

Archaic Italian into Modern Italian

Alessio Borgi

Sapienza University of Rome

borgi.1952442@studenti.uniroma1.it

Abstract

Machine translation from archaic to modern Italian presents unique challenges due to significant differences in vocabulary, grammar, and style between historical and contemporary language variants. Recent advances in Natural Language Processing (NLP), particularly with the advent of first transformer-based architectures and most recently using large language models (LLMs), have enabled the development of automatic systems capable of handling these complex translation tasks. In this work, we investigate the effectiveness of several open-source models for translating sentences from non-modern to modern Italian, and we evaluate their outputs using an LLM-as-a-Judge paradigm, under different judging schemes, focusing on both translation quality and human-LLM agreement. We also perform a gold-label VS LLM-as-a-judge correlation study to assess the effectiveness of the judging techniques employed. The code is available at ¹.

1 Automatic Machine Translation Models

For the task of automatic translation from archaic to modern Italian, we experimented with a total of four models, encompassing both *transformer-based* (2 both from Meta AI) and *(LLM)-based* (one from Meta and one from Google) approaches.

- **mBART-large-50-many-to-many-mmt**(Liu et al., 2020): A multilingual sequence-to-sequence model pre-trained pretrained via Denoising auto-encoding on monolingual corpora in 50 languages (mBART-50), consisting of 12-layer encoder + 12-layer decoder Transformer (≈ 610 M parameters).
- **NLLB (No Language Left Behind)-200-3.3B**(Team et al., 2022): A model specifically designed to cover a wide range of low-resource languages, including historical and

regional variants, making it well-suited for the archaic-to-modern Italian translation task.

- **LLAMA-2-7b-chat-hf**(Touvron et al., 2023): An open-access LLM with strong multilingual capabilities and adaptable instruction-following abilities from Meta.
- **Gemma-2b-it**(Team et al., 2024): A recent LLM architecture optimized for generative and translation tasks from Google.

1.1 LLMs Prompting Techniques

We explored four prompting techniques:

1. **Zero-shot translation**: Directly prompting the model to translate without any example.
2. **Few-shot translation**(Brown et al., 2020): Providing the model with several example input-output translation pairs (*In-Context Learning*).
3. **Chain-of-thought (CoT)**(Wei et al., 2023): In addition to the Few-shot, with this technique, I am encouraging the model to reason through intermediate steps during the translation process explicitly.
4. **ReAct**(Yao et al., 2023): In addition to the Few-shot, with this technique, I am combining reasoning and acting by prompting the model to iteratively generate rationales and partial outputs to improve translation quality.

2 LLM-as-a-Judge Evaluation

To assess the quality of the archaic-to-modern Italian translations produced by our models, we adopted the *LLM-as-a-Judge* (Zheng et al., 2023) paradigm, which leverages LLMs as automatic evaluators, prompting them to score translation outputs according to detailed rubrics. We explored three distinct evaluation methodologies:

¹<https://github.com/alessioborgi/AMT-AutomaticMachineTranslation>

Single-Criteria, *Multi-Criteria* and *Debate-and-Consensus* Evaluation. In all three cases, we make use of Gemini-2.0-Flash API (Team et al., 2025) as a Judge.

2.1 Single-Criteria Evaluation (General)

Here, Gemini is used to assign a *single quality score*, ranging from 1 (worst) to 5 (best), for each translation. The evaluation rubric, presented directly in the model prompt, can be seen in Appendix A.1, and covers the full spectrum of translation quality from unacceptable to perfect, providing a general, holistic judgment of translation quality.

2.2 Multi-Criteria Evaluation

Here, we extended the evaluation to a multi-criteria setting, by having each translation to be scored independently across four critical dimensions: **Adequacy**, **Fluency**, **Style**, and **Completeness**. Each criterion is rated from 1 to 5 according to a dedicated rubric. The detailed prompt can be seen in Appendix A.2. This methodology allows for a more granular analysis of translation strengths and weaknesses.

2.3 Debate-and-Consensus: Reference-Free Dual-Judge Evaluation

To further enhance reliability and robustness, we adopted a reference-free, self-improving framework based on the *debate and consensus* paradigm. Each translation is initially evaluated by two strong local LLMs, Phi3.5 (Abdin et al., 2024) from Microsoft and OpenELM (Mehta et al., 2024) from Apple, each scoring the four criteria above. Their independent scores are then submitted, together with the original sentence and translation, to Gemini. Gemini acts as a moderator, debating the merits of each judge’s scores and outputting a consensus set of scores. This process simulates an expert panel, aiming to reduce the single-model bias. The prompt used for debating can be found in Appendix A.3.

3 Experiments & Results

To quantitatively assess the reliability of LLM-based evaluation, we manually annotated a set of *gold labels* for the first 30 translations produced by each model and prompting strategy, following the same scoring rubrics used in the LLM-as-a-Judge paradigm. We performed this analysis both for the general (single-criteria) setting and for the multi-criteria one.

The *Evaluation Metrics* used are: *Cohen’s Kappa*, *Pearson Correlation*, *Spearman’s* and *Kendall’s Rank Correlations*, *Exact Match*, *Confusion Matrix*. Further info can be found in Appendix C.

Fig. 1 and Fig.2 (see Appendix) display the mean single and multiple-criterion evaluation scores (gold labels) across all models and prompting strategies. The results highlight that NLLB and LLaMA consistently outperform other models. Among prompting approaches, few-shot and chain-of-thought (CoT) techniques further boost LLaMA’s performance, with LLaMA-fewshot and LLaMA-fewshot-CoT achieving the highest mean scores overall. Conversely, mBART and Gemma in zero-shot settings yield significantly lower mean scores, suggesting their translations are less faithful or less fluent without additional context or guidance. In the multi-criteria viewpoint, few-shot and chain-of-thought (CoT) result in best-in-class results in all four evaluation dimensions. The fluency and completeness scores show the greatest overall gains with advanced prompting. Gemma lags behind the other models in zero-shot and few-shot settings, particularly on adequacy and style, though its scores improve noticeably with CoT and ReAct prompting.

For the general evaluation, all results and agreement metrics between gold labels and LLM-as-a-Judge scores for the various model-prompting pairs are reported in Tables 3, 4, 5, and 6, in the Appendix.

Overall, the LLM-as-a-Judge paradigm demonstrates strong agreement with expert gold labels, both in single- and multi-criteria evaluation. Agreement is generally highest for adequacy and fluency, with slightly lower scores for style and completeness, likely reflecting the greater subjectivity in those criteria. The results confirm the reliability of automatic LLM-based evaluation for Italian archaic-to-modern translation, and highlight the usefulness of prompting techniques such as few-shot and CoT to improve alignment with human judgment.

References

- Marah Abidin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benham, Misha Bilenko, Johan Björck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024. [Phi-3 technical report: A highly capable language model locally on your phone](#). *Preprint*, arXiv:2404.14219.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. [Language models are few-shot learners](#). *Preprint*, arXiv:2005.14165.
- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. [Multilingual denoising pre-training for neural machine translation](#). *Preprint*, arXiv:2001.08210.
- Sachin Mehta, Mohammad Hossein Sekhavat, Qingqing Cao, Maxwell Horton, Yanzi Jin, Chenfan Sun, Iman Mirzadeh, Mahyar Najibi, Dmitry Belenko, Peter Zatloukal, and Mohammad Rastegari. 2024. [Openelm: An efficient language model family with open training and inference framework](#). *Preprint*, arXiv:2404.14619.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1332 others. 2025. [Gemini: A family of highly capable multimodal models](#). *Preprint*, arXiv:2312.11805.
- Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, and 89 others. 2024. [Gemma: Open models based on gemini research and technology](#). *Preprint*, arXiv:2403.08295.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. [No language left behind: Scaling human-centered machine translation](#). *Preprint*, arXiv:2207.04672.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *Preprint*, arXiv:2307.09288.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. [Chain-of-thought prompting elicits reasoning in large language models](#). *Preprint*, arXiv:2201.11903.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. [React: Synergizing reasoning and acting in language models](#). *Preprint*, arXiv:2210.03629.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). *Preprint*, arXiv:2306.05685.

A LLM-as-a-Judge Prompt Templates

In this section, we will focus on the Evaluation prompts used from the LLM-as-a-Judge paradigm.

A.1 Single-Criteria Evaluation Prompt

The Single-Criteria Evaluation is based on a scale from 1(worst) to 5(best).

<p>You are an expert evaluator of machine</p> <p>→ translations from Archaic Italian to</p> <p>→ Modern Italian.</p> <p>For each translation, assign a score from 1</p> <p>→ (worst) to 5 (best), using this rubric:</p> <ol style="list-style-type: none"> 1: Completely unacceptable translation. The <ul style="list-style-type: none"> → translation has no pertinence with the → original meaning; the generated sentence → is either gibberish or makes no sense. 2: Severe semantic errors, omissions, or <ul style="list-style-type: none"> → substantial additions on the original → sentence. The errors are semantic and → syntactic in nature. Its still something → no human would ever write. 3: Partially wrong translation. The translation <ul style="list-style-type: none"> → is lackluster; it contains errors, but → mostly minor errors, like typos, or → small semantic errors. 4: Good translation. The translation is mostly <ul style="list-style-type: none"> → right, substantially faithful to the → original text, but the style does not → perfectly match the original sentence; → still fluent and comprehensible, and → could be semantically acceptable. 5: Perfect translation. The translation is <ul style="list-style-type: none"> → accurate, fluent, complete and coherent. → It retained the original meaning as much → as it could.
--

Evaluate ONLY the translation quality according
→ to these guidelines.

Original (Archaic Italian): {sentence}

Translation (Modern Italian): {translation}

Your score (1-5):

A.2 Multi-Criteria Evaluation Prompt

The Multi-Criteria Evaluation is based on always a scale from 1(worst) to 5(best) but in 4 different translation analyses: Adequacy, Fluency, Style and Completeness.

You are an expert evaluator of machine
→ translations from Archaic Italian to
→ Modern Italian.

For each translation, assign a score from 1
→ (worst) to 5 (best) on the following
→ four criteria. Here is the meaning of
→ each score for each criterion:

Adequacy:

- 1 - The translation does not capture the
→ original meaning at all.
- 2 - The translation is mostly wrong; the main
→ meaning is lost, but there are rare
→ fragments of meaning.
- 3 - Some meaning is preserved, but important
→ information is lost or altered.
- 4 - Most meaning is present, with only minor
→ issues; very little is lost.
- 5 - All essential meaning from the original is
→ preserved.

Fluency:

- 1 - The translation is unreadable or
→ ungrammatical; clearly machine-generated.
- 2 - The translation has severe grammar errors,
→ unnatural phrasing, or frequent
→ awkwardness.
- 3 - Some awkwardness or minor grammar issues,
→ but still understandable.
- 4 - Mostly fluent and grammatical, only rare
→ awkward or unnatural expressions.
- 5 - Perfectly fluent, fully natural Italian.

Style:

- 1 - The tone/register is completely lost or
→ inappropriate.
- 2 - The style is mostly lost; it is awkward or
→ inappropriate for the context.
- 3 - The style is partially preserved but
→ inconsistent or awkward.
- 4 - The style is almost fully preserved, with
→ only minor slips.
- 5 - The style, tone, and register are perfectly
→ matched to the original.

Completeness:

- 1 - Major parts are omitted or unnecessary
→ parts are added.
- 2 - The translation is incomplete; many
→ elements are missing or excessive
→ additions present.

- 3 - Minor omissions/additions, but most
→ information is present.
- 4 - Almost everything is present, with only
→ trivial information missing or added.
- 5 - Complete; nothing important is lost or
→ added.

Output ONLY the four scores as numbers 1-5, in
→ exactly this format (no extra text):

Adequacy: <score>

Fluency: <score>

Style: <score>

Completeness: <score>

Original (Archaic Italian): {sentence}

Translation (Modern Italian): {translation}

A.3 Dual-Judge and Debate-and-Consensus Prompts

For both *Phi 3.5* and *OpenELM*, we used the same prompt used for the others (Appendix A.2). Gemini is instead prompted with the following one:

Two expert judges scored this translation
→ independently.

Expert 1 scores:

Adequacy: {s1['Adequacy']}

Fluency: {s1['Fluency']}

Style: {s1['Style']}

Completeness: {s1['Completeness']}

Expert 2 scores:

Adequacy: {s2['Adequacy']}

Fluency: {s2['Fluency']}

Style: {s2['Style']}

Completeness: {s2['Completeness']}

Original (Archaic Italian): {sentence}

Translation (Modern Italian): {translation}

Please debate which scores are most accurate

→ and, if any should change,
output ONLY the final four scores in exactly
→ this format (no extra text):

Adequacy: <15>

Fluency: <15>

Style: <15>

Completeness: <15>

B Additional Results: Gold-Label Statistics

B.1 Mean Evaluation Scores by Model and Prompting Type

Here, we report the extended results got in Fig. 1 and Fig.2.

Table 1: Gold-label mean scores for each model and prompting strategy (multi-criteria).

Model/Prompt	Adequacy	Fluency	Style	Completeness
mBART	2.73	3.13	2.60	2.93
NLLB	4.37	4.43	4.27	4.50
Gemma (Zero-Shot)	2.73	3.87	2.80	3.03
LLaMA (Zero-Shot)	4.27	4.57	4.17	4.43
Gemma (Few-Shot)	3.40	4.00	3.40	4.07
LLaMA (Few-Shot)	4.40	4.53	4.30	4.57
Gemma (Few-Shot CoT)	3.73	4.03	3.73	4.10
LLaMA (Few-Shot CoT)	4.33	4.50	4.23	4.53
Gemma (Few-Shot ReAct)	3.63	4.20	3.63	3.90
LLaMA (Few-Shot ReAct)	4.00	4.27	3.97	4.17

C Evaluation Metrics: Explanation and Extended Results

The *Evaluation Metrics* used are:

- **Cohen’s Kappa:** Measures the degree of agreement between categorical labels assigned by the LLM and the gold annotations, while correcting for chance agreement.
- **Pearson Correlation:** Assesses the linear correlation between the numeric scores given by the LLM and those provided by human annotators.
- **Spearman’s and Kendall’s Rank Correlations:** Evaluate the consistency of ranking between LLM and gold annotations, capturing monotonic relationships even in the presence of non-linearities.
- **Exact Match:** Reports the proportion of cases where the LLM-as-a-Judge score exactly matches the human-assigned gold label.
- **Confusion Matrix:** Provides a detailed breakdown of agreement and disagreement patterns, illustrating where the LLM tends to over- or under-score compared to human judgment.

Speaking about the *Metrics’ Range*, we have that: Cohen’s κ (-1 to 1 ; 1 = perfect agreement, 0 = chance, < 0 = systematic disagreement), Pearson, Spearman, and Kendall (-1 to 1 ; 1 = perfect correlation), Exact Match (0 to 1 ; 1 = all predictions identical to gold). These complementary metrics allow for a comprehensive assessment of the agreement between automated and manual evaluation, both at the aggregate and granular level.

Table 2: Mean Gold-Label Scores - Single Criterion.


For each model and prompting strategy, we report the Single-Criterion evaluation.

Model/Prompt	Mean Score
mBART	2.53
NLLB	4.13
Gemma (Zero-Shot)	2.53
LLaMA (Zero-Shot)	4.03
Gemma (Few-Shot)	3.40
LLaMA (Few-Shot)	4.30
Gemma (Few-Shot CoT)	3.70
LLaMA (Few-Shot CoT)	4.23
Gemma (Few-Shot ReAct)	3.60
LLaMA (Few-Shot ReAct)	3.97

Example Debate & Consensus Log

In the following, we show two different results of the Debate and Consensus technique, one positive and one negative. We take the LLaMA CoT model example in both as a point of reference.

Debate & Consensus Log: Full Agreement (Low Scores) Prompt to LLM (Debate & Consensus):

Two expert judges scored this translation
 independently.

Expert 1 scores:
Adequacy: 1
Fluency: 1
Style: 1
Completeness: 1

Expert 2 scores:
Adequacy: 1
Fluency: 1
Style: 1
Completeness: 1

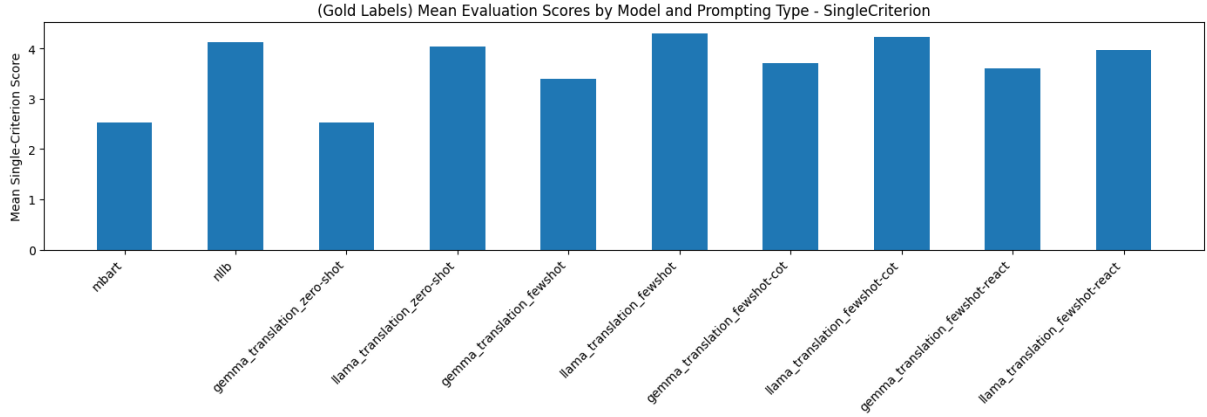


Figure 1: **(Gold Labels) Mean evaluation Scores - Single-criterion Setting.** NLLB and LLaMA, across all prompting variants, clearly lead in mean human scores, confirming their strong capacity for accurate archaic-to-modern Italian translation. Notably, few-shot and chain-of-thought prompting methods (especially with LLaMA) consistently enhance performance, sometimes surpassing even the transformer-based NLLB. In contrast, mBART and Gemma yield the lowest mean scores in zero-shot or minimal context settings, indicating limitations when used without prompt engineering. These findings emphasize the importance of both model selection and prompt design for optimal translation outcomes.

Table 3: **Mean Gold-Label Scores - Multi-Criteria.** For each model and prompting strategy, we report the Multi-Criteria evaluation.

Model/Prompt	Adequacy	Fluency	Style	Completeness
mBART	2.73	3.13	2.60	2.93
NLLB	4.37	4.43	4.27	4.50
Gemma (Zero-Shot)	2.73	3.87	2.80	3.03
LLaMA (Zero-Shot)	4.27	4.57	4.17	4.43
Gemma (Few-Shot)	3.40	4.00	3.40	4.07
LLaMA (Few-Shot)	4.40	4.53	4.30	4.57
Gemma (Few-Shot CoT)	3.73	4.03	3.73	4.10
LLaMA (Few-Shot CoT)	4.33	4.50	4.23	4.53
Gemma (Few-Shot ReAct)	3.63	4.20	3.63	3.90
LLaMA (Few-Shot ReAct)	4.00	4.27	3.97	4.17

Original (Archaic Italian): quella guerra ben
 ↳ fatta l' opera perch etc. Et dall' altra
 ↳ parte Aiaces era uno cavaliere franco e
 ↳ prode all' arme, di gran guisa, ma non
 ↳ era pieno di grande senno
 Translation (Modern Italian): La guerra ben
 ↳ fatta l'opera perch era un uomo di
 ↳ piccole dimensioni.

Please debate which scores are most accurate
 ↳ and, if any should change,
 output ONLY the final four scores in exactly
 ↳ this format (no extra text):

Adequacy: <15>
 Fluency: <15>
 Style: <15>
 Completeness: <15>

LLM Response:

The translation is severely flawed. The
 ↳ original text describes a well-executed
 ↳ war and then contrasts it with Aiaces, a
 ↳ brave but not particularly intelligent
 ↳ knight. The translation only mentions a
 ↳ well-executed war and a small man,
 ↳ completely missing the second part of
 ↳ the original text and misinterpreting
 ↳ the first.

Adequacy: 1
 Fluency: 1
 Style: 1
 Completeness: 1

Debate & Consensus Log: Strong Disagreement Resolved Prompt to LLM (Debate & Consen-

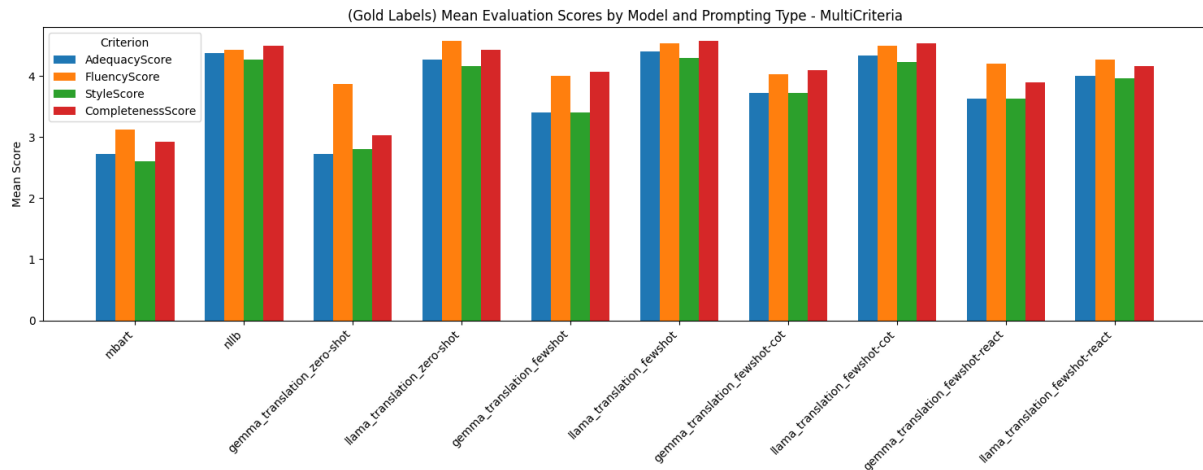


Figure 2: (Gold Labels) Mean evaluation Scores - Multi-criteria setting. Both NLLB and LLaMA (across all prompting strategies) consistently lead in mean human scores for all criteria, demonstrating robust translation quality, especially when enhanced by few-shot and chain-of-thought (CoT) prompts. Notably, LLaMA’s few-shot and CoT prompting achieves scores that are comparable to or even exceed those of the best transformer-based models, indicating the effectiveness of prompt-based reasoning for large language models. In contrast, mBART and Gemma (particularly in zero-shot) score substantially lower, with evident deficits in adequacy and style. From this, one can deduce that achieving high translation quality depends on both model architecture and sophisticated prompting, and that evaluating translations across multiple criteria provides a more granular and informative measure of a system’s strengths and weaknesses.

sus):

Two expert judges scored this translation
 ↳ independently.

Expert 1 scores:

Adequacy: 1

Fluency: 1

Style: 1

Completeness: 1

Expert 2 scores:

Adequacy: 5

Fluency: 4

Style: 5

Completeness: 5

Original (Archaic Italian): contra lui e contra

↳ le sue sorelle e contra il reame e

↳ contra l' alto pregio della sua

↳ ingenerazione e della sua familia

Translation (Modern Italian): Contra lui e

↳ contro le sue sorelle, contro il reame e

↳ contro l'alto pregio della sua

↳ ingenerazione e della sua famiglia.

Please debate which scores are most accurate

↳ and, if any should change,

output ONLY the final four scores in exactly

↳ this format (no extra text):

Adequacy: <15>

Fluency: <15>

Style: <15>

Completeness: <15>

The translation is a very close modernization
 ↳ of the original archaic Italian. Expert
 ↳ 2's scores seem more accurate. The
 ↳ translation is highly adequate, fluent,
 ↳ and stylistically appropriate, and it
 ↳ completely conveys the meaning of the
 ↳ original. Expert 1's scores are too low.

Adequacy: 5

Fluency: 4

Style: 5

Completeness: 5

LLM Response:

Table 4: **Gold Labels VS LLM-as-a-Judge Agreement/Correlation Metrics - Single-Criteria Evaluation.** LLaMA (especially in the zero-shot configuration) consistently demonstrates the highest correlation and agreement with gold annotations, as reflected by the strongest Pearson, Spearman, and exact match scores. mBART also achieves competitive results. In contrast, Gemma’s few-shot and chain-of-thought prompting strategies exhibit notably weaker alignment, with low kappa and exact match values. These results underscore the importance of both model architecture and prompting approach for maximizing the reliability of automatic translation evaluation. For all models, moderate agreement with human judgment is observed, but there is substantial variability depending on the strategy.

Model/Prompt	Cohen’s κ	Pearson	Spearman	Kendall	Exact Match
mBART	0.35	0.66	0.69	0.64	0.50
NLLB	0.26	0.57	0.63	0.56	0.47
Gemma (Zero-Shot)	0.25	0.32	0.42	0.36	0.47
LLaMA (Zero-Shot)	0.48	0.78	0.78	0.72	0.67
Gemma (Few-Shot)	0.03	0.30	0.22	0.20	0.10
LLaMA (Few-Shot)	0.25	0.69	0.66	0.61	0.53
Gemma (Few-Shot CoT)	0.08	0.31	0.32	0.29	0.17
LLaMA (Few-Shot CoT)	-0.02	0.03	0.10	0.09	0.30
Gemma (Few-Shot ReAct)	0.07	0.47	0.50	0.45	0.20
LLaMA (Few-Shot ReAct)	0.24	0.65	0.67	0.58	0.40

Table 5: **Gold labels and LLM-as-a-Judge Agreement/Correlation Metrics - Multi-Criteria.** For all four criteria (Adequacy, Fluency, Style, Completeness) in multi-criteria evaluation ($n = 30$ for all cases), these results show that LLaMA (especially with few-shot and chain-of-thought prompting) and mBART consistently achieve the highest levels of agreement with gold-standard human annotations across adequacy, fluency, style, and completeness. The boost observed with advanced prompting strategies, particularly for LLaMA, demonstrates the practical importance of prompt engineering in automated translation evaluation. In contrast, Gemma-based models display more variable and generally lower alignment with human judgment, underlining differences in model capability. Overall, style and completeness are persistently more challenging for automatic assessment than adequacy and fluency, as shown by lower kappa and correlation scores.

Criterion	Model/Prompt	κ	Pearson	Spearman	Kendall	Ex. Match
Adequacy	mBART	0.37	0.66	0.67	0.60	0.50
	NLLB	0.29	0.39	0.48	0.45	0.60
	Gemma Zero-Shot	0.14	0.39	0.45	0.38	0.30
	LLaMA Zero-Shot	0.30	0.61	0.62	0.57	0.53
	Gemma Few-Shot	0.10	0.30	0.34	0.28	0.17
	LLaMA Few-Shot	0.32	0.69	0.65	0.60	0.57
	Gemma CoT	0.15	0.37	0.39	0.33	0.23
	LLaMA CoT	0.28	0.52	0.50	0.45	0.33
	Gemma ReAct	0.15	0.45	0.47	0.40	0.27
	LLaMA ReAct	0.20	0.62	0.64	0.56	0.47
Fluency	mBART	0.15	0.50	0.52	0.46	0.33
	NLLB	0.22	0.51	0.57	0.52	0.30
	Gemma Zero-Shot	0.07	0.42	0.48	0.41	0.27
	LLaMA Zero-Shot	0.19	0.59	0.61	0.55	0.37
	Gemma Few-Shot	0.03	0.27	0.33	0.25	0.13
	LLaMA Few-Shot	0.28	0.70	0.68	0.61	0.50
	Gemma CoT	0.09	0.31	0.37	0.30	0.17
	LLaMA CoT	0.21	0.56	0.54	0.48	0.27
	Gemma ReAct	0.12	0.44	0.46	0.39	0.23
	LLaMA ReAct	0.18	0.64	0.66	0.57	0.37
Style	mBART	0.36	0.60	0.61	0.54	0.50
	NLLB	0.20	0.39	0.48	0.41	0.47
	Gemma Zero-Shot	0.09	0.31	0.35	0.28	0.17
	LLaMA Zero-Shot	0.15	0.55	0.57	0.49	0.23
	Gemma Few-Shot	0.03	0.23	0.29	0.21	0.07
	LLaMA Few-Shot	0.23	0.68	0.64	0.58	0.43
	Gemma CoT	0.07	0.24	0.30	0.24	0.10
	LLaMA CoT	0.17	0.49	0.51	0.44	0.20
	Gemma ReAct	0.09	0.33	0.38	0.31	0.17
	LLaMA ReAct	0.11	0.62	0.63	0.54	0.30
Completeness	mBART	0.33	0.68	0.67	0.59	0.47
	NLLB	0.31	0.59	0.61	0.53	0.47
	Gemma Zero-Shot	0.12	0.41	0.43	0.36	0.20
	LLaMA Zero-Shot	0.18	0.57	0.58	0.50	0.23
	Gemma Few-Shot	0.04	0.28	0.30	0.23	0.13
	LLaMA Few-Shot	0.25	0.65	0.63	0.56	0.40
	Gemma CoT	0.08	0.35	0.37	0.30	0.17
	LLaMA CoT	0.14	0.51	0.53	0.45	0.20
	Gemma ReAct	0.09	0.37	0.39	0.32	0.17
	LLaMA ReAct	0.13	0.59	0.60	0.51	0.23

Table 6: **Gold labels and LLM-as-a-Judge Agreement/Correlation Metrics - Multi-Criteria Debate & Consensus.** For all four criteria (Adequacy, Fluency, Style, Completeness) in the multi-criteria setting ($n = 30$ for all cases). In particular, for subjective aspects such as Style and Completeness, the D&C approach shows variable and generally low agreement with gold labels. Notably, only the LLaMA zero-shot configuration achieves high scores across all metrics, indicating the continued importance of strong base models. The results reinforce that D&C aggregation is not a panacea for reference-free evaluation and that substantial challenges remain in automating the assessment of nuanced linguistic features.

Criterion	Model/Prompt	κ	Pearson	Spearman	Kendall	Ex. Match
Adequacy	mBART	0.13	0.49	0.45	0.40	0.33
	NLLB	0.16	0.49	0.48	0.42	0.33
	Gemma Zero-Shot	0.26	0.41	0.46	0.38	0.40
	LLaMA Zero-Shot	0.07	0.31	0.28	0.26	0.23
	Gemma Few-Shot	0.09	0.00	-0.08	-0.07	0.20
	LLaMA Few-Shot	-0.05	0.38	0.29	0.24	0.17
	Gemma CoT	0.05	0.28	0.36	0.31	0.17
	LLaMA CoT	-0.03	-0.08	-0.01	-0.01	0.17
	Gemma ReAct	0.09	0.43	0.49	0.42	0.20
	LLaMA ReAct	-0.01	0.55	0.56	0.48	0.13
Fluency	mBART	-0.04	0.43	0.40	0.35	0.13
	NLLB	0.07	0.48	0.46	0.41	0.27
	Gemma Zero-Shot	0.03	0.20	0.26	0.22	0.27
	LLaMA Zero-Shot	0.12	0.25	0.35	0.31	0.43
	Gemma Few-Shot	0.01	0.00	0.00	0.00	0.17
	LLaMA Few-Shot	0.05	0.23	0.17	0.16	0.13
	Gemma CoT	0.16	0.49	0.51	0.46	0.37
	LLaMA CoT	-0.04	-0.11	-0.05	-0.04	0.20
	Gemma ReAct	0.09	-0.01	0.02	0.02	0.23
	LLaMA ReAct	-0.05	0.08	0.07	0.06	0.23
Style	mBART	0.10	0.35	0.29	0.27	0.33
	NLLB	-0.09	0.47	0.46	0.41	0.07
	Gemma Zero-Shot	0.31	0.39	0.42	0.36	0.47
	LLaMA Zero-Shot	0.08	0.05	0.09	0.06	0.23
	Gemma Few-Shot	0.00	0.05	-0.02	-0.01	0.07
	LLaMA Few-Shot	-0.14	0.24	0.22	0.18	0.07
	Gemma CoT	0.04	0.35	0.33	0.30	0.17
	LLaMA CoT	-0.14	-0.17	-0.03	-0.03	0.03
	Gemma ReAct	-0.04	0.53	0.51	0.47	0.07
	LLaMA ReAct	-0.07	0.19	0.18	0.16	0.17
Completeness	mBART	0.24	0.31	0.36	0.31	0.40
	NLLB	0.09	0.09	0.15	0.13	0.43
	Gemma Zero-Shot	0.17	0.47	0.48	0.40	0.33
	LLaMA Zero-Shot	0.08	-0.07	0.00	0.00	0.40
	Gemma Few-Shot	0.05	0.05	0.03	0.03	0.13
	LLaMA Few-Shot	0.09	0.20	0.12	0.10	0.17
	Gemma CoT	-0.11	0.10	0.09	0.08	0.00
	LLaMA CoT	0.09	0.20	0.12	0.10	0.17
	Gemma ReAct	0.05	0.50	0.51	0.45	0.17
	LLaMA ReAct	-0.00	0.25	0.25	0.21	0.23