CulturalIA: Ancient to Modern Italian Automatic Translation

Kevin Giannandrea and Leonardo Mariut
Sapienza University of Rome
{giannandrea.2202212,mariut.1986191}@studenti.uniroma1.it

1 Introduction

This project addresses the task of translating ancient Italian into modern Italian using three models: Gemma 2B-it, Minerva350M-finetuned, and OpusMT. The generated translations were rated by human annotators and further evaluated using Prometheus and Gemini 2.0 Flash Lite. Finally, a comparative analysis between human and automatic evaluations was conducted using agreement metrics.

2 Methodology

In this section, we describe the datasets used by the models, the translation phase and the evaluation rubrics applied.

2.1 Dataset Exploration

The annotation dataset consists of 97 samples of sentences in ancient Italian. Initially, each sentence was manually translated to create a set of gold reference translations, which serve as a benchmark for evaluating the automated translations.

Additionally, we constructed an auxiliary dataset by extracting and splitting the entire *Divina Commedia* into individual sentences, each paired with their modern Italian translations.

2.2 Translation phase

Three different models were employed to generate translations from ancient Italian to modern Italian: Gemma 2B-it, a fine-tuned version of Minerva 350M, and OpusMT. Each model processed the same set of 97 sentences to produce corresponding translations.

Specifically, for Gemma 2B-it and Minerva350M-finetuned, we adopted a context learning approach, where a few translation examples were included in the prompt to guide the model. For OpusMT, we employed a two-step translation strategy: first translating the ancient Italian sentences into English, and then translating the resulting English into modern Italian.

2.3 Evaluation

To evaluate the quality of the generated translations, we employed two large language models as automatic judges: **Prometheus Eval 2.0 7B** and **Gemini 2.0 Flash Lite**. Each model assessed the outputs produced by the three translation systems using a rubric based on two main dimensions: *semantic fidelity* and *grammatical correctness*, with scores ranging from 1 (poor) to 5 (excellent).

In addition, for the evaluation of **Gemma 2B-it** specifically, we tested an alternative rubric focusing on *fluency* and *naturalness* of the translation, deliberately ignoring semantic alignment. This secondary rubric used a 3-point scale.

3 Experiments

We conducted our experiments on the 97-sentence ancient Italian dataset, running each translation system under comparable conditions to assess both quantitative performance and qualitative behavior.

With OpusMT, we employed a pivot-language strategy: each sentence was first translated into English using the opus-mt-it-en model, then converted back into modern Italian via opus-mt-en-it. This two-step pipeline ran on an NVIDIA GeForce 1660 Ti Mobile GPU (6 GB VRAM) with a batch size of 8. The system frequently produced literal renderings that suffered from English–Italian structural mismatches. Nouns were mis-gendered and verb forms often incorrect. We also experimented with German as the intermediate language but found the errors even more pronounced, likely reflecting greater syntactic divergence between Italian and German.

Minerva-350M showed significantly lower performance compared to the other systems. The base model was unable to understand the task and failed to produce relevant or meaningful translations. Its outputs were often nonsensical, indicating a complete lack of familiarity with the linguistic structures and vocabulary typical of ancient Tuscan Italian. To address this, the model underwent a fine-tuning process using two datasets. The first dataset was specifically built for this task and consisted of direct translations from ancient Tuscan Italian into modern Italian. These translations were manually prepared by extracting content from La Divina Commedia, focusing on some of its most significant cantos. Each canto was split into terzetti, which were then paired with their corresponding modern Italian paraphrases. This pairing allowed the model to learn the correspondence between archaic and contemporary forms of the language. To ensure quality and variety, all three canti of the poem were used, each containing both the original verses and professionally written modern paraphrases. The second dataset used for fine-tuning was LIMA. Although LIMA does not contain Italian translations, it was included to help the model generalize better and reinforce language structure, coherence, and stylistic fluency in Italian outputs. Fine-tuning used a float16-quantized checkpoint, batch size 4, for three epochs on an NVIDIA RTX 4070 Ti Super GPU (16 GB VRAM), requiring about ten minutes. Post-fine-tuning, Minerva-350M demonstrated improved lexical mappings—correctly rendering archaic expressions like "lo porco cenghiare" as "cinghiale"—but its output still lacked fluid syntax and cohesive structure. We further applied in-context learning by providing example translations in the prompt. While this yielded modest gains in formatting and consistency, semantic fidelity remained weaker than expected.

For Gemma 2B-it, we leveraged a context-learning approach: each prompt included three human-translated sentence pairs to guide generation. This model achieved the smoothest overall style but exhibited occasional mistranslations when context examples did not closely match the target sentence structure. Collectively, these experiments highlight the trade-offs between zero-shot prompting, fine-tuning, and pivot-based translation for the specialized task of ancient Italian modernisation.

In the evaluations step, first, we loaded Prometheus eval 2.0 7B in a Kaggle environment equipped with two NVIDIA T4 GPUs (32 GB total VRAM) to perform inference. Using the human "gold" translations as reference, Prometheus scored according to our 1–5 rubric based on fidelity and grammatical correctness and assigned a score to all three translation systems. Additionally, the scoring process was also conducted without a gold reference; these results are presented in Table 2.

In addition, we employed Gemini 2.0 Flash Lite under the same rubric and hardware setup. For Gemma 2b-it, we also tried an alternative rubric focusing on fluency and naturalness of the translation. Both LLM-based judges produced scores across the 97 sentences and three systems, enabling a direct comparison with human annotations and subsequent computation of agreement metrics.

4 Results

The agreement and correlation metrics between each translation model and the two LLM judges are summarized in Table 1. We also plotted score distributions for each judge and our annotations, revealing notable differences in how LLMs and humans assign ratings. Overall, agreement remains very low across the board, indicating that LLMs as judges do not reliably reproduce human-style assessments of semantic fidelity and grammatical correctness. These results lead us to reject the hypothesis that LLM-based evaluation would closely track human judgments when evaluating translations from Ancient Italian to Modern Italian. In fact, the pivot-based OpusMT system, which we expected to perform worst, shows the highest agreement with Gemini 2.0.

In Appendix 6, we compare histograms of score distributions: Our scores are relatively uniformly distributed between 2 and 5, whereas both LLM judges tend to cluster their ratings more narrowly (e.g., Gemini 2.0 rarely utilizes the extreme ends of the scale). These findings suggest that the LLM-as-a-Judge models struggle to properly score translations from ancient Italian, likely due to the niche nature and under-representation of archaic Italian in their training data. It is also important to note that the translation models employed demonstrated considerable limitations in generating ancient-to-modern Italian translations, as evidenced in our experiments. This points towards a need for future work to explore more powerful foundation models combined with more extensive domain-specific fine-tuning to elevate the overall quality of such translations.

5 Appendix A

Translation Model	**LLM Judge**	**Cohen's Kappa**	**Spearman's Correlation**
gemma2b_it_context_learning	gemini-2.0-flash-lite	0.0033	0.2463
gemma2b_it_context_learning	prometheus	-0.0470	0.0406
opusMT	gemini-2.0-flash-lite	0.0427	0.3939
opusMT	prometheus	0.0985	0.0790
minerva350M_finetuned	gemini-2.0-flash-lite	-0.0356	-0.0459
minerva350M_finetuned	prometheus	0.0148	0.1026

Table 1: Agreement and Correlation Metrics Between Human and LLM Judge Ratings

Original (Tuscan)	**Modern Italian Translation**	**Prometheus Feedback**
Non d' altra forza d' animo fue ornato Ponzio Aufidiano, romano cavaliere.	Non ha altra forza d'animo che l'onore di Ponzio Aufidiano, cavaliere romano.	[Score: 5] [The translation is a faithful and fluent rendition of the original text. The translation is not only accurate but also stylistically elegant, making it a perfect translation.]
Se questo piace a tutti e se 'l tempo hae bisogno d' ompeio per cavaliere e non per compagno, non riterrò più i fati.	Se questo piace a tutti e se il tempo ha bisogno di ompeio come cavaliere e non come compagno, non riterrò più i fati.	[Score: 4] [The translation is faithful to the original text, with no significant errors or omissions. However, there are minor stylistic mismatches, so the translation is good but not perfect.]
Officio di questa arte pare che sia dicere appostatamente per fare credere, fine è far credere per lo dire.	L'arte di questa disciplina sembra essere quella di inventare una falsa immagine per far credere che sia vera.	[Score: 5] [The translation provided is clear and accurate, capturing the essence of the original sentence; therefore, based on the criteria outlined in the score rubric, the translation is perfect.]
cose ch'io sapeva che erano fatte in Italia.	cose che io sapevo erano fatte in Italia.	[Score: 1] [The translation provided is a direct and literal translation but is not natural or fluent for an English speaker. The translation is unacceptable.]
A Milano fue ripressa la malvagità d' una donna in simile bugìa, in questo medesimo tempo di questo signore della republica, in questo modo:	A Milano fu ripresa la malvagità di una donna in simile bugìa, in questo medesimo tempo di questo signore della Repubblica, in questo modo:	[Score: 1] [End]

Table 2: Examples of Prometheus evaluations for Gemma translations, without gold references. These feedbacks generally disagree with human judge's scoring.

6 Appendix B

Gemma2b-it

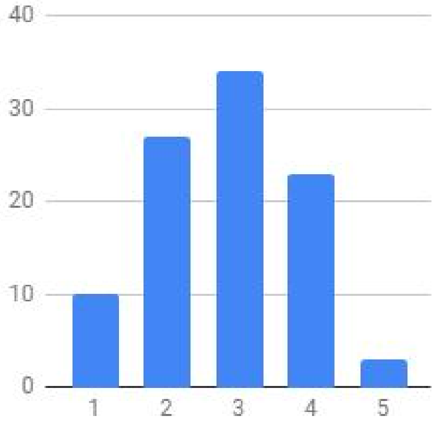

Figure 1: Histogram of score distributions for Human Annotations (Gemma2b-it).

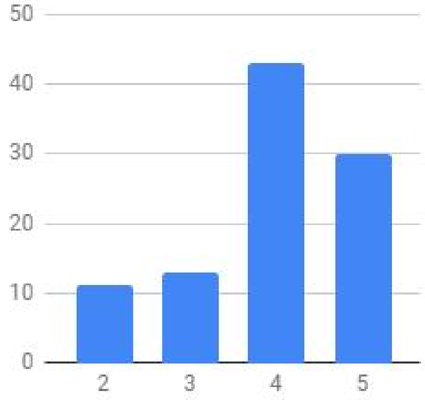

Figure 2: Histogram of score distributions for LLM Judge Gemini 2.0 Flash Lite (Gemma2b-it).

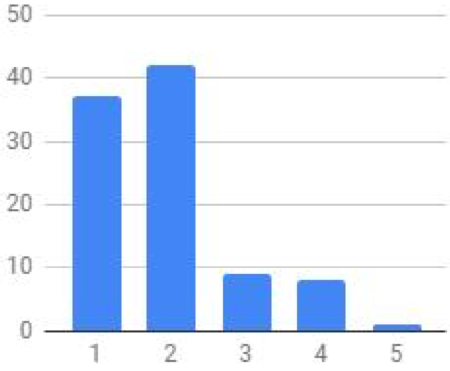

Figure 3: Histogram of score distributions for LLM Judge Prometheus Eval 7B (Gemma2b-it).

OpusMT

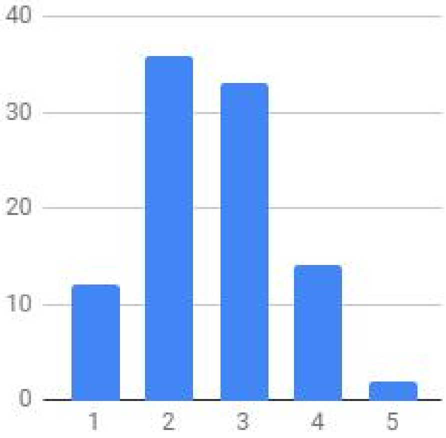

Figure 4: Histogram of score distributions for Human Annotations (OpusMT).

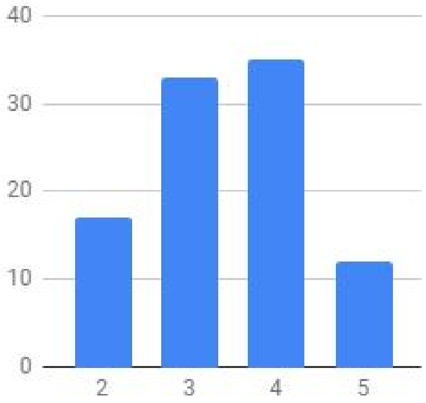

Figure 5: Histogram of score distributions for LLM Judge Gemini 2.0 Flash Lite (OpusMT).

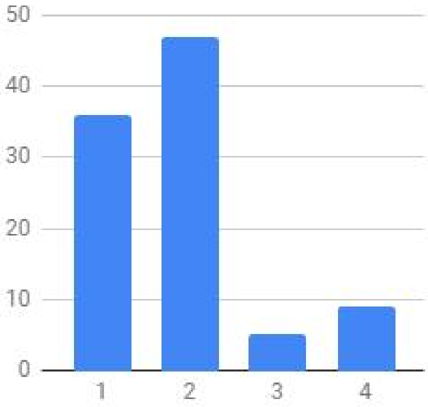

Figure 6: Histogram of score distributions for LLM Judge Prometheus Eval 7B (OpusMT).

Minerva350M-finetuned

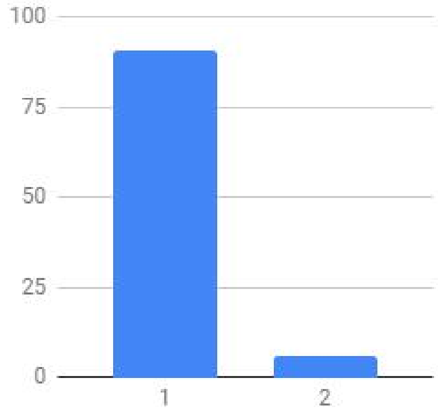

Figure 7: Histogram of score distributions for Human Annotations (Minerva350M-finetuned).

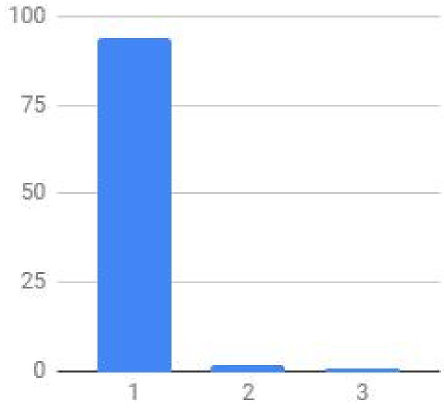

Figure 8: Histogram of score distributions for LLM Judge Gemini 2.0 Flash Lite (Minerva350M-finetuned).

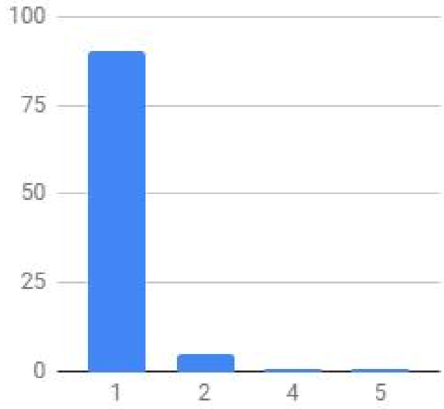

Figure 9: Histogram of score distributions for LLM Judge Prometheus Eval 7B (Minerva350M-finetuned).

References

Dante Alighieri. 1308. Inferno. http://www.letteratura-italiana.com/pdf/divina%20commedia/02%20Inferno.pdf.

Dante Alighieri. 1315. Purgatorio. http://www.letteratura-italiana.com/pdf/divina%20commedia/03%20Purgatorio.pdf.

Dante Alighieri. 1320. Paradiso. http://www.letteratura-italiana.com/pdf/divina%20commedia/04%20Paradiso.pdf.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, and 1 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33.

GAIR. 2023. Lima dataset. https://huggingface.co/datasets/GAIR/lima.

Google. 2024a. Gemini 1.5 flash 2.0 documentation. https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash.

Google. 2024b. Gemma: Open models (2b, 7b). https://huggingface.co/google/gemma.

Helsinki-NLP. 2020a. Opus-mt english to italian. https://huggingface.co/Helsinki-NLP/opus-mt-en-it.

Helsinki-NLP. 2020b. Opus-mt italian to english. https://huggingface.co/Helsinki-NLP/opus-mt-it-en.

PrometheusEval. 2024a. Prometheus-7b v2.0 model. https://huggingface.co/prometheus-eval/prometheus-7b-v2.0.

PrometheusEval. 2024b. Prometheus eval github repository. https://github.com/prometheus-eval/prometheus-eval.

SapienzaNLP. 2024. Minerva-350m-base-v1.0. https://huggingface.co/sapienzanlp/Minerva-350M-base-v1.0.