

Homework 2 - Ancient to Modern Italian Automatic Translation

A. Infantino 1922069

A. Di Chiara 1938462

F. Fragale 2169937

1 Introduction

In this work, we explore the use of large language models (LLMs) to translate sentences from archaic Italian (13th–15th century) to modern Italian. We compare different models and prompting strategies, and then evaluate the translations quality using both human-written gold references and LLM-based judges.

2 Methodology

2.1 LLMs for translation

We decided to test three multilingual instruction-tuned language models with different sizes with the aim to compare how models with different capacities perform on translating from archaic to modern Italian.

LLaMa: First we tried **LLaMa-3-8b-instruct** model that we initially tested using zero-shot prompts. However we observed that the model struggled to correctly translate the ambiguities present in the dataset. To address this issue, we tried the model with few-shot prompts, where we noticed some improvements. Finally, in an attempt to further enhance translation quality, we introduced POS tagging using SpaCy, integrating it into the few-shot prompt (see Table 1).

Minerva: We selected **Minerva-7B-Instruct**. We started using a basic few-shot prompt and then progressively introduced more detailed instructions and improved example quality. We selected three prompt configurations: one minimal, one intermediate, and one optimized. All prompts were written entirely in Italian to reflect the nature of the source texts and the native language of the model (see Table 2).

Qwen: The third model we tested was **Qwen2.5-3B-Instruct**. At first, we used a few-shot prompting approach with prompts written in English, how-

ever it tended to give mixed language answers, so we decided to use Italian prompts only.

Although this approach yielded acceptable results, in an attempt to improve performance, we tried a method inspired by the TEaR framework (Translate, Estimate, Refine) (Feng et al., 2024)¹. Specifically, we asked the model to: (1) translate the input using the same few-shot prompt; (2) evaluate the quality of its own translation; (3) use a prompt to refine the translation based solely on the feedback previously provided (see Table 3).

2.2 LLMs as Judges

To evaluate the quality of the modern translations, we used two different judges: **Gemini 2.0 Flash** and **M-Prometheus-7B**. Each model received the original sentence in archaic Italian and the translation into modern Italian. Both models followed the same rubric (see Table 4). For Prometheus, we also included a gold reference translation (made by us).

3 Experimental setup

The examples used for few-shot prompting were AI-generated and/or drawn from different authors (see Table 5). Each prompt was tailored to the specific model. All generations are performed with deterministic decoding (do_sample=False) to ensure consistent and reproducible outputs.

LLaMa: We used conversational turn prompts and the model outputs were generated via the Chat API. Each response is extracted from the assistant message, specifically from the 'content' field that contains the actual text of the reply.

Minerva: We used a role-based conversational prompt format (system, user, assistant), provided as a single input block. A true multi-turn setup

¹Our method is only inspired by the TEaR pipeline. We do not use the MQM-style prompts proposed in the original paper, but rather a custom refinement prompt.

was avoided, as it showed no quality gains and increased latency. We used the text-generation pipeline from the transformers library and outputs were extracted from the generated text using regular expressions to isolate the assistant’s final response.

Qwen: We imported Qwen from HuggingFace and we used conversational turn prompts in the text-generation pipeline. Model outputs were extracted from the content of the last `<lim_start>assistant` block.

LLMs-as-Judges We used Gemini 2.0 Flash through the google.generativeai library (API) and we extracted the model’s output by locating the line starting with `**Score**`. M-Prometheus-7B instead was imported from Hugging Face. The output was extracted using a regular expression to match a pattern like `[RESULT] <score>`.

4 Results

4.1 Qualitative Analysis

LLaMa: The model using the few-shot prompt achieved strong results in terms of fluency, tone, and the handling of ambiguous and historical terms (e.g., interpreting "franco" as "libero" rather than "French"). Preliminary attempts using POS tagging and zero-shot yielded unsatisfactory results, reinforcing the effectiveness of the few-shot approach. However, some limitations emerged with more syntactically complex sentences, where parts of the meaning were occasionally lost or misinterpreted (see Table 6).

Minerva: By comparing the previously described prompts, we obtained that the optimized prompt provided the best results (see Table 7). The model demonstrated a strong ability to preserve semantic content while modernizing lexical choices and adapting syntactic structures. Nonetheless, it occasionally misinterpreted culturally specific references or generated overly paraphrased translations.

Qwen: The TEaR-like prompt provided the best results, allowing the model to correct its own mistakes in most cases and producing overall better translations (see Table 8). However, some issues remained. For example, the model sometimes failed with cultural or historical references (e.g. 'Cecilia' not interpreted as 'Sicilia'), or added content not present in the original sentence (see Table 9).

4.2 Quantitative Analysis

For each model, we performed this analysis using the prompt that provided the best results in our qualitative analysis and selected 30 random sentences from the original dataset.

Human annotators agreement: Two human annotators rated each translation using the same 1–5 rubric given to LLMs judges, and we used Cohen’s Kappa coefficient to measure inter-annotator agreement. We obtained $\kappa = 0.907$ for LLaMA, $\kappa = 0.847$ for Qwen, and $\kappa = 0.814$ for Minerva. The results indicate a strong agreement for all the models.

Human and Judges agreement: We used Cohen’s Kappa and Mean Absolute Error (MAE). The results are shown in Table 10.

Even if Gemini shows lower MAE and higher Kappa values, it exhibited a tendency to overestimate translation quality, often assigning high scores even to inaccurate or mistranslated outputs (see Fig. 1 and Tab. 11). Prometheus seems to be more conservative in its evaluation and it follows human scoring patterns more closely. It avoid to overestimate translations, which suggests that it better replicates how humans actually evaluate (See Fig. 2 and Tab. 12).

Score mean and performance evaluation: To better understand how each model performed, we computed the average human and Prometheus ratings for all 30 sentences. We also counted how many sentences were rated between 4 and 5 by both human annotators and Prometheus. The results are summarized in Table 13.

5 Conclusion

Overall, all the evaluated LLMs performed well in translating from archaic to modern Italian, generally producing coherent outputs.

LLaMA achieved the highest average Prometheus and human score, indicating stronger alignment with human preferences. Despite Minerva showing good performance, it tended to over-paraphrase the sentences, obtaining lower scores, while Qwen showed the weakest performance in both metrics. Our analysis also shows that Prometheus is a more reliable automatic judge, and avoids overestimating translation quality unlike Gemini.

References

Zhaopeng Feng, Yan Zhang, Hao Li, Bei Wu, Jiayu Liao, Wenqiang Liu, Jun Lang, Yang Feng, Jian Wu, and Zuozhu Liu. 2024. [TEaR: Improving LLM-based Machine Translation with Systematic Self-Refinement](#). *Preprint*, arXiv:2402.16379.

6 Appendices

LLaMA Prompt	Content
Zero shot	<p>You are a professional linguist specializing in philology and the translation of archaic Italian (Florentine, 1300–1500) into modern Italian. Your task is to translate sentences written in archaic Italian into contemporary, fluent Italian. Your translations must be:</p> <ul style="list-style-type: none"> - Faithful to the original meaning. - Written in a natural and idiomatic modern style, - Clear and immediately comprehensible to a modern reader avoid archaic or overly literal structures. - You may restructure the syntax and update the vocabulary as needed to improve readability, as long as the core meaning and tone remain intact. - Provide only the translated sentence. Do not include explanations or commentary. <p>Input: "content": f'Input: sentence'</p>
Few shot	<p>You are a professional linguist specializing in philology and the translation of archaic Italian (Florentine, 1300–1500) into modern Italian. Your task is to translate sentences written in archaic Italian into contemporary, fluent Italian. Your translations must be:</p> <ul style="list-style-type: none"> - Faithful to the original meaning. - Written in a natural and idiomatic modern style, - Clear and immediately comprehensible to a modern reader avoid archaic or overly literal structures. - You may restructure the syntax and update the vocabulary as needed to improve readability, as long as the core meaning and tone remain intact. - Provide only the translated sentence. Do not include explanations or commentary. <p>Few-shot examples: 3 sentences with optimal responses.</p> <p>Input: "content": f'Input: sentence'</p>
Few shot + Pos Tagging	<p>You are an expert Italian translator and philologist. Your task is to translate literary and poetic texts from archaic Italian (Florentine, 1300–1500) to fluent Modern Italian.</p> <ul style="list-style-type: none"> - Preserve the full meaning of the original without omissions or unjustified additions. - Use correct and elegant grammar (verb tenses, agreement, punctuation) - Adapt archaic or obsolete words and expressions into natural, idiomatic modern equivalents. - Render metaphors, poetic turns, and complex structures with clarity and sensitivity. - Maintain the tone, emotional register, and stylistic elegance of the original. <p>Avoid literal translations that sound unnatural or archaic in contemporary Italian.</p> <ul style="list-style-type: none"> - Provide only the translated sentence. Do not include explanations or commentary. - Do not invent or transliterate unknown or archaic words: interpret their meaning from context and render with modern equivalents. <p>Few shot+pos tagging examples: 3 sentences with optimal responses and pos tagging.</p> <p>Input: "content": f"Translate: {sentence} POS tagging: {pos_tags}"</p>

Table 1: LLaMA prompt structures: comparison between zero-shot, few-shot, and few-shot with POS tagging prompts, aimed at improving translation quality from archaic to modern Italian through added context and linguistic cues

Minerva Prompt	Content
Basic Prompt (low detail, 2 examples)	<p>Sei un esperto linguista umano, specializzato in filologia e traduzione dall’italiano antico (Fiorentino 1300–1400) all’italiano moderno.</p> <ul style="list-style-type: none"> - Traduci solo la frase, senza spiegazioni o commenti. - Mantieni il significato originale adattando la struttura all’italiano moderno. <p>Few-shot examples:</p> <ol style="list-style-type: none"> 1. “Così che il core ne tremava tutto.” → “Il mio cuore tremava completamente.” 2. “Li occhi miei, ché fermi stettero alquanto, pur pur mi si chiuser per paura.” → “I miei occhi, che rimasero fissi per un istante, alla fine si chiusero per la paura.” <p>Input: Utente: <frase da tradurre> Assistente:</p>
Intermediate Prompt (high detail, 2 examples)	<p>Sistema: Sei un linguista professionista specializzato in filologia e nella traduzione dall’italiano arcaico (fiorentino – XIV–XV secolo) all’italiano moderno.</p> <ul style="list-style-type: none"> - Il risultato deve essere fedele al significato originale, naturale e scorrevole, chiaro e comprensibile. - Elimina ambiguità o elementi stilistici obsoleti. - Se necessario, cambia la struttura sintattica mantenendo significato e tono. <p>Few-shot examples:</p> <ol style="list-style-type: none"> 1. “E però che la seconda suole talora per la grande provedenza fare timoroso...” → “Poiché la seconda, per la sua grande previdenza, può talora incutere timore...” 2. “Egli, vedendo ciò, tacque, e con volto pietoso si trasse indietro.” → “Egli, vedendo ciò, tacque e con volto pietoso si ritrasse.” <p>Input: Utente: <frase da tradurre> Assistente:</p>
Optimized Prompt (high detail , 4 examples)	<p>Sistema: Sei un linguista professionista specializzato in filologia e nella traduzione dall’italiano arcaico (fiorentino – XIV–XV secolo) all’italiano moderno.</p> <ul style="list-style-type: none"> - Il risultato deve essere fedele al significato originale, naturale e scorrevole, chiaro e comprensibile. - Elimina ambiguità o elementi stilistici obsoleti. - Adatta sintassi, lessico e stile a un lettore moderno. <p>Few-shot examples:</p> <ol style="list-style-type: none"> 1. Consecutive clause + postponed subject. 2. Causal clause + archaic adjectives. 3. Impersonal main clause + comparative adverbial. 4. Impersonal past tense + dependent clause. <p>Input: Utente: <frase da tradurre> Assistente:</p>

Table 2: Minerva-7B prompt configurations used in the experiments. Prompts differ in the number of examples and the amount of stylistic and structural guidance provided.

Qwen Prompt	Content
Translation (few-shot)	<p>Sei un traduttore e filologo italiano. Traduci dall'italiano antico (Fiorentino, 1200–1400) all'italiano moderno.</p> <ul style="list-style-type: none"> - Il significato della frase originale deve essere rispettato senza omissioni o aggiunte improprie. - La frase deve essere grammaticalmente corretta (tempi verbali, punteggiatura, accordi morfologici come genere, numero e concordanza) - Il lessico e la struttura sintattica devono essere naturali e adatti a un lettore contemporaneo italiano. - I termini arcaici e non più in uso devono essere sostituiti con termini moderni appropriati - Il tono e lo stile devono essere moderni. - I riferimenti storici, culturali o semantici devono essere interpretati correttamente e adattati al contesto. <p>Few-shot examples: 4 sentences with optimal responses.</p> <p>Input: Traduci: <frase da tradurre></p>
Evaluation	<p>Sei un esperto di filologia e traduzione italiana. Fornisci una valutazione critica sulla qualità di una traduzione dall'italiano arcaico (Fiorentino, 1200-1400) all'italiano moderno. Presta attenzione ai seguenti aspetti per la valutazione:</p> <ul style="list-style-type: none"> - Il significato della frase originale è stato rispettato? (niente omissioni o aggiunte improprie) - La frase è grammaticalmente corretta (tempi verbali, punteggiatura, accordi morfologici come genere, numero e concordanza)? - Il lessico e la struttura sintattica sono naturali e adatti a un lettore contemporaneo italiano? - I termini arcaici e non più in uso sono stati sostituiti con termini moderni appropriati? - Il tono e lo stile sono moderni, scorrevoli e simili a un testo scritto oggi? - Sono stati evitati costrutti rigidi, impersonali o troppo fedeli all'originale antico? - I riferimenti storici, culturali o semantici sono stati interpretati correttamente e adattati al contesto? <p>Input: Original sentence + First translation</p> <p>Output: Feedback</p>
Refinement	<p>Sei un linguista e filologo italiano esperto in traduzione dall'italiano arcaico all'italiano moderno. Dati una frase in italiano arcaico (Fiorentino, 1200-1400), la sua traduzione in italiano moderno e una valutazione della traduzione, migliora la traduzione sulla base della valutazione data.</p> <p>Input: First translation + Feedback</p> <p>Output: Fornisci SOLO la versione corretta della traduzione basandoti sulla valutazione data senza fornire altre spiegazioni.</p>

Table 3: Qwen prompt structure. The few-shot prompt was re-used in the TEaR-like prompt. The model responded better to imperative forms (you must, you have to...) rather than negative forms (don't do this, don't include...). The original sentence was not included in the refinement prompt as it helped the model to better refine the initial translation.

Score	Description
1	The translation is incorrect or meaningless.
2	The translation is very inaccurate or misleading. It is too literal and fails to preserve the original meaning.
3	Understandable translation, but with significant issues in vocabulary, structure, or historical interpretation. The main idea is present but distorted.
4	Good translation that preserves the meaning but contains awkward phrasing or slightly outdated structures for modern Italian.
5	Excellent translation that is fluent, accurate, and faithful to both the meaning and tone of the original. Uses natural modern Italian.

Table 4: Scoring rubric used by both Gemini and M-Prometheus for evaluating translation quality.

Sentences	Ambiguities	Author
E però che la seconda suole talora per la grande provedenza fare timoroso, e la prima per l'ardire rendere altrui matto, assai utile è la presente distinzione.	Complex syntactic structure; archaic terms like "providenza"; dislocated sentence components.	AI-generated
Niuna cosa si dimostra onesta che con suo contrario non si disonesti.	Semantic ambiguity in the word "onesta"; depends on historical and moral context.	Francesco Petrarca
Era una giovane di maravigliosa bellezza e di costumi onesti, la quale con leggiadria e onestà si portava.	Elevated and moralizing tone; "onesta" implies modesty and social decorum.	Francesco Petrarca
Diffinizione d'una cosa è dicere ciò che quella cosa è, per tali parole che non si convegano ad un'altra cosa.	General statement with archaic vocabulary ("dicere", "convegano").	Brunetto Latini
Vergine bella, che di sol vestita, coronata di stelle, al sommo Sole piacesti sì, che n te sua luce ascose; amor mi spinge a dir di te parole.	The phrase is ambiguous, blending divine and poetic meanings in 'Sole,' 'luce,' and 'amore'.	Francesco Petrarca
Niuno savio uomo così leggermente si lascia trasportare dalla ventura.	Causal clause and archaic adjective.	Dante Alighieri
E però che la moltitudine non sapea discernere il vero dal falso, si lasciò menare per inganno.	Consecutive clause and postponed subject.	Giovanni Boccaccio
Parve a molti che colui avesse parlato con arte più che con verità.	Main impersonal clause and comparative adverbial phrase.	Giovanni Boccaccio
Coloro che il senno lor non adoperano a buono fine, degni sono di misera sorte.	Impersonal past tense and dependent clause.	Giovanni Boccaccio

Table 5: List of examples used in our prompts with corresponding linguistic ambiguities and authors.

Original Sentence	Few-shot	Few shot +POS tagging	Zero Shot	Note
Quella guerra ben fatta l'opera perché etc. Et dall'altra parte Aiace era uno cavaliere franco e prode all'arme, di gran guisa, ma non era pieno di grande senno.	Quella guerra era stata ben condotta ecc. Dall'altra parte, Aiace era un cavaliere libero e valoroso in battaglia, ma non era dotato di grande saggezza.	Quell'opera era stata correttamente condotta, come si diceva... E dall'altro lato, Aiace era un cavalier francese e prode in armi, di gran lunga, ma non era pieno di grande saggezza	Quella guerra era stata fatta con successo, perciò eccetera. Dall'altra parte, Aiace era un cavaliere francese e valoroso con le armi, enormemente, ma non era pieno di grande intelligenza.	The few-shot prompt translates the word 'franco' in the correct way.
Tarentini, i quali erano nati di quegli di Lacedemonia et facta da lloro nobile cittade de' Greci.	I Tarentini, che erano originari di Sparta e figli della nobile città dei Greci fondata da loro.	Tarentini, i quali erano nati di quegli di Lacedemonia et facta da lloro nobile cittade de' Greci.	Greci di Taranto, che erano nati dai Lacedemoni e provenivano dalla illustre città dei Greci	The few-shot prompt has historical context.
Quando li serpenti invellenava di giorno alcuno Romano, allora iera la maraviglia a vedere come li Psille si combattevano al veleno, ché elli imolavano tutto inazzi della loro salive.	Quando i serpenti attaccavano un romano di giorno, era una meraviglia vedere come i pitoni si difendevano dal veleno, poiché li uccidevano con il loro morso.	Quando li serpenti invellenava di giorno alcuno Romano, allora iera la maraviglia a vedere come li Psille si combattevano al veleno, ché elli imolavano tutto inazzi della loro salive	When the snakes were aware that a Roman was nearby during the day, then it was a wonder to see how the Psille were fighting in the venom, because they were sacrificing everything for the sake of their saliva.	All prompts struggle with complex sentences, and in particular, the model with zero-shot prompting translates the sentence into English

Table 6: Qualitative comparison of selected translations generated using few-shot, POS tagging, and zero-shot prompting. This section highlights examples where the few-shot approach outperforms the others, as well as a case where all models produce an incorrect translation. Notably, when using the few-shot prompt combined with POS tagging, the model frequently failed to translate certain words, often leaving them unchanged in the output.

Original Sentence	Detailed prompt with optimized examples	Prompt not detailed	Prompt detailed with two examples	Note
Però che or chi spererebbe quello che eziandio questi che non vogliono ancora credere in Cristo, già veggiono con noi, e perché nol possono negare, stridono colli denti.	Tuttavia coloro che non credono ancora in Cristo, pur vedendolo con i propri occhi, lo negano con i denti perché non possono ammetterlo.	Però chi spera anche quelli che non credono ancora in Cristo, già vedono con noi e perché non possono negarlo, stridono coi denti.	Tuttavia, coloro che non credono ancora in Cristo, già vedono con noi e perché non possono negarlo, digrignano i denti.	The prompt with optimized examples is most semantically coherent and stylistically natural.
Non lo volle conoscere per nimico. Qesta è quella, la quale diede ardire al profeta Natan a riprendere con grande autoritate quello re, il quale avea peccato.	Non volle conoscerlo come nemico. Questa è la stessa persona che diede coraggio al profeta Natan nel rimproverare severamente quel re, colpevole di aver commesso un grave peccato.	Questa è quella, la quale diede coraggio al profeta Natan ad affrontare con grande autorità il re, il quale aveva peccato.	Questa è quella, la quale diede coraggio al profeta Natan ad affrontare con grande autorità il re, il quale aveva peccato.	Only the first prompt (detailed and with optimized examples) includes a translation of the initial clause.
Quando li serpenti invelenava di giorno alcuno Romano, allora iera la meraviglia a vedere come li Psille si combattevano al veleno, ché elli imolavano tutto inazzi della loro salive	Quando i Romani erano assaliti dai serpenti durante il giorno, era meraviglia a vedere come le api si difendevano dal veleno, poiché esse immergevano nel proprio pungiglione tutta la saliva dei rettili.	Quando i serpenti avvelenati invadevano le strade romane, era meraviglia a vedere come gli insetti si affrontavano tra loro, poiché essi iniettavano veleno ovunque, uccidendo tutti coloro che incontravano.	Quando i serpenti avvelenavano alcuni Romani durante il giorno, era meraviglia a vedere come le pulci si combattevano tra loro, poiché esse imollavano tutto il veleno del loro sangue.	None of the three prompts delivers a correct or faithful translation of the original sentence.
Sappiate veramente, che se noi non ci fossimo sì affrettati del fuggire, noi saremmo tutti morti.	Sapete bene che se avessimo agito con maggiore rapidità, avremmo potuto salvarci tutti.	Sappiate, che se noi non avessimo agito così rapidamente, saremmo tutti morti.	Sappiate, che se noi non avessimo agito così rapidamente, saremmo tutti morti.	The optimized prompt introduces a conditional interpretation not present in the original, while the two simpler prompts are more accurate and faithful to the text.

Table 7: Comparative examples of Minerva-7B translations under different prompting strategies. While optimized prompts yield more accurate and stylistically coherent outputs, certain domain-specific or culturally anchored sentences remain challenging across all configurations.

Original sentence	Few-shot translation (conv-turn)	Correction	Note
Crudele, [...] a neuno cavaliere perdoni	[...] e perdonerà a nessun cavaliere	[...] non perdonerà a nessun cavaliere	Added a double negation
Io spero in messer Iesù di mandare [...]	Spero che messer Iesù mi invii subito Timoteo, così da potermi sentire in pace.	Prego Gesù di inviare a te Timoteo, così da potermi sentire in pace.	Subject-object inversion and idiomatic interpretation
Acciocché quegli, il quale ora per le [...]	Perché quell’persona, il cui coraggio [...]	Perché quella persona, il cui coraggio [...]	Grammatical correction
In vano si domanda chi ha scritto questo libro [...]	[...] l’autore fosse l’Esprit Santo.	[...] l’autore fosse il Spirito Santo.	Language correction

Table 8: Qualitative comparison of selected translations generated using few-shot prompting and after the refinement. Here we can see some cases when the TEaR-like prompt improved the performance.

Original Sentence	Few-shot translation (conv-turn)	Correction	Note
Marco Cornelio [...] si riservò di parlare all’ultimo.	[...] si riservò di parlare all’ultimo.	[...] si tenne per sé di parlare all’ultimo.	Improper adding
[...] e perché nol possono negare, stridono colli denti	[...] strillano con le loro labbra	[...] stridono colli denti con le loro labbra	Partial correction
Se questo piace a tutti [...] non riterrò più i fati.	[...] non riterrò più i fatti.	[...] non considererò più le destinazioni.	Improper semantic interpretation

Table 9: Qualitative comparison of selected translations generated using few-shot prompting and after the refinement. Here we can see some cases where the TEaR-like prompt worsened the performance. In general, it added additional informations, but it still preserved the original meaning of the sentence.

Model	Gemini κ	Gemini MAE	Prometheus κ	Prometheus MAE
LLaMa	0.330	0.700	0.265	0.767
Minerva	0.264	0.667	0.396	0.700
Qwen	0.288	0.667	0.255	0.500

Table 10: Automatic evaluation scores (κ and MAE) for Gemini and Prometheus across LLMs.

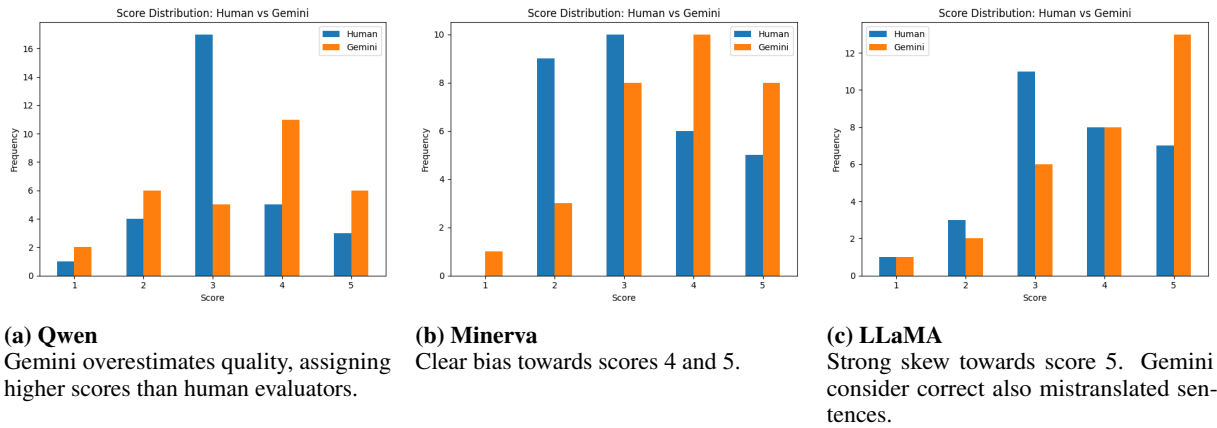


Figure 1: Score distribution: human annotators vs Gemini.

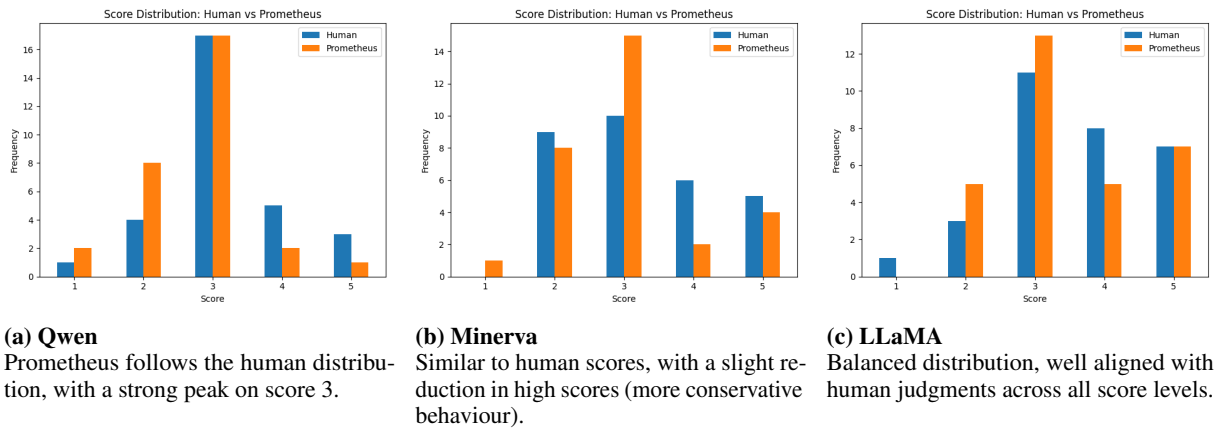


Figure 2: Score distribution: human annotators vs Prometheus.

Model	Original Sentence	Translation	Gemini score	Human score
LLaMA	noi iscaciati e dipartiti per debito dela cittade, e tutti iscaciati da fama e da ventura buona.	Noi ci siamo allontanati e partiti dalla città a causa del dovere, e tutti ci siamo allontanati anche per la fama e la buona sorte	4	2
Minerva	uno luogo si mandano lancioni; la quale cosa i cavalieri l' appellano capo di porco	Un luogo viene chiamato "capo di porco" dai cavalieri.	4	2
Qwen	Dio, per cui tutte le cose sono dispepe e giudicate.	Dio, per cui tutte le cose sono dispensate e giudicate.	5	3

Table 11: Some cases where Gemini gave high scores to mistranslated sentences.

Model	Original Sentence	Translation	Prometheus score	Human score
LLaMA	ne salì in su l'argine del fosso, e in su lo steccato, se da alto si potessero difendere, o per alcuna maniera passare oltre e scampare.	Salì sull'argine del fossato e sul muro, e cercò di difendersi o di passare oltre, nella speranza di evitare l'agguato.	3	4
Minerva	udire di belle novelle et odorare di belli fiori, come et quando et quanto si conviene	ascoltare buone notizie e annusare bei fiori, come e quando e quanto si addice.	3	5
Qwen	Quindi il dolore deriva quando l'uso dell'arma e la fatica vengono rifiutati, con un grande disonore di esser trattati come pecore.	È doloroso da sentire quando gli uomini rifiutano di combattere e di faticare, e finiscono con l'essere come pecore, nel massimo disonore	2	3

Table 12: Some cases where Prometheus it's too conservative in its evaluations.

Model	Human Avg.	4–5 Ratings (Human)	Prometheus Avg.	4–5 Ratings (Prometheus)
LLaMA	3.57	14	3.47	12
Minerva	3.27	10	3.00	6
Qwen	3.20	8	2.73	3

Table 13: Average scores and number of high-rated translations (4–5) for each model. LLaMa outperform the other models achieving higher scores overall.