Evaluating Large Language Models for Old Occitan–Italian Machine Translation

Mario Casillo
DIPSUM
University of Salerno
Fisciano, Italy
mcasillo@unisa.it

Sabrina Galano
DIPSUM
University of Salerno
Fisciano, Italy
sgalano@unisa.it

Annachiara Guerra
DiSPaC
University of Salerno
Fisciano, Italy
anguerra@unisa.it

Domenico Santaniello
DiSPaC
University of Salerno
Fisciano, Italy
dsantaniello@unisa.it

Giusy Strollo
DIIn
University of Salerno
Fisciano, Italy
gstrollo@unisa.it

Annapia Tanzola
DIIn
University of Salerno
Fisciano, Italy
atanzola@unisa.it

***Abstract*— Large Language Models (LLMs) have recently achieved remarkable results in the field of machine translation, particularly for high-resource languages. However, their effectiveness in the context of historical and low-resource languages remains largely underexplored. This study investigates the use of Large Language Models for the translation of Old Occitan into Italian. Specifically, it provides a comparative evaluation of five LLMs and examines the impact of two in-context learning strategies, namely zero-shot and few-shot prompting, on translation quality. The models are assessed using COMET-based automatic evaluation, complemented by a qualitative analysis of translation errors. This study contributes to understanding the capabilities of current LLMs and the role of prompting strategies for machine translation in historical low-resource settings.**

Keywords*— *Machine Translation, Large Language Models, Low- Resource Languages

I. INTRODUCTION

Machine translation has undergone significant evolution over time. From the first systems based on linguistic rules, known as Rule-Based Machine Translation (RBMT), the field progressively shifted towards corpus-based approaches, including Example-Based Machine Translation (EBMT), Statistical Machine Translation (SMT), eventually leading to Neural Machine Translation (NMT) systems [1]. The introduction of Large Language Models (LLMs) has represented a further step forward in the field of machine translation. Thanks to their ability to process vast amounts of linguistic data and interact with users, these models have opened new perspectives for improving translation quality. However, machine translation continues to pose a significant challenge, especially for low-resource languages, for which the availability of training data is insufficient. To fully exploit the potential of Large Language Models in machine translation there are several methodological approaches, including In-Context Learning (ICL), which enables models to learn from the context provided in the prompt [2]. The evaluation of the quality of machine translation is a fundamental aspect. To this end, several automatic evaluation metrics have been developed, among the most widely used is BLEU. This metric assesses translation quality by comparing the n-grams in the generated text with those in a reference translation [3]. Although BLEU has long been the standard for automatic translation evaluation, it has some limitations, as it relies mainly on lexical overlap and does not always adequately capture semantic and contextual aspects. To overcome these limitations, neural metrics have been developed in recent years, such as BLEURT, a BERT-based neural metric that aims to reproduce human assessments of the quality of generated text [4], and COMET (Crosslingual Optimized Metric for Evaluation of Translation) introduced in 2020. COMET leverages multilingual neural language models and is trained to predict scores assigned by human evaluators, thus providing an estimate of translation quality that is closer to human judgment than traditional metrics [5]. This study represents an exploratory investigation within a broader research effort aimed at examining the potential of Large Language Models for supporting the translation and accessibility of historical low-resource languages. In particular, this work focuses on the comparative evaluation of existing Large Language Models applied to the traslation of Old Occitan, a historical low-resource language. It examines the effectiveness of In-Context Learning in zero-shot and few-shot settings, comparing the performance achieved in the two scenarios through automatic evaluation based on the COMET metric and a qualitative error analysis. Within the Smart City context, digital technologies can contribute to the valorization of cultural heritage. In this perspective, Artificial Intelligence and Large Language Models can support the processing, translation, and dissemination of historical resources, helping to make them more accessible.

II. THE PROPOSED APPROACH

Machine translation based on Large Language Models has shown particularly promising results for high-resource languages; however, its application to low-resource languages continues to represent a significant challenge. Old Occitan, the medieval language that served as a vehicle for much of the literary production of its time, constitutes a particularly interesting case study in this context. The experimental setup involves the use of four works written in the langue d'oc. Three of them belong to the narrative genre: Blandin di Cornovaglia, Jaufre, and Las aventuras de monsenher Guillem de la Barra. These texts represent a relatively uncommon genre within medieval Occitan literature, which is traditionally characterized by the predominance of lyric poetry. To provide a broader representation of the language,

the corpus also includes a selection of songs by the troubadour Bernart de Ventadorn. The analysis is conducted on textual segments corresponding to translation units extracted from the selected works. The experiment involves five Large Language Models selected on the basis of the existing scientific literature concerning Old Occitan [6] and the application of Large Language Models to low-resource languages. Two prompting strategies are considered. In the zero-shot setting, the prompt contains only an instruction requesting the translation of the text from Old Occitan into Italian. In the few-shot setting, examples of translation from Occitan into Italian are provided within the prompt in order to guide the model during the translation process. This will make it possible to evaluate the impact of including translation examples on translation performance. Translation quality will be assessed using the COMET metric, specifically through two evaluation models: wmt22-comet-da [7] and XCOMET-XL. The latter, in addition to providing an overall quality score, produces a structured annotation of translation errors inspired by the MQM (Multidimensional Quality Metrics) framework. For each detected error, the model identifies the affected segment, the error category, its severity, and a confidence score ranging from 0 to 1 [8]. This automatic evaluation will be complemented by a qualitative analysis of the errors conducted on a sample of translations.

III. RESULTS AND DISCUSSION

The results of the automatic evaluation of the overall model performance according to the wmt22-comet-da metric show that GPT-5 (0.742) achieves the highest score, followed by Phi4-14B (0.708), Qwen2.5-VL-32B (0.694), Claude Sonnet 4.5 (0.667), and Mistral Nemo-12B (0.624). The evaluation performed with XCOMET-XL shows a similar trend, confirming GPT-5 (0.378) as the best-performing model, followed by Phi4-14B (0.290), Claude Sonnet 4.5 (0.280), Qwen2.5-VL-32B (0.275), and Mistral Nemo-12B (0.203).

The effect of prompting strategies was analyzed by comparing the zero-shot and few-shot settings (Table I). The results show that the impact of including translation examples varies depending on the model considered. GPT-5, Phi4-14B, and Qwen2.5-VL-32B show limited variations between the two settings according to both evaluation metrics. Conversely, Claude Sonnet 4.5 and Mistral Nemo-12B benefit the most from few-shot prompting, with improvements of +0.052 and +0.046, respectively, according to wmt22-comet-da, and +0.067 and +0.051, respectively, according to XCOMET-XL. The qualitative analysis conducted on the generated translations shows that few-shot prompting mainly improves translation fluency, particularly for models exhibiting more significant variations between the two settings. Conversely, models that already produce relatively fluent translations in the zero-shot condition, such as GPT-5, Phi4-14B, and Qwen2.5-VL-32B, show more limited improvements after the introduction of examples.

However, an important limitation concerns the applicability of automatic evaluation metrics to historical low-resource languages. Although COMET provides an estimation of translation quality based on multilingual data, it does not directly support Old Occitan as a source language, while Italian is supported as the target language. Therefore, the obtained scores should be interpreted with caution in this specific translation scenario. Moreover, although XCOMET-XL provides automatic error annotations inspired by the MQM framework, its qualitative assessments proved insufficiently reliable for Old Occitan–Italian translation. This highlights the need to integrate such tools with human evaluation in order to accurately identify and categorize translation errors in historical low-resource language contexts.

TABLE I. RESULTS

Models	**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
Qwen2.5VL-32B	Zero-shot	0.692	0.276
	Few-shot	0.697	0.273

REFERENCES

[1] H. Wang, H. Wu, Z. He, L. Huang, and K. W. Church, "Progress in Machine Translation," *Engineering*, vol. 18, pp. 143-153, 2022.

[2] Y. Wang, J. Zhang, T. Shi, D. Deng, Y. Tian, and T. Matsumoto, "Recent Advances in Interactive Machine Translation with Large Language Models," *IEEE Access*, vol. 12, pp. 179353-179382, 2024.

[3] K. Papineni, S. Roukos, T. Ward, and 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*, pp. 311-318, 2002.

[4] T. Sellam, D. Das, and A. Parikh, "BLEURT: Learning Robust Metrics for Text Generation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp.7881-7892, 2020.

[5] R. Rei, C. Stewart, A. Farinha, and A. Lavie, "COMET: A Neural Framework for MT Evaluation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 2685-2702, 2020.

[6] M. Schöffel, M. Wiedner, E. G. Arias, P. Ruppert, C. Heumann, and 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*, pp. 334–349, 2025.

[7] R. Rei, J. G. C. de Souza, D. M. Alves, C. Zerva, A. C. Farinha, T. Glushkova, A. Lavie, L. Coheur, and 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), pp.578-585, 2022.

[8] N. M. Guerreiro, R. Rei, D. van Stigt, L. Coheur, P. Colombo, and A. F. T. Martins, "xCOMET: Transparent Machine Translation Evaluation through Fine-grained Error Detection," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 979-995, 2024.