Archaic to Modern Italian Automatic Translation

Federico Siciliani, Gianmarco Corsi

1 Introduction

This project explores the application of Large Language Models for the task of translating archaic Italian sentences into modern Italian. The aim is not only to perform these translations but also to evaluate the quality of the generated modern Italian text using LLMs and manual evaluation, and finally compare them.

2 Methodology

The initial phase of this project focuses on generating modern Italian translations from archaic Italian sentences using a set of large language models (LLMs) and customized prompts. Specifically, three different LLMs will be selected, and five distinct prompts will be applied to each model to translate the archaic sentences in the dataset. The selected models have a similar number of parameters, they differ in architecture, training data, and design features. This cross-validation approach aims to explore how these characteristics influence performance on the translation task. Additionally, it helps identify which types of prompts are most effective for each model. The following phase involves evaluating the quality of the generated translations. This includes both automated and manual assessments. Two additional LLMs will be employed as evaluators, each using the same prompt to assess the translations. In addition to the automated scoring, a manual evaluation will be conducted. After that, a comparative analysis between human judgments and LLM-based evaluations will be carried out to evaluate the reliability of the LLMs judge.

3 Experimental setup

The dataset used in this task consists of about 100 sentences written in archaic Italian from the 13th century. For the translation task, three large language models (LLMs) with almost the same number of parameters each were selected, using the Ollama library. Although similar in size, these models differ significantly in architecture and training data, making them suitable for comparative evaluation:

- **LLaMA 3 (8B)**: A model developed by Meta, known for its strong performance across a variety of natural language understanding tasks and trained on a broad multilingual corpus.

- **Gemma (7B)**: A model by Google, optimized for instruction-following and conversational tasks.

- **Cerbero 7B**: A model developed with a focus on high performance in the Italian language, trained primarily on Italian texts and capable of understanding historical linguistic features.

To evaluate the influence of prompt design on translation quality, five different prompting strategies were tested with each model:

1. **Standard prompt**: A basic, direct instruction to translate the archaic sentence into modern Italian.

2. **Detailed prompt**: A more comprehensive instruction that includes contextual information and a clearer definition of the translation task.

3. **Few-shot prompt**: Several examples of archaic-to-modern translations are provided before the target sentence, guiding the model through analogy.

4. **Role-based prompt I**: The model is instructed to act as an expert in historical linguistics and translation.

5. **Role-based prompt II**: A simulated dialogue where the model plays the role of a teacher translating sentences requested by a student.

After translation, the outputs are evaluated by two separate LLMs used as scoring agents. The primary evaluation model is **M-Prometheus 3B**, a model designed for high-quality response evaluation and alignment scoring. The second one is Gemini, more specifically **gemini-1.5-flash-latest**, which was used via API calls. Each evaluator has to assign a score from 1 to 5 based on the following criteria: semantic equivalence, grammatical correctness, fluency, and preservation of the original nuance.

3.1 Experiments

During the initial trials with the proposed prompts, some models began generating outputs that included not only the translated sentence but also explanations of the translation. In some cases, parts of the output were even translated into English rather than modern Italian. To address this issue and enforce consistency, we added a clarification line to each prompt: *"The output must be ONLY the translated sentence in modern Italian."*

Following this modification, most models followed the instruction, and only a small number of outputs included extra content or English translations. However, specific issues persisted. The Gemma model, when using the detailed prompt, included explanatory text alongside the translation. Similarly, LLaMA 3 tended to produce longer outputs, which also include the translation explanation, particularly with the detailed and role-based I prompts.

4 Results

4.1 M-Prometheus (3B)

According to evaluations performed by the M-Prometheus 3B model, Cerbero 7B achieved the highest overall performance in the translation task, followed by Lama 3, with Gemma 7B showing the lowest performance among the three models.

In terms of prompting strategies, the few-shot prompt provided the best results for LLaMA 3, demonstrating the model's ability to generalize from examples. The detailed prompt and role-based prompt I also led to consistently strong translations across all models, while the basic prompt resulted in the weakest performance overall, indicating the importance of context and clear task definition.

Specifically for the Gemma model, the role-based prompt II performed particularly poorly, suggesting that this model may struggle with complex role-play instructions.

4.2 Gemini (1.5-flash-latest)

From the results, we observe that Gemini generally assigns lower scores compared to Prometheus. However, it remains consistent with Prometheus in the relative ranking of the three models: LLaMA3 receives the highest scores, while Gemma receives the lowest, according to both evaluators. A divergence with Prometheus can be seen when evaluating the combination of LLaMA3 as the model and detailed as the prompt: it receives the lowest score from Prometheus but the highest from Gemini. A similar inconsistency appears with the detailed prompt used with Cerbero, which achieves the best score according to Prometheus but the worst according to Gemini.

4.3 Manual evaluation

We manually assessed a subset of the translations. Specifically, we selected 20 samples for each combination of model and prompts. Each translation was scored on a scale from 1 to 5 based on the previous criteria. The results showed that **LLaMA 3** produced the highest-quality translations overall. It demonstrated a strong ability to convert archaic vocabulary into appropriate modern Italian equivalents. **Cerbero 7B** ranked second, delivering accurate translations that generally preserved both meaning and syntactic structure. However, in some cases, it retained archaic terms that should have been modernized. The lowest performance was observed in **Gemma 7B**, where more translations than the other models exhibited incorrect syntax or altered the intended meaning of the original sentence. Overall, the two LLMs demonstrate reliability in evaluating translations. However, the variance between their evaluations and the manual scores is quite high, indicating that they tend to assign different grades compared both to human judgments and to each other.

4.4 Variance by Group

We also calculated the variance of the mean evaluation scores, first grouped by model and then by prompt. This analysis allows us to determine whether the choice of model or prompt has a greater influence on the evaluator's scoring. Results indicate that, for both LLM evaluators, the model has a stronger impact on the assigned scores than the prompt, especially with the Gemini scores.

A Appendix

Prompt	Cerbero7B	Gemma 7B	LLaMA 3
Base	4.59	4.12	4.40
Detailed	4.77	4.87	4.18
Few-shot	4.56	4.29	4.81
Role-based I	4.75	4.18	4.69
Role-based II	4.64	3.97	4.55

Table 1: Evaluation scores from M-Prometheus.

Prompt	Cerbero7B	Gemma 7B	LLaMA 3
Base	3.70	3.37	3.84
Detailed	3.38	3.69	3.94
Few-shot	3.73	3.51	3.85
Role-based I	3.68	3.39	3.79
Role-based II	3.58	3.45	3.72

Table 2: Evaluation scores from Gemini 1.5.

Model	Prompt	Same Scores	Variance
Cerbero	base	7.00	1.92
Cerbero	detailed	6.00	2.88
Cerbero	few_shot	7.00	1.50
Cerbero	role-based	4.00	2.20
Cerbero	teacher_student	2.00	2.56
Gemma	base	6.00	1.62
Gemma	detailed	6.00	0.88
Gemma	few_shot	10.00	0.84
Gemma	role-based	6.00	1.52
Gemma	teacher_student	6.00	0.64
LLaMA 3	base	2.00	0.90
LLaMA 3	detailed	5.00	1.58
LLaMA 3	few_shot	8.00	0.98
LLaMA 3	role-based	9.00	0.47
LLaMA 3	teacher_student	2.00	1.29

Table 3: Comparison between Gemini and manual scores: number of identical ratings and mean variance.

Model	Prompt	Same Scores	Variance
Cerbero	base	6.00	1.14
Cerbero	detailed	4.00	1.69
Cerbero	few_shot	4.00	2.22
Cerbero	role-based	2.00	0.95
Cerbero	teacher_student	8.00	1.73
Gemma	base	5.00	2.57
Gemma	detailed	3.00	0.95
Gemma	few_shot	6.00	0.99
Gemma	role-based	4.00	2.02
Gemma	teacher_student	7.00	1.11
LLaMA 3	base	5.00	1.51
LLaMA 3	detailed	7.00	1.31
LLaMA 3	few_shot	8.00	0.85
LLaMA 3	role-based	8.00	1.32
LLaMA 3	teacher_student	8.00	1.73

Table 4: Comparison between Prometheus and manual scores: number of identical ratings and mean variance.

Judge	Group By	Variance
Gemini	Model	0.02008
Gemini	Prompt	0.00145
Prometheus	Model	0.03029
Prometheus	Prompt	0.02601

Table 5: Variance of mean scores grouped by model and prompt for each evaluator.