Report Homework2 NLP2025

Francesco Casacchia 1698281 **Luca Franzin 1886634**

1 Introduction

In this report, we performed a comparative analysis of the output from three LLMs instructed to translate sentences from Archaic Italian to modern Italian. Additionally, to evaluate these models' performance, an additional LLM was utilized following the **'LLM-as-a-judge**' paradigm. This judge model evaluated the translations of the three models by comparing them against the golden label (our correct translation), in accordance with a specific Rubric's score system. Finally, we analyzed a selected number of translations, providing our own scores, computing various coefficients, and highlighting the strengths and weaknesses of all models.

2 Methodology

2.1 The Models

Regarding the models we used the follwing LLMs:

1. **DeepSeek-R1-0528-Qwen3-8B-bnb-4bit** this is a **reasoning model** made by distilling the famous DeepSeek-R1-0528 with Qwen3-8B. This particula version is quantized at 4bit.

2. **Mistral-7B-Instruct-v0.2**, this is a dense model famous for being quick and easy to deploy.

3. **Qwen3-32B**, this is our **biggest reasoning model**. This model is very well known for its high level reasoning capabilities and accuracy. Our version of this model has been quantized at 4bit for better performances.

4. **M-Prometheus-7B** this is a "LLM-as-a-judge" model specifically designed for comparisons. This is a very common model used for similar judging-based tasks.

All models were use locally and were downloaded from HuggingFace according to their respective open licenses.

2.2 Obtaining the Translations

In order to use the models to translate from Arcaic italian to modern italian we used the following prompt's structure: for the **user prompt** we simply used the sentence to translate without any additional commands, for the **system prompt** we used the same prompt for all three models.

- "*Sei un traduttore dall'italiano antico all'italiano moderno. Traduci una frase in italiano moderno e rispondi solo con: La traduzione è:<traduzione>. Non aggiungere altro pena la morte. Usa solo l'italiano e nessun'altra lingua nella traduzione.*"

Our findings indicate that using Italian rather than English prompts significantly reduced the likelihood of models erroneously translating sentences into English. Moreover, the specific prompt string: **"La traduzione è:<traduzione>"** was used to force the models to generate the translation after this designated text. It was also necessary to refine the output to remove additional notes or explanations that models often provided despite system instructions, especially concerning the Mistral model. Interestingly, the Mistral model demonstrated **slightly more compliance** after the inclusion of the sentence "pena la morte," effectively threatening the model with severe penalty. Nevertheless, no similar improvements were noted in the other models.

2.3 Obtaining the scores

In order to make Prometheus give us a score of "how good" the translation is, we first translate ourself all the sentences in the database, building this way a **GoldenLabel** dataset.

Then we use Prometheus with Rubric system that forces the model to focus on the quality of the translations yielding a score:

- ***Criteria**:"Rubric: Archaic to Modern Italian Translation Quality"*
- ***Score 1**:"The translation fails to convey the core meaning or introduces incorrect information. It's fundamentally broken."*
- ***Score 2**:The translation contains significant grammatical or lexical errors or introduce new words that leads to a distorted or inaccurate understanding of the original meaning."*
- ***Score 3**:"The translation preserves the original meaning accurately, but its presentation is flawed. It is difficult to read due to unnatural phrasing, incorrect modern syntax, or other stylistic errors."*
- ***Score 4**:"The translation is accurate, grammatically correct, and almost entirely fluent. It faithfully preserves the original meaning."*
- ***Score 5**:"The translation is grammatically perfect, accurate, and reads as completely natural, fluent modern Italian. It effectively modernizes all archaic elements and skillfully captures the tone and nuances of the original text."*

We experimented with **various Rubrik's scores**, ranging from the most simplistic (e.g., a categorical 'very bad' to 'bad' system) to highly complex frameworks. Our findings indicated that a rubric offering a simple yet comprehensive description yielded the most optimal results.

3 Experiments

To ensure compliance with system prompts, we deployed a range of prompting strategies across the models. Our observations indicate: **Deepseek**, by virtue of being a reasoning model, consistently followed directives without additional information. **Qwen3** similarly exhibited robust complience to system instructions. In contrast, **Mistral** frequently appended additional notes or information to its translations, often in explicit violation of the system prompt. To circumvent this, only the translation content was extracted from Mistral's output, omitting any supplementary sentences. A similar challenge was encountered with Premetheus' output. However, its 7B iteration demonstrated notably higher compliance, simplifying the extraction of scores. Moreover, we evaluated a smaller model as well: the Prometheus-3B model. However, we immediatly **noticed severely degraded performance** regarding compliance, precision of translation descriptions, and compliance to rubric-based scoring.

4 Results

Table 1: Average of the Judge for all dataset and average of our scores using the 20 selected sentences

Model	Judge's scores	Our scores
Mistral	3.08	2.2
Deepseek	3.02	2.7
Qwen	3.0	3.55

Table 2: The Cohen's Kappa Coefficient, Spearman's correlation, MSE on our scores with scores of the Judge

Model	Cohen's	Spreamn's	MSE
Mistral	0.0058	0.2134	1.5500
Deepseek	0.2691	0.1781	1.0500
Qwen	-0.0332	-0.1307	2.1500

Our translation scoring **diverges significantly** from Premetheus's results. On average, our scores for Qwen are significantly higher, for Mistral significantly lower, and for Deepseek slightly higher. While Premetheus occasionally assigns appropriately low scores to Mistral translations, it rarely rates Qwen's translations highly. Qwen notably exhibits superior semantic and syntactic comprehension, a likely benefit of its number of parameters and reasoning capabilities. Deepseek generally demonstrates a sound understanding of sentence structure, though it frequently struggles with precise terminology and idiomatic expressions. In contrast, Mistral frequently fails to accurately interpret sentence structure, leading to almost literal and contextually inaccurate translations, which consistently **yield the lowest quality**. Nevertheless, Premetheus frequently assigned Mistral favorable scores, despite the latter's erroneous translations. We attribute this to Premetheus's insufficient comprehension of Archaic Italian sentence structures. This is further sustained by its tendency outputs low to average results to good Qwen's traslations that adapt the original sentence's structure to make it more sound in modern italian.