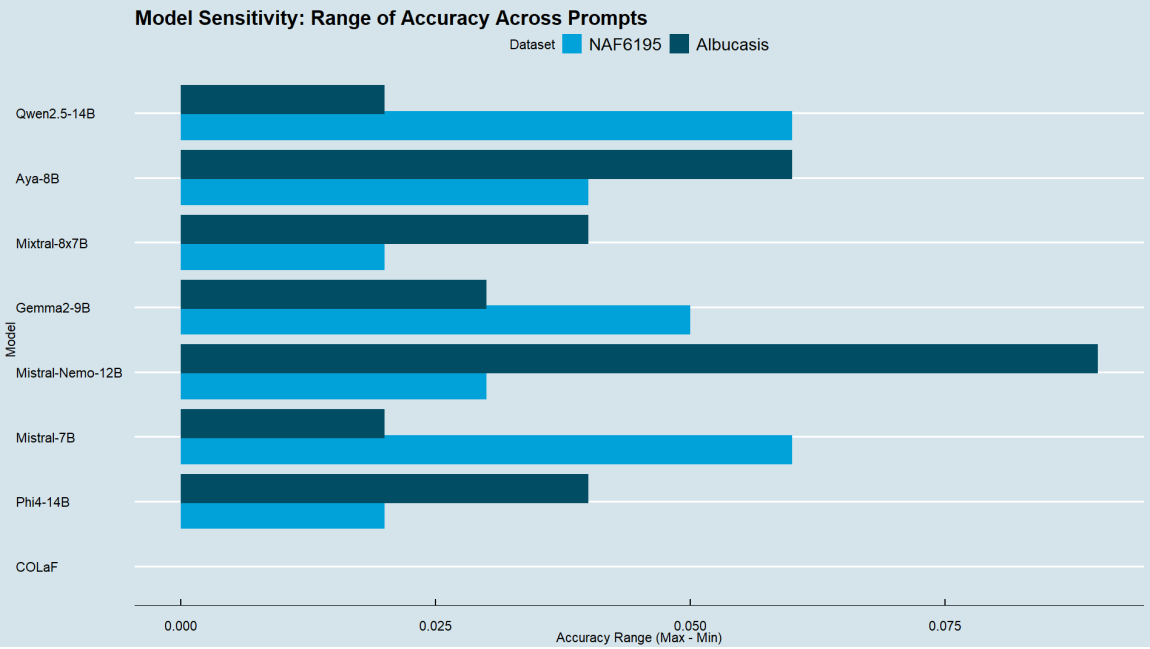


Figure 6: Range of accuracy (Max - Min) per model across prompts.

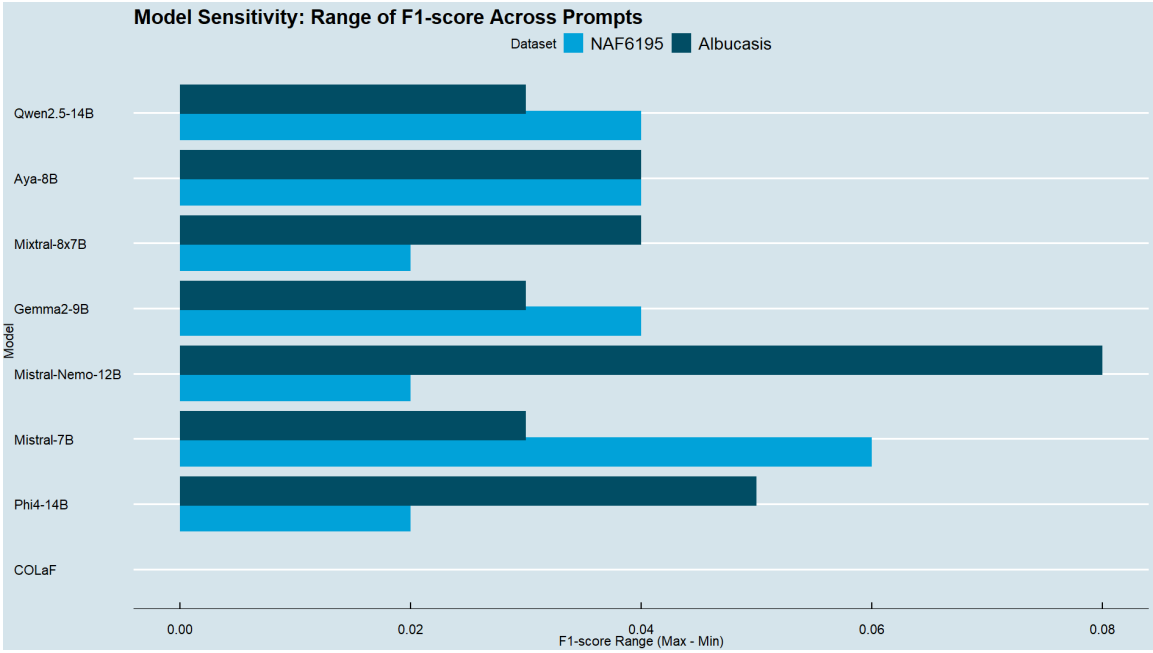


Figure 7: Range of F1-score (Max - Min) per model across prompts.

E Further results

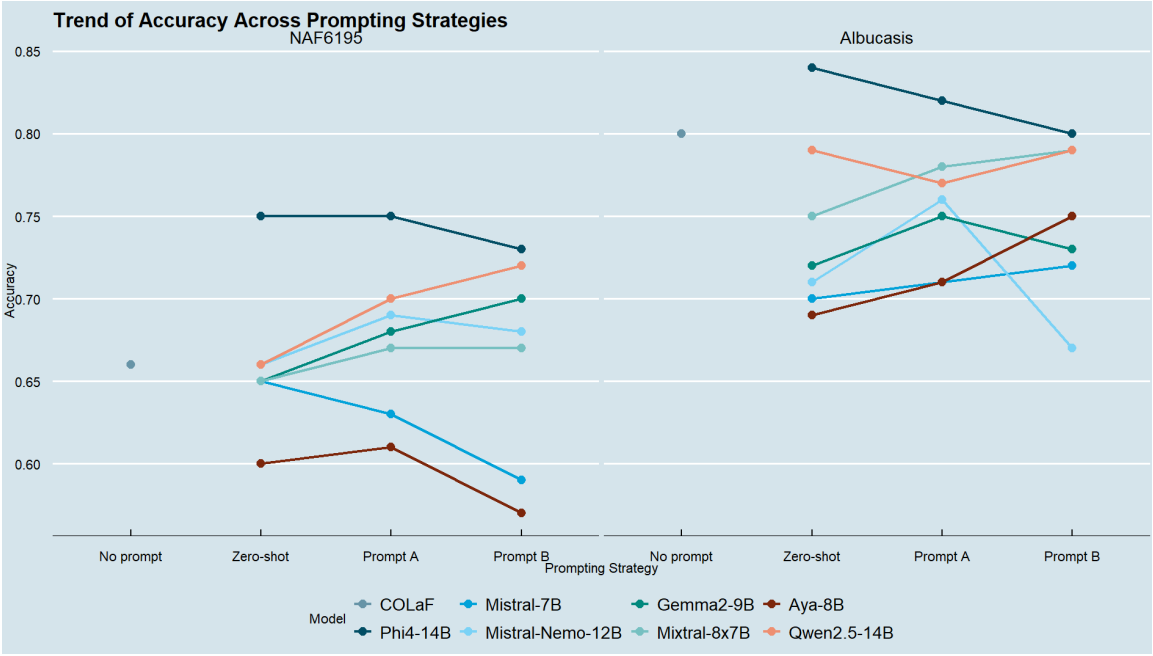


Figure 8: Accuracy behavior vs. choice of prompting strategies.

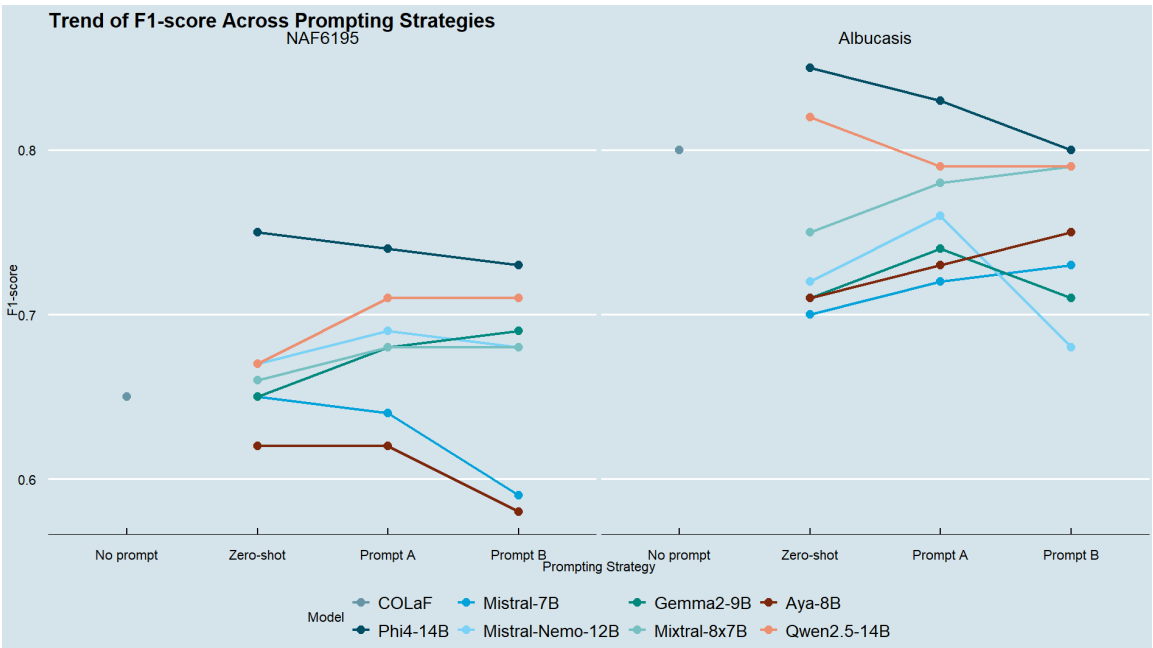


Figure 9: F1-score behavior vs. choice of prompting strategies.