Machine Translation of Low-Resource Historical Languages: A Comparative Study of LLMs on Old Occitan

Francesco Colace[1,†], Sabrina Galano[2,†], Angelo Lorusso[3,†], Domenico Santaniello[4,*,†], Giusy Strollo[1,†] and Annapia Tanzola[1,†]

[1]DIIn, University of Salerno, Fisciano (SA) - Italy
[2]DIPSUM, University of Salerno, Fisciano (SA) - Italy
[3]Department of information science and technology, Pegaso University, Napoli, Italy
[4]DiSPaC, University of Salerno, Fisciano (SA) - Italy

Abstract

Large Language Models (LLMs) are increasingly being used in machine translation. However, their ability to translate low-resource historical languages, such as Old Occitan, remains a significant challenge. This study investigates the translation performance of five LLMs, including both open-source and proprietary models, for translating Old Occitan into Italian, while also evaluating the impact of in-context learning on translation quality. The evaluation was conducted using two COMET-based metrics, complemented by a qualitative analysis of a subset of the translations. The results show that GPT-5 achieves the best performance among the evaluated models, while few-shot prompting yields moderate improvements over zero-shot prompting. Nevertheless, translation errors persist across all models. Furthermore, the qualitative evaluation performed with XCOMET-XL showed limitations when applied to this language pair. This study provides a systematic evaluation of the use of LLMs for translating a low-resource historical language and offers insights into the strengths and limitations of both prompting strategies and automatic evaluation metrics.

Keywords

Old Occitan, Large Language Models (LLMs), Machine Translation, Low-Resource Languages, In-Context Learning

1. Introduction

Large Language Models (LLMs) are artificial intelligence models used for a wide range of natural language processing tasks, including machine translation [2, 3]. They achieve strong performance on high-resource languages but face difficulties when translating low-resource languages [4]. Our study focuses on the use of LLMs for the machine translation of a low-resource language: Old Occitan. Also known as langue *d'oc*, it flourished in the Middle Ages between the eleventh and thirteenth centuries as the language of troubadour literature, before being gradually replaced by French and reduced to a dialectal system still spoken today in southern France [5]. The application of AI techniques to the preservation and valorization of cultural heritage has recently attracted growing scientific interest [6, 7]. This work extends this line of research to the domain of linguistic and literary heritage, focusing on the machine translation of a historical low-resource language. Among the existing methodological approaches for using Large Language Models in machine translation is in-context learning, which allows language models to learn from instructions provided in the prompt. This technique has also been applied to translation scenarios involving low-resource languages [4]. Another important aspect concerns the evaluation of machine translation. Several metrics have been developed: the most widely used standard metric is BLEU, introduced in 2002, which compares the n-grams of the generated translation with those of the reference translation, but has limitations in capturing semantic similarity. [8] Other metrics have been developed to overcome BLEU's limitations, such as METEOR, which is not limited to exact word matches but also recognizes morphological variants and synonyms. [9] In recent years,

NL4AI 2026: Ninth Workshop on Natural Language for Artificial Intelligence, October 6-9th, 2026, Perugia, Italy [1]
*Corresponding author.
†These authors contributed equally.
fcolace@unisa.it (F. Colace); sgalano@unisa.it (S. Galano); angelo.lorusso@unipegaso.it (A. Lorusso); dsantaniello@unisa.it (D. Santaniello); gstrollo@unisa.it (G. Strollo); atanzola@unisa.it (A. Tanzola)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-05

CEUR-WS.org/Vol-4315/paper8.pdf

a major advance has been the development of neural metrics such as COMET, introduced in 2020, which leverages multilingual neural language models to estimate translation quality, moving closer to human judgment. [10] COMET supports Italian, the target language of this study, but not Old Occitan. Consequently, in this work we evaluate the machine translation of Old Occitan into Italian using LLMs. The contributions of this work are as follows:

- We evaluate five LLMs on the translation of literary texts in Old Occitan, in a zero-shot setting and in a few-shot in-context learning setting, where translation examples are provided in the prompt;
- We evaluate the results using the automatic metric COMET;
- We complement this with a qualitative error analysis inspired by the Multidimensional Quality Metric (MQM) taxonomy.

2. Related works

Machine translation based on Large Language Models (LLMs) is a rapidly growing research area, with particular attention to low-resource language settings. In this context, various strategies have been explored to improve model performance, including in-context learning, investigated by Pei et al. for the translation of Manchu, a low-resource language. The results show that the selection of appropriate parallel examples within the prompt can significantly improve model performance, highlighting the crucial role of prompt quality [11]. In parallel, a central role in the evaluation of machine translation systems is played by neural metrics, including COMET, currently one of the metrics with the highest correlation with human judgments [12]. COMET has been trained on human evaluations derived from schemes such as Direct Assessment (DA), Human-mediated Translation Edit Rate (HTER), and Multidimensional Quality Metrics (MQM), thereby learning to approximate the judgments expressed by human annotators. Regarding its architecture, COMET is based on a multilingual encoder, specifically XLM-RoBERTa [10]. However, its applicability remains limited in terms of language coverage. With specific reference to langue d'oc, Schöffel et al. analyzed POS tagging for Old Occitan, showing that prompt quality significantly influences model performance, although challenges related to the morphological and syntactic complexity of the language remain [13]. Subsequent studies have extended the analysis to different tasks. Staffieri evaluated GPT-3.5 on the translation of troubadour poetic texts, observing good semantic adequacy despite some limitations [14]. Di Meo further explored the capabilities of LLMs on medieval Romance languages, including langue d'oc, considering tasks ranging from translation to exegetical and prosodic analysis, highlighting overall promising but not yet fully reliable performance [15].

3. Methodology

This section presents the methodology adopted in this study. The experimental framework is based on a manually curated corpus of Old Occitan texts derived from critical editions and on the comparative evaluation of multiple Large Language Models. The following sections describe the dataset creation and the models employed in the experiments.

3.1. Dataset Creation

The corpus consists of four works in Old Occitan, three narrative and one lyric, all drawn from critical editions. The selected works represent different genres of Old Occitan literature, including narrative and lyric texts, and exhibit different textual organizations, providing a diverse textual dataset for model evaluation.

Blandin di Cornovaglia is an Arthurian verse romance composed by an unknown author around the mid-fourteenth century. The work adopts the metrical structure of the octosyllable with paired rhymes [16].

Las aventuras de monsenher Guillem de la Barra, composed in 1318 by Arnaut Vidal de Castelnaudary, has come down to us through a single manuscript preserved at the Musée Condé in Chantilly. The work, comprising over five thousand octosyllabic verses, unfolds across two distinct thematic levels: a first part of epic-hagiographic character and a second in the register of the courtly and chivalric romance [17].

Jaufre is an Arthurian romance structured according to the typical model of the genre: a first part devoted to the chivalric trial and a second centred on the amorous dimension. The work displays strong intertextuality with the romances of Chrétien de Troyes, with continuous references and reworkings [18].

The songs of Bernart de Ventadorn represent the only lyric text in the corpus. Five songs were selected: *Ara no vei luzir solelh, Ges de chantar no m'pren talans, La dousa votz ai auzida, Lo tems vai e ven e vire and Can vei la lauzeta mover*. The central theme of the lyrics is courtly love, understood as amorous vassalage, characterised by the poet's absolute devotion towards an idealised woman [19].

The Italian reference translations were taken from the corresponding critical editions used as sources for the Old Occitan texts. The translations were segmented to correspond to the respective Old Occitan segments, ensuring alignment between the source text and the reference translation for automatic evaluation. The evaluation corpus is composed of 120 segments in Old Occitan. A reference Italian translation is also provided for each segment. The first half of the sample was extracted consecutively from the opening section of each work, so as to include the introductory sequences. The remaining half was extracted non-consecutively, drawing fragments from central and final sections, thus including representative episodes in order to capture the narrative and stylistic variety present in the texts. Each narrative segment comprises one to three sentences; since these are narrative texts with strong temporal and semantic continuity, the sequences are interconnected and difficult to isolate without loss of meaning. For the translation experiments, each segment was treated as a single input instance. The distribution of segments is as follows: 40 segments for *Blandin di Cornovaglia*, 30 for *Las aventuras de monsenher Guillem de la Barra*, and 30 for *Jaufre*. As for the five songs by Bernart de Ventadorn, the sampling units are the coblas, for a total of 20 segments.

Overall, the corpus contains 8,584 whitespace-delimited tokens in the Old Occitan source texts and 8,695 tokens in the corresponding Italian reference translations. The average segment length is 71.5 tokens for the Old Occitan texts (SD = 25.2; median = 68; range = 29–210) and 72.5 tokens for the Italian reference translations (SD = 27.3; median = 68.5; range = 24–221). Since the exact number of subword tokens depends on the tokenizer adopted by each LLM model, additional tokenizer-specific measurements were performed for the models whose tokenizers were directly available. These yielded approximately 14,200–15,500 tokens for the Old Occitan source corpus and 13,500–15,700 tokens for the Italian reference translations, corresponding to approximately 28,000–31,000 subword tokens for the complete aligned corpus. These results confirm that the source and reference portions of the dataset are closely balanced in terms of overall textual length.

3.2. Models

The five Large Language Models evaluated are: Phi-4-14B, developed by Microsoft; [20] Qwen2.5-VL-32B, developed by Alibaba; [21] GPT-5, developed by OpenAI; [22] Claude Sonnet 4.5, developed by Anthropic; [23] and Mistral Nemo-12B, developed by Mistral AI and NVIDIA [24]. GPT-5 has been employed, in previous versions, in the translation of literary texts both in Occitan and in other languages. Phi-4-14B, Qwen2.5, and Mistral Nemo-12B were selected on the basis of Schoffel's study on POS tagging of Old Occitan; in the present work, the 32-billion-parameter version of Qwen was adopted. Finally, Claude Sonnet 4.5 was included for its multilingual capabilities.

4. Experimental setup

The experimental setup evaluates five Large Language Models on Old Occitan translation. The following sections describe the model access and generation settings, the prompting strategies, and the evaluation

methodology, including both automatic and human-based approaches.

4.1. Model access and generation settings

The models were accessed through two different setups. The open-source models Phi-4-14B, Qwen2.5-VL-32B, and Mistral Nemo-12B were installed and run locally on a workstation and queried through the open-source Open WebUI platform, using Ollama as the backend runtime. For these three models, the generation parameters were kept constant across all experiments: temperature was set to 0.8, Top-k to 40 and Top-p to 0.9. GPT-5 and Claude Sonnet 4.5 were accessed through their respective web-based conversational interfaces, using the default settings available and without modifying the generation parameters. The same translation prompt was used for all five models in order to ensure comparable input conditions.

4.2. Prompting strategies

Two in-context learning techniques were evaluated. In the zero-shot setting, the prompt included only the translation instruction, specifying the language pair. In the few-shot setting, five translation examples were provided before the source text, randomly selected from the Old Occitan work *Flamenca* [25]. Each example consisted of the original Old Occitan text and its corresponding reference translation into Italian, together with the instruction to perform the translation based on the examples provided. The choice of *Flamenca* was motivated by the availability of a critical edition and a corresponding reference translation, as well as by the fact that the work was not included in the corpus used for evaluation. This prevented any overlap between the examples provided to the models in the few-shot setting and the data used for testing. The same five examples, randomly selected from the work, were kept unchanged across all models in the few-shot condition, ensuring consistent prompting conditions and enabling a coherent comparison of the effect of the prompting strategy across models. The full prompts used in the two settings are reported in Table 1.

Table 1
Structure of the prompts used in the translation experiments.

Zero-shot	**Few-shot**
Traduci dall'occitano antico all'italiano: Occitano antico: [text] Italiano: [text]	Traduci dall'occitano antico all'italiano: Esegui il compito basandoti sugli esempi forniti: Occitano antico: [text] Italiano: [text] Occitano antico: [text] Italiano: [text] Occitano antico: [text] Italiano: [text] Occitano antico: [text] Italiano: [text] Occitano antico: [text] Italiano: [text] Occitano antico: [text] Italiano: [text]

4.3. Evaluation

The evaluation of the generated translations is conducted using both automatic and human-based approaches. Automatic evaluation is applied to the full set of translations. Human evaluation is instead performed on a subset of translations and is based on a typological analysis of errors. The two evaluation levels jointly capture both overall translation quality and specific error patterns.

4.3.1. Automatic evaluation

The LLMs were evaluated using the automatic metric COMET, which assesses translation quality from a semantic perspective. COMET is based on XLM-R, which supports Italian as the target language but not Old Occitan as the source language. Two models were employed. The first, wmt22-comet-da, evaluates the translation by comparing it with both the source text and the reference translation, adopting an approach primarily based on the latter [26]. The second, XCOMET-XL, is designed to combine a quality score with error recognition according to a framework consistent with the MQM taxonomy [27].

4.3.2. Human evaluation

A subset of the translations was subjected to qualitative evaluation to complement the results of the COMET metric. The evaluation was inspired by the taxonomy of the Multidimensional Quality Metrics (MQM) framework, from which the categories considered most relevant to the case study were selected. The taxonomy was not used to assign numerical quality scores to the translations, but rather as a structured tool to identify and describe the main types of errors present in the model outputs and to highlight potential qualitative differences across models and prompting strategies. As a preliminary step, the MQM taxonomy was applied to a sample of translations, including a count of the identified errors. Although this approach allowed for a systematic classification of errors, an exhaustive count of errors for each translation proved particularly time-consuming. Therefore, in the final evaluation, the taxonomy was retained as a descriptive framework for qualitative analysis, without systematically counting errors for each output.

The qualitative analysis was conducted on the evaluation sample consisting of 30 segments, translated by the five LLMs according to the two prompting strategies considered, for a total of 300 outputs. The evaluation was carried out by an annotator with expertise in Old Occitan and medieval texts and subsequently reviewed by a second expert. The second expert checked the classifications, contributing to the validation of the qualitative analysis. Since the second expert did not independently annotate the entire sample, no formal inter-annotator agreement measure was calculated. The qualitative evaluation focused on two categories considered relevant to the analysis of the translations produced by the models. In particular, the categories **Accuracy** and **Fluency** were selected because they distinguish two fundamental dimensions of translation quality: the accurate rendering of the meaning of the source text and the correctness, naturalness, and fluency of the Italian output. **Accuracy** was particularly relevant for assessing the accurate transfer of meaning and the difficulties associated with Old Occitan vocabulary, especially in cases where the same lexical form may allow for different interpretations depending on the context. **Fluency**, on the other hand, was considered to assess the models' ability to reformulate the translated content according to the linguistic conventions of Italian, without compromising the correctness and naturalness of the output.

The categories adopted are the following:

Accuracy: concerns cases in which the content does not correctly convey the meaning of the source text. The subcategories considered include omissions, additions, and mistranslations. Inadequate lexical choices due to polysemy were also annotated, namely cases in which a word is translated in a plausible but contextually inappropriate manner.

Fluency: encompasses errors that make the text less fluent and natural, even when the content is correct. This category includes grammatical errors, non-idiomatic constructions, word order that does not conform to Italian syntactic structures, and literal translations.

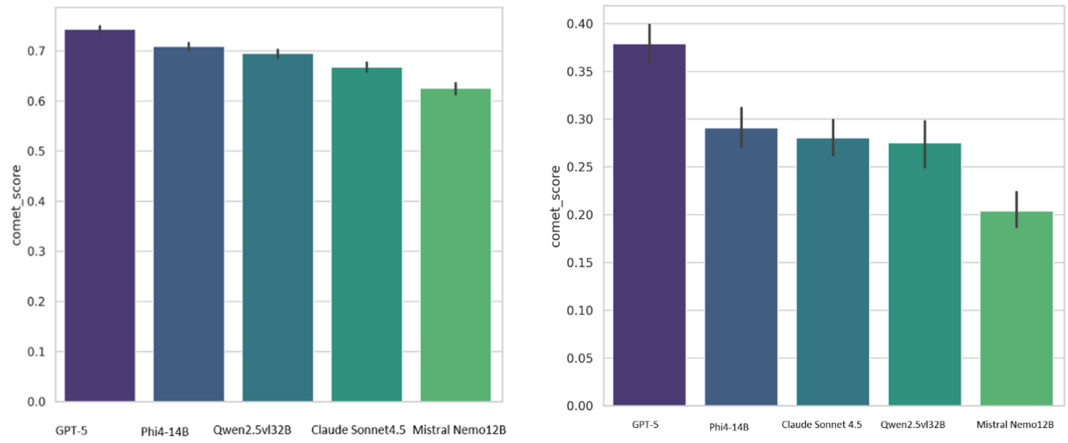


Figure 1: Average scores achieved by the LLMs to wmt22-comet-da (a) and XCOMET-XL (b).

5. Results

The results are shown in Table 2. In the wmt22-comet-da model, GPT-5 confirms itself as the best-performing model on the entire corpus, followed by Phi-4-14B, Qwen2.5-VL-32B, Claude Sonnet 4.5 and Mistral Nemo-12B.The XCOMET-XL model yields overall different scores, while still confirming GPT-5 in first place, followed by Phi-4-14B, Claude Sonnet 4.5 and Qwen2.5-VL-32B. Mistral Nemo-12B achieves the lowest score in both models. The figure 1 shows a graphical representation of the results.

Table 2
COMET scores obtained by the evaluated models.

Model	wmt22-comet-da	XCOMET-XL
GPT-5	0.742	0.378
Claude Sonnet 4.5	0.667	0.280
Mistral Nemo-12B	0.624	0.203
Phi-4-14B	0.708	0.290
Qwen2.5-VL-32B	0.694	0.275

Regarding the impact of the prompting strategy, the results are shown in Table **??**. In the wmt22-comet-da model, the variations between zero-shot and few-shot settings are negligible for GPT-5, Phi-4-14B and Qwen2.5-VL-32B. Claude Sonnet 4.5 (+0.052) and Mistral Nemo-12B (+0.046) benefit most from the inclusion of examples in the prompt. A similar pattern is observed in the XCOMET-XL model, where the most significant variation is recorded for Claude Sonnet 4.5 (+0.067) and Mistral Nemo-12B (+0.051). The Figure 2 shows a graphical representation of the results.

5.1. Results by textual genre

An additional aspect examined was the possible influence of textual genre on model performance. As shown in Figure 3, the three narrative texts obtained higher scores than the lyric text on both evaluation metrics. In particular, for wmt22-comet-da, the narrative texts recorded scores ranging from 0.683 to 0.711 (Text 1: 0.711; Text 2: 0.683; Text 3: 0.696), whereas the lyric text obtained a score of 0.633. A similar, and more pronounced, pattern was observed for XCOMET-XL: the narrative texts obtained scores ranging from 0.226 to 0.405 (Text 1: 0.405; Text 2: 0.226; Text 3: 0.280), whereas the lyric text reached a score of 0.142.

Table 3
Evaluation results of different models using zero-shot and few-shot prompting strategies.

Model	Strategy	wmt22-comet-da	XCOMET-XL
GPT-5	Zero-shot	0.742	0.378
	Few-shot	0.743	0.379
Claude Sonnet 4.5	Zero-shot	0.641	0.246
	Few-shot	0.693	0.313
Mistral Nemo-12B	Zero-shot	0.601	0.178
	Few-shot	0.647	0.229
Phi-4-14B	Zero-shot	0.700	0.289
	Few-shot	0.716	0.292
Qwen2.5-VL-32B	Zero-shot	0.692	0.276
	Few-shot	0.697	0.273

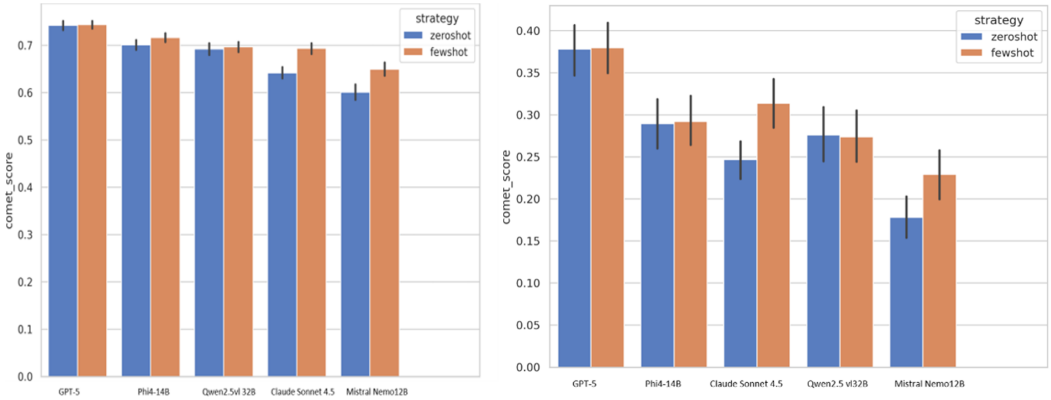


Figure 2: Comparison of zero-shot and few-shot performance for each LLM according to wmt22-comet-da (a) and XCOMET-XL (b).

Overall, these results suggest a possible association between textual genre and translation performance, with the narrative texts consistently showing higher scores than the lyric text. However, this difference should be interpreted with caution, as the comparison is based on three narrative texts and only one lyric text and therefore does not allow the effect of genre to be isolated from the specific characteristics of the individual works. The results reported in the figures reflect model performance across the corpus as a whole, without distinguishing between the zero-shot and few-shot conditions. Therefore, the specific contribution of the few-shot examples cannot be isolated. Moreover, since all examples used in the few-shot setting were selected from the narrative work *Flamenca*, the genre of the examples provided in the prompts represents a potential contributing factor to the observed difference between the narrative and lyric texts. Accordingly, the results should be interpreted by considering both possible genre-related characteristics and the potential influence of the distribution of the examples used in the few-shot setting.

5.2. Qualitative analysis of XCOMET-XL output

XCOMET-XL makes errors in its annotations, in some cases flagging non-existent errors or classifying them inaccurately. This is reflected in overall lower average scores compared to the other COMET model adopted. Table 4 reports a representative example.

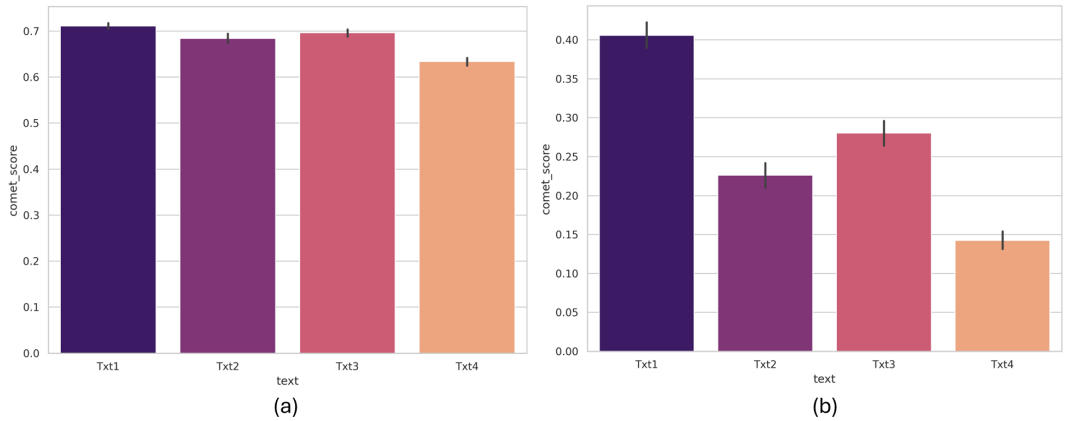


Figure 3: Translation difficulty for text in the wmt22-comet-da model (a) and XCOMET-XL model (b).

Table 4
Example of XCOMET-XL analysis. The Reference column reports the reference translation taken from the critical edition of Blandin di Cornovaglia, edited by S. Galano (2004) [16].

MT	Allora si meravigliarono quando trovarono il segugio e dissero l'un l'altro: «Questo è un'indubbia avventura!» Allora Blandinet disse: «Seguiamolo fino a notte fonda, e vedremo quale cammino prenderà o quale avventura ci mostrerà.» Allora il segugio, correndo veloce, entrò in un torrente e trovò una cavità che si addentrava nella terra. Si infilò al suo interno così rapidamente che da quel momento nessuno poté più vederlo.
Reference translation	Allora si meravigliarono quando trovarono in quel luogo il segugio e si dissero: «Questa è senza dubbio un'avventura!» Allora Blandin aggiunse: «Seguiamolo fino a notte e vedremo quale strada farà e a quale avventura ci condurrà.» Allora il cane, di corsa, entrò in un torrente, e qui trovò una grotta scavata nella terra, e infila dentro la testa per prima dopodiché non lo si poté più scorgere.
Output	Major errors: • segugio
Score	0.375
Comment	The noun "segugio" is classified as a major error, despite the translation being correct. Some annotations furthermore concern elements that do not alter the overall meaning of the translation. The translation produced is on the whole faithful to the original; however, the COMET score is very low, most likely due to the inaccurate annotations.

5.3. Error analysis zero-shot

Phi-4-14B exhibits limitations primarily in terms of accuracy, with mistranslations and omissions that fail to adequately convey the original content. Regarding fluency, the model tends to produce overly literal translations and word order that closely mirrors the structure of the source language. Qwen2.5-VL-32B shows accuracy-related limitations, including mistranslations, omissions, and additions that are not present in the source text. In terms of fluency, the model displays changes in verb tenses compared with the reference translation and uses word order that is often non-idiomatic. Claude Sonnet 4.5 presents accuracy issues similar to those of the previous models, including mistranslations, omissions, and untranslated words. With respect to fluency, the model adopts a literal translation approach and produces unnatural word order. Changes in verb tenses are also observed. GPT-5 demonstrates a more balanced approach between fidelity and readability. It preserves the main content more accurately while producing smoother and more fluent translations. Nevertheless, some inaccuracies remain, including additions, mistranslations, and changes in verb tenses, most of which are of minor severity. Mistral Nemo-12B exhibits the highest number of errors, both in terms of accuracy, with mistranslations,

omissions, and untranslated words, and fluency, with a highly literal translation approach that results in unnatural word order and syntax, accompanied by morphological errors. Major-severity errors predominate.

5.4. Few-shot error analysis

The qualitative analysis shows that few-shot prompting leads to the most noticeable improvements in terms of fluency. Models that already produced relatively fluent translations in the zero-shot setting—Phi-4-14B, Qwen2.5-VL-32B, and GPT-5—exhibit more limited gains, consistent with what is also observed in the COMET scores. Table 5 e 6 reports a representative example comparing translations produced in the zero-shot and few-shot settings, along with a qualitative commentary on the observed errors.

Table 5
Error analysis in zero-shot and few-shot for proprietary models.

Model	Zero-shot	Comment	Few-shot	Comment
Claude Sonnet 4.5	E quando ebbe un gran tempo andato, egli vide una grande chiarezza e là lontano c'era una dimora nella quale c'era un assai bel portale. A quel portale egli se ne andò e lì un portiere trovò che gli aprì subito la porta e disse: «Entrate in questa corte ché voi qui troverete avventura, se la volete.»	In terms of fluency, the machine translation "quando ebbe un gran tempo andato" sounds unnatural in Italian. Furthermore, the constructions "a quel portale egli se ne andò" and "lì un portiere trovò" exhibit unidiomatic word order. Regarding accuracy, the term "orta" is translated as the noun "corte," resulting in an incorrect lexical choice.	E quando ebbe a lungo camminato, egli vide una grande chiarezza e laggiù lontano vide un palazzo nel quale c'era un bellissimo portale. A quel portale egli se ne andò e lì trovò un portiere che gli aprì subito la porta e disse: «Entrate in questo orto ché voi qui troverete avventura, se la volete.»	The translation presents a more natural syntax. For example, the strongly literal rendering "quando ebbe un gran tempo andato" is reformulated as "quando ebbe a lungo camminato." Furthermore, in this case the word order is "lì trovò un portiere," which sounds more natural.
GPT-5	E quando ebbe viaggiato per molto tempo, vide una grande chiarità, e là in lontananza c'era un palazzo che aveva un portale molto bello. A quel portale si avvicinò e lì trovò un portinaio, che subito gli aprì la porta e gli disse: «Entra in questo luogo, perché qui troverai un'avventura, se la vorrai.»	The machine translation shows a mistraslation error, as the term "giardino" is incorrectly rendered as "luogo."	E dopo aver camminato a lungo, vide in lontananza una grande luce, e là, più avanti, una dimora con un portale magnifico. Si avvicinò a quel portale e vi trovò un portinaio, che subito gli aprì la porta e gli disse: «Entra in questo giardino, perché qui troverai un'avventura, se la desideri.»	Compared to the zero-shot setting, it correctly translates "ort" as "giardino."

5.5. Hallucination

In the few-shot setting, Phi-4-14B model renders "Jaufre" as "Gauvain". This can be classified as an extrinsic hallucination, since "Gauvain" is neither present in the source text nor in the prompt examples. Although both Jaufre and Gauvain are knights of the Round Table, they are distinct characters; therefore, the translation introduces a figure belonging to the Arthurian cycle. The error is reported in Table 7. This behaviour may be influenced by prior knowledge encoded during pre-training.

6. Conclusion

This study provides an evaluation of Large Language Models (LLMs) for machine translation from Old Occitan into Italian, examining the impact of prompting while combining automatic evaluation with qualitative analysis. The results show that, although GPT-5 achieves the best overall performance, translation errors persist in terms of both accuracy and fluency. Compared with the zero-shot setting, providing examples in the prompt primarily improves translation fluency and, in some cases, leads to

Table 6
Error analysis in zero-shot and few-shot for open-source models.

Model	Zero-shot	Comment	Few-shot	Comment
Phi-4-14B	E dopo aver viaggiato per molto tempo, vide una grande luce e non distante da lui c'era un ostello con un bellissimo portale. Al portale si diresse e lì incontrò un portiere che gli aprì immediatamente la porta e disse: «Entra in questa alta dimora perché qui troverai avventura, se vuoi».	There are accuracy issues: the noun "giardino" is translated as "dimora," and the location expression "in lontananza" as "non distante da lui." Furthermore, the polysemous lemma "hostal" is translated as "ostello," a lexical choice that is not appropriate in this context. In terms of fluency, the construction "al portale si diresse" shows a word order that is not idiomatic.	E dopo aver percorso un lungo tratto, vide una grande luce e in quella direzione c'era una casa con un portone molto bello. Al portone si avvicinò ed incontrò un guardiano che aprì subito la porta e disse: «Entra in questa dimora perché qui troverai avventura se vuoi».	The translation of "hostal" as "casa" is more appropriate in this context and matches the noun used in the reference translation; however, the rendering of "giardino" as "dimora" still remains.
Qwen2.5-VL-32B	Dopo essere andato per molto tempo, vide una grande luce e, più avanti, un ostello con un bellissimo portale. Si avvicinò al portale e trovò un portiere che gli aprì subito la porta e disse: «Entra, perché qui troverai un'avventura, se la desideri.»	In terms of fluency, the translation "dopo essere andato per molto tempo" makes the text less natural. Regarding accuracy, the polysemous lemma "hostal" is translated as "ostello," a lexical choice that is not appropriate in this context. In addition, the term "giardino" is omitted.	Dopo aver camminato a lungo, videro una grande luce e, poco distante, un ostello con un bellissimo portale. Si avvicinarono al portale e trovarono un portiere che subito li aprì la porta e disse: «Entrate, perché qui troverete un'avventura, se la desiderate.»	The verb is rendered in the plural instead of the singular. Furthermore, the construction "li aprì la porta" contains a grammatical error.
Mistral Nemo-12B	E dopo tanto tempo, egli vide una grande chiarezza e arrivò a un ostello con un portale molto bello. A quel portale, egli entrò e trovò un portiere che gli aprì subito la porta e disse: «Entra in questa altra sala perché qui troverai avventura, se la vorrai.»	The machine translation shows omissions of the terms "camminare" and "lontananza." Furthermore, a translation error is observed in the case of "sala," while the polysemous lemma "hostal" is translated as "hostel," a lexical choice that is not appropriate in this context.	E quando ebbero camminato per molto tempo, videro una grande luminosità e scorsero un bellissimo castello con un portale molto bello. A quel portale si avvicinarono e trovarono un portiere che subito gli aprì la porta e disse loro: «Entrate in questo castello perché qui troverete l'avventura se la volete.»	The term "camminare" is not omitted in the few-shot setting; however, the machine translation shows an incorrect shift from plural to singular.

Table 7
Hallucination of the Phi-4-14B model

Zero-shot	Few-shot
Ma Brunisend non dorme né riposa, bensì riflette su un altro motivo: di Jaufre, quale possa essere il suo amore, perché né oro né argento né ricchezza ha preso da lui neanche un denaro. Perché mai aveva visto un cavaliere che le piacesse tanto e per cui Amore l'avesse così stretta come per questo faceva. E diceva, se potesse vederlo il giorno dopo, che suo marito sarebbe stato ingannato.	Ma Brunisenda non dorme né riposa, al contrario si dedica ad altre preoccupazioni: di Gauvain considera il suo amore, perché per lui non ha mai ricevuto né oro né argento né ricchezza alcuna. Infatti, non aveva mai visto un cavaliere che le piacesse tanto o da cui Amore l'avesse legata così strettamente come Gauvain. E dice che, se potesse vederlo, sarebbe sedotto anche il suo marito.

more appropriate lexical choices. Furthermore, the qualitative evaluation provided by XCOMET-XL proved unreliable for this language pair, highlighting the need to complement automatic metrics with human qualitative evaluation. Future work will explore the extension of the dataset, the inclusion of further Large Language Models, and the comparison of additional evaluation metrics in order to better assess the capabilities and limitations of LLMs for historical language translation.

Declaration on Generative AI

The authors declare that no generative AI tools were used in the writing of this manuscript. AI-assisted technologies (such as Grammarly and ChatGPT) were solely used for grammar and spelling corrections, and as part of the translation experiments detailed in the methodology section. After using these services, the authors reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] A. Bondielli, G. Bonetta, E. Leonardelli, I. Siragusa, Preface to the Ninth Workshop on Natural Language for Artificial Intelligence (NL4AI), in: Proceedings of the Ninth Workshop on Natural Language for Artificial Intelligence (NL4AI 2026) co-located with 24th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2026), 2026.

[2] W. X. Zhao, K. Zhou, J. Li, T. Tang, Z. Dong, Y. Hou, B. Zhang, Y. Min, J. Zhang, P. Liu, X. Wang, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, Y. Hu, J.-Y. Nie, J.-R. Wen, A survey of large language models, Frontiers of Computer Science 20 (2026) 2012627. URL: https://doi.org/10.1007/s11704-026-60308-3. doi:10.1007/s11704-026-60308-3.

[3] A. Castiglione, F. Colace, M. Fusco, A. Giordano, D. Santaniello, C. Valentino, Data-driven innovation in cybersecurity for e-business information systems: A case study on cve report generation via llm, in: J. Kim, K.-L. Tsai, F. Colace, I. You (Eds.), International Symposium on E-Business Information Systems Evolution, Springer Nature Switzerland, Cham, 2027, pp. 49–62. doi:10.1007/978-3-032-26616-3_4.

[4] Y. Wang, J. Zhang, T. Shi, D. Deng, Y. Tian, T. Matsumoto, Recent advances in interactive machine translation with large language models, IEEE Access 12 (2024) 179353–179382. doi:10.1109/ACCESS.2024.3487352.

[5] A. Roncaglia, La lingua dei trovatori: profilo di grammatica storica del provenzale antico, Edizioni dell'Ateneo, Roma, 1965.

[6] M. Casillo, F. Colace, B. B. Gupta, A. Lorusso, D. Santaniello, C. Valentino, The Role of AI in Improving Interaction With Cultural Heritage, IGI Global, 2023, pp. 107–136. doi:10.4018/978-1-6684-9999-3.ch006.

[7] F. Colace, R. Gaeta, A. Lorusso, M. Pellegrino, D. Santaniello, New ai challenges for cultural heritage protection: A general overview, Journal of Cultural Heritage 75 (2025) 168–193. URL: https://www.sciencedirect.com/science/article/pii/S1296207425001517. doi:https://doi.org/10.1016/j.culher.2025.07.019.

[8] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318. doi:10.3115/1073083.1073135.

[9] S. Banerjee, A. Lavie, METEOR: An automatic metric for MT evaluation with improved correlation with human judgments, in: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Association for Computational Linguistics, Ann Arbor, Michigan, 2005, pp. 65–72. URL: https://aclanthology.org/W05-0909/.

[10] R. Rei, C. Stewart, A. C. Farinha, A. Lavie, COMET: A neural framework for MT evaluation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 2685–2702. doi:10.18653/v1/2020.emnlp-main.213.

[11] R. Pei, Y. Liu, P. Lin, F. Yvon, H. Schuetze, Understanding in-context machine translation for low-resource languages: A case study on manchu, in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vienna, Austria, 2025, pp. 8767–8788. doi:10.18653/v1/2025.acl-long.429.

[12] M. Freitag, R. Rei, N. Mathur, C.-K. Lo, C. Stewart, E. Avramidis, T. Kocmi, G. Foster, A. Lavie, A. F. T. Martins, Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust, in: Proceedings of the Seventh Conference on Machine Translation (WMT), Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 2022, pp. 46–68. doi:10.18653/v1/2022.wmt-1.2.

[13] M. Schöffel, M. Wiedner, E. G. Arias, P. Ruppert, C. Heumann, M. Aßenmacher, Modern models, medieval texts: A POS tagging study of old occitan, in: Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities, Association for Computational Linguistics, Albuquerque, USA, 2025, pp. 334–349. doi:10.18653/v1/2025.nlp4dh-1.30.

[14] M. Staffieri, GPT-3.5: la linguistica nelle mani dell'AI. esperimenti di analisi sintattica su guglielmo ix e raimbaut d'aurenga, Cognitive Philology 16 (2023) 1–24.

[15] E. D. Meo, ChatGPT tra lirica ibero e galloromanza: l'intelligenza artificiale alla prova con jaufre raudel, don denis e thibaut de champagne, Cognitive Philology 16 (2023) 1–20.

[16] S. Galano (Ed.), Blandin di Cornovaglia, Edizioni dell'Orso, Alessandria, 2004.

[17] A. de Castelnaudary, Las aventuras de monsenher Guillem de la Barra, Carocci, Roma, 2014.

[18] C. Lee (Ed.), Jaufre, Carocci, Roma, 2006.

[19] B. de Ventadorn, Canzoni, Carocci, Roma, 2003.

[20] Microsoft, et al., Phi-4 technical report, 2024. URL: https://arxiv.org/abs/2412.08905. arXiv:2412.08905.

[21] Alibaba Cloud, et al., Qwen2.5-VL technical report, 2025. URL: https://arxiv.org/abs/2502.13923. arXiv:2502.13923.

[22] OpenAI, A. Singh, et al., OpenAI GPT-5 system card, 2025. URL: https://arxiv.org/abs/2601.03267. arXiv:2601.03267.

[23] Anthropic, Claude sonnet 4.5 system card, 2025. URL: https://assets.anthropic.com/m/12f214efcc2f457a/original/Claude-Sonnet-4-5-System-Card.pdf.

[24] Mistral AI, et al., Mistral nemo 12b technical report, 2024. URL: https://arxiv.org/abs/2407.11935. arXiv:2407.11935.

[25] M. Mancini (Ed.), Flamenca, Carocci, Roma, 2006.

[26] R. Rei, J. G. C. de Souza, D. M. Alves, C. Zerva, A. C. Farinha, T. Glushkova, A. Lavie, L. Coheur, A. F. T. Martins, COMET-22: Unbabel-ist 2022 submission for the metrics shared task, in: Proceedings of the Seventh Conference on Machine Translation (WMT), 2022, pp. 578–585. doi:10.18653/v1/2022.wmt-1.52.

[27] N. M. Guerreiro, R. Rei, D. van Stigt, L. Coheur, P. Colombo, A. F. T. Martins, xCOMET: Transparent machine translation evaluation through fine-grained error detection, Transactions of the Association for Computational Linguistics 12 (2024) 979–995. doi:10.1162/tacl_a_00683.