Evaluating the Effect of Prompt Language on LLM-based Translation: Evidence from Spanish<>Italian Translation

Antonella Bove
First Author - PhD student
Ca' Foscari University of Venice
antonella.bove@unive.it

Paola Di Cataldo
Landoor
paola.dicataldo@landoor.com

Davide Maestroni
Landoor
davide.maestroni@landoor.com

Abstract

The integration of large language models (LLMs) into translation practice has substantially reshaped translation workflows (Kornacki and Pietrzak, 2025). Since translation quality depends partly on how these models are prompted, prompt design deserves closer attention as a key stage of the LLM-augmented translation process. This study investigates Spanish<>Italian translation with GPT 5.1 in the advertising and biomedical domains. It examines whether prompt language affects the translations generated by the model, focusing on whether prompts written in the target language generate translations that are preferred over those generated from English prompts, given that English is the language most prevalent in the model's training data (Armengol-Estapé et al., 2022). In this study, preference is operationalised in terms of *post-editing usefulness*, that is, the perceived suitability of a translation as a starting point for subsequent human post-editing. Three prompt templates, varying in complexity and informational content, were tested. The translations were first screened for textual similarity, and only the translations generated from the template that produced the greatest variation across outputs were subsequently selected for human evaluation. Human preferences were collected through a pairwise-comparison task. The findings indicate that translations generated from prompts written in the target language tend to receive more favorable preference judgments than those produced using English-language prompts.

1 Introduction

The widespread adoption of LLMs into translation practice has reshaped translation workflows. Indeed, neural machine translation (NMT) systems and large language models (LLMs) not only differ in training data and purposes (Brown et al., 2020; Hendy et al., 2023; Balashov, 2025; Forcada, 2017); they also give rise to fundamentally different translation workflows (Kornacki and Pietrzak, 2025). In a conventional NMT-based translation workflow, human intervention tends to concentrate in two stages: pre-editing (when source text is prepared so the MT system is more likely to produce a better translation) and post-editing (which implies revising the raw output into a publishable target text, ensuring quality). In LLM-augmented translation workflows, by contrast, a new stage comes into play: *prompt design* (Yamada, 2023). In that stage, the translator and linguist has to instruct the machine by providing it with clear and precise instructions to obtain the desired output. Domain, purpose, target audience, style and terminology constraints are only some of the information that make up a *translation brief* (Reiss & Vermeer, 1984/2013) and that should be included in a well-designed prompt (Kornacki and Pietrzak, 2025; He, 2024; Yamada, 2023). This is particularly important because LLM's behaviour is highly sensitive to prompting: while such sensitivity allows the same model to be adapted to a wide range of Natural Language Processing (NLP) tasks, even minor variations can produce substantial differences in the generated output (Zhang et al., 2023). However, because prompting has only recently emerged as a field of study, there is still a clear need for research on its optimal formulation. This need is especially evident in machine translation,

where research on prompting remains comparatively sparse and fragmented and a lot of questions remain to be answered. In this study, we aim to contribute to this line of research by examining a core aspect of prompt design for LLM-augmented-translation workflows, that is, prompt language. Specifically, this study asks whether the language of the prompt influences the perceived usefulness of LLM-generated translations for human post-editing. In particular, it compares prompts written in English, the language most prevalent in model training data (Armengol-Estapé et al., 2022)[1], with prompts written in the target language of the translation.

This investigation was conceived and developed within an ongoing PhD project in collaboration with Landoor, a Translate One Company. The project investigates the emerging translation workflows enabled by large language models, with particular attention to how these technologies reshape post-editing processes and the role of linguists in assessing and refining machine-generated outputs. The involvement of an LSP gives the study a clear applied orientation, informing a methodological design situated at the intersection of academic research and professional translation practice. Accordingly, beyond addressing a gap in current research on prompting for translation, the study aims to address aspects that are relevant to real-world translation workflows.

2 Related research

A growing body of research has examined how prompt design affects LLM performance across different NLP tasks (Aldosari & Altuwairesh, 2025; Briakou et al., 2024; Wu et al., 2023). However, most of this work focuses on prompting techniques, while comparatively little attention has been paid to the language of the prompt. This gap is even more pronounced in machine translation, where studies on prompt language in LLM-based translation remain especially scarce.

Lin et al. (2022) provide evidence that prompt language can measurably affect LLM performance. Focusing on natural language inference and commonsense reasoning tasks, they found that English prompts generally yield better results than prompts written in other languages. At the same time, they report notable exceptions: Spanish is highly competitive for natural language inference tasks, while Chinese performs comparably well for commonsense reasoning. For the machine translation task, by contrast, the prompt formats used are essentially non-verbal (a source sentence followed by a mask token), reducing the opportunity to investigate the impact of language on that specific task. Zhang et al. (2023) explicitly investigate prompt-language effects on translation tasks. However, their prompts include relatively little verbal instructional content compared with the more detailed prompts typically used in professional translation workflows. Their analysis primarily contributes to research on the impact of in-context learning, rather than on the effect of language itself. Enomoto et al. (2025) provide complementary evidence through a human-in-the-loop approach to prompt construction. Their results indicate that whether English or target-language instructions are preferable depends on the task. Moreover, they find that semantically equivalent instructions written in different languages often lead to divergent generations, with more than half of outputs differing between language conditions.

Taken together, these studies suggest that prompt language is a consequential variable in LLM-based NLP tasks and that, in the context of translation, it warrants more systematic investigation under prompting conditions that are more representative of professional translation workflows. Moreover, as Yamada (2026) insightfully observes, the majority of this research originates in the field of NLP and, consequently, approaches translation from an engineering perspective. While undoubtedly valuable, further research by translation scholars and linguists is needed to assess LLM outputs in ways that reflect the complexities of real translation practice to better understand how these models can be integrated into professional translation settings.

3 Methodology

The methodology adopted in this study consisted of five main stages. It began with the selection of materials, followed by prompt design, model

[1] Although Armengol-Estapé et al. (2022) explicitly focus on GPT-3, the OpenAI figures they cite concerning the predominance of English in the training data remain a useful indirect reference point for subsequent GPT models. However, to the best of our knowledge, OpenAI has not publicly disclosed a language-by-language distribution of the training data used for models in the GPT-5 family.

configuration and generation procedures, text-similarity pre-screening, and, lastly, human evaluation. Each stage is described in detail below.

3.1 Materials selection

Corpora of authentic source texts were compiled for two domains (advertising and biomedical) and two translation directions (Spanish→Italian and Italian→Spanish). For each domain–direction combination, ten texts were selected resulting in a corpus of 10 advertising texts originally written in Spanish, 10 advertising texts originally written in Italian, 10 biomedical texts originally written in Spanish, and 10 biomedical texts originally written in Italian. The selected texts were checked to ensure that no Italian or Spanish translations were publicly available online, thereby reducing the likelihood that the outputs generated during the experiment would reproduce pre-existing translations. Details about the corpora are provided in Appendix A.

The choice of advertising and biomedical domains was motivated by both practical and methodological considerations. Within the partner company involved in this study, these domains represent areas in which translation demand is particularly frequent, making them especially relevant for investigating real translation workflows. At the same time, the two domains reflect different communicative and translational constraints: advertising texts typically allow for greater creativity and adaptation, whereas biomedical texts require terminological precision and adherence to specialized discourse conventions. Including two highly distinct domains also allows us to investigate the extent to which the findings are robust across different translation settings.

3.2 Prompt design

Prompt creation followed a "human-in-the-loop instruction construction procedure" adapted from Enomoto et al. (2025). First, we established a set of informational elements to be included in each prompt. Then, we produced two parallel versions of each prompt —one in English and one in the target language (Spanish or Italian). Unlike Enomoto et al. (2025), where instructions were generated using an LLM, all prompts in this study were drafted manually in both languages. Each English-target language pair was then checked to ensure content equivalence (i.e., identical instructions and contextual information). Finally, the prompts were reviewed by the authors to confirm linguistic adequacy. This procedure was designed to keep informational content constant while varying only the language of the prompt. We tested three prompt templates: a *basic* template (P1), a *persona* template (P2), and a *functional* template (P3)[2]. Prompts were written both in English, representing the language most prevalent in the model's training data and in the target language of the translation, i.e., Italian or Spanish. When Spanish was the target language, the European Spanish[3] variety was specified.

Below are the three templates under investigation. For a full example, see Appendix B.

P1	Translate from [SL] into [TL].
P2	You are an expert translator specialized in [domain] translation. Translate the following text from [SL] into [TL].
P3	Translate from [SL] into [TL] the following text intended for [purpose] purposes for the official website of [sender, brief description]. The text is addressed to [addressee] and will be published [medium and date]. Use a [style] style.

Table 1. Prompt templates in English

3.3 Model configuration and generation procedure

All generations were conducted with a fixed snapshot of GPT-5.1, namely `gpt-5.1-2025-11-13`. As described by OpenAI, snapshots allow researchers to lock a specific model version so that performance and behavior remain stable across runs. All generations were produced via API with temperature set at zero (T=0) and reasoning effort set at 'none'[4]. The zero-temperature setting was adopted to minimize sampling variability and isolate, as far as possible, the effect of prompt

[2] We label this prompt *functional* because it draws on Functionalist approaches from Translation Theory (cfr. Reiss, K., & Vermeer, 1984/2013). Specifically, it serves as a succinct translation *brief* by consolidating the essential contextual and task-related information needed to guide the translation process effectively.

[3] We specified this variety because it corresponds to the Spanish variety spoken by the evaluators in our pool.

[4] Reasoning effort was set at none because OpenAI allows explicit control over sampling parameters (e.g., temperature) only when reasoning is disabled for GPT-5 family models.

language. Translations were collected in two sessions (January 30 and February 4, 2026).

The model was selected to reflect contemporary professional practice at the time the study was conceived, as GPT-5.1 was both the most recent version available and one widely used by the professional translators involved in the broader collaboration underlying this research. Given the rapid pace of development in AI systems, maintaining a single fixed model version was also necessary to ensure experimental consistency and reproducibility, avoiding potential variability introduced by model updates.

3.4 Text similarity pre-screening

The evaluation of the impact of prompt language on the translations relied on expert human preference judgements. Since human evaluation is resource-intensive, we first conducted a quantitative screening step to avoid submitting near-identical outputs to raters. Similarity between translations was measured using *Levenshtein distance* (Levenshtein, 1966). This metric quantifies the minimum number of edit operations (i.e., deletion, insertion and substitution) required to transform one string *a* into another string *b*. We are aware that it is a relatively surface-level measure, as it does not capture semantic relationships between texts. Nevertheless, we consider it suitable for the present purpose because it is used only as a preliminary screening step before human evaluation.

The quantitative analysis indicates that the effect of prompt language on LLM-generated translations increases with prompt complexity. Specifically, the lower the informational content of the prompt, the smaller the impact of prompt language. Conversely, as the informational density and complexity of the prompt increase, the influence of prompt language becomes more pronounced, resulting in greater formal divergence between the translations. As an example, in the advertising domain (Italian→Spanish direction), the mean raw *Levenshtein distance* between translations generated from P1 is 30.4.; this increases to 33.9 for translations generated from P2 and rises sharply to 82.5 for those generated from P3 (see Appendix C for details on the other directions and domains). Moreover, the *Levenshtein distance* between translations generated from Italian and English prompts is greater than that observed between translations generated from Spanish and English prompts. However, this difference may simply reflect the fact that the Spanish source texts are, on average, slightly longer than the Italian source texts. A further consideration worth noting is that, in the biomedical domain, language appears to have a stronger overall impact than in the advertising domain, as reflected in the larger distances observed between the means. However, these differences remain relatively small. Ultimately, expert human judgment will determine whether such minimal differences are negligible or still have an impact on the perceived suitability of a translation as a starting point for subsequent post-editing. Based on the *Levenshtein distance* values, we selected only the translations generated from the *functional*-template prompt for human evaluation, as it yielded the largest differences across texts.

3.5 Human evaluation task

For human evaluation, we implemented a within-subject, pairwise comparison task. We ran four separate experiments, corresponding to each domain–direction combination (advertising ES→IT, advertising IT→ES, biomedical ES→IT, biomedical IT→ES). In each experiment, participants completed ten trials. In every trial, they first read the source text displayed at the top of the screen and then compared two candidate translations presented underneath side by side ("Translation A" and "Translation B"). Items were counterbalanced to reduce position effects: in each experiment, 5 trials displayed the translation generated with the prompt in target-language on the left and 5 trials displayed the translation generated with the English prompt on the right, following an alternating order. Items were then randomized to mitigate fatigue-related biases. Participants were not informed of the purpose of the experiment, meaning that they were not told what the comparison was intended to test. To facilitate comparison, differences between the two translations were highlighted. After reading both versions, participants indicated their preference by choosing one of three options: *Translation A, Translation B*, or *No preference*. The *No preference* option was reserved for cases in which the two translations were perceived as genuinely equivalent, with no meaningful preference for either version. Participants were instructed to indicate the translation they considered to be *a better initial version for a subsequent human post-editing phase*. In line with the broader aims of the

PhD project and the applied orientation of the study, preference was therefore operationalised in terms of *post-editing usefulness*, understood as the perceived viability of a given version as a starting point for subsequent post-editing. This criterion was both accessible and consistent for participants to apply, simulating a natural decision-making process within a realistic professional scenario, while also providing insights of practical relevance for professional translation workflows. Based on the *post-editing usefulness* criterion, this approach is not intended to replace established quality-assessment research, but rather to complement it.

3.6 Platform, ethics, and participants

The experiment was implemented in Gorilla and received approval from the Ethics Committee of Ca' Foscari University of Venice. Participants signed an informed consent form for personal data processing and completed a pre-task questionnaire collecting age range, nationality, professional profile, years of translation experience (if applicable), native language, and source and target working languages. Participants included professional translators, university professors in Translation Studies, and MA-level graduates in Translation Studies. Participant details for each experiment are reported in the corresponding *Results* section.

4 Results

Below, the results are reported by domain and translation direction. The results are presented in tables structured as follows:

- TextID column indicates the identifier of the evaluated text;
- "IT"/"ES", "EN", and "NP" columns report, respectively, the number of times participants preferred the translation generated with the target-language prompt (Italian/Spanish), the English prompt, or indicated no preference (NP).
- Majority-preferred version (MPV): to support interpretation and readability of the findings, a majority rule column was added. Specifically, the preferred version was defined as the one selected by most participants. Accordingly, this column indicates the language of the prompt in which the majority-preferred translation was generated.

4.1 Advertising domain

For the advertising domain in the ES>IT translation direction, evaluations were collected from 10 participants: 5 professional translators, 3 holders of a master's degree in translation and interpreting, and 2 university professors of translation and/or interpreting. Seven participants were native speakers of Italian, while three participants were native speakers of Spanish. All participants reported ES>IT as part of their working language combination; therefore, their evaluations were deemed eligible and were included in the analysis.

TextID	IT	EN	NP	MPV
T01	5	3	2	IT
T02	6	2	2	IT
T03	6	2	2	IT
T04	2	5	3	EN
T05	7	1	2	IT
T06	7	2	1	IT
T07	7	1	2	IT
T08	5	2	3	IT
T09	4	5	1	EN
T10	2	6	2	EN

Table 2. Results for the advertising domain in the ES>IT direction.

For the IT>ES direction, evaluations were collected from 11 participants: 6 professional translators, 3 holders of a master's degree in translation and interpreting, and 2 university professors of translation and/or interpreting. Seven participants were native speakers of Italian, while four were native speakers of Spanish. Since all participants indicated the IT>ES combination in their working languages, all preferences were eligible and included in the analysis.

TextID	ES	EN	NP	MPV
T01	6	2	3	ES
T02	2	7	2	EN
T03	6	4	1	ES
T04	4	4	3	ND
T05	6	4	1	ES
T06	4	1	6	ES
T07	6	4	1	ES
T08	6	2	3	ES
T09	4	6	1	EN
T10	3	5	3	EN

Table 3. Results for the advertising domain in the IT>ES translation direction.

4.2 Biomedical domain

For the biomedical domain in the ES>IT language direction, the following participants took part in

the evaluation: 5 professional translators, 3 holders of a master's degree in translation and interpreting, and 2 university professors of translation and/or interpreting. Seven participants were native speakers of Italian, while three were native speakers of Spanish. All participants reported ES>IT as part of their working language combination; therefore, their preferences were deemed eligible and were included in the analysis.

TextID	IT	EN	NP	MPV
T01	7	2	1	IT
T02	7	1	2	IT
T03	1	4	5	NP
T04	5	5	0	ND
T05	8	1	1	IT
T06	6	1	3	IT
T07	5	4	1	IT
T08	0	7	3	EN
T09	5	2	3	IT
T10	5	2	3	IT

Table 4. Results for the biomedical domain in the ES>IT translation direction.

In the biomedical domain, for the IT>ES translation direction, the evaluation involved five professional translators, three holders of a master's degree in translation and interpreting, and three university professors of translation and/or interpreting. Seven participants were native speakers of Italian, while four were native speakers of Spanish. Since all participants reported IT>ES in their working language combinations, their preferences were considered valid and were therefore included in the analysis.

TextID	IT	EN	NP	MPV
T01	0	7	4	EN
T02	7	2	2	ES
T03	5	0	6	NP
T04	4	4	3	ND
T05	6	1	4	ES
T06	6	3	2	ES
T07	1	4	6	NP
T08	4	5	2	EN
T09	8	1	2	ES
T10	10	0	1	ES

Table 5. Results for the biomedical domain in the IT>ES translation direction.

5 Discussion

The discussion is organised in two parts. The first part examines the descriptive results at item level, focusing on the majority-preferred version across domains and translation directions. The second part complements this descriptive account with statistical analyses of individual preference judgments, assessing whether the observed tendency remains statistically supported while accounting for participant- and item-level variability.

Across domains and translation directions, the descriptive results indicate a recurrent preference for translations generated with prompts formulated in the target language. In the advertising domain, translations generated with Italian prompts were majority-preferred in seven out of ten ES>IT texts (7/10), whereas English-prompt outputs were majority-preferred in only two cases (2/10). A similar pattern emerged in the IT>ES direction, where translations generated with Spanish prompts were majority-preferred in six out of ten texts (6/10), compared with three cases in which translations generated with English prompts were preferred by the majority (3/10). One text (T04) produced no clear majority preference, as both versions received the same number of preference indications (1/10).

This tendency was also observed in the biomedical domain. In the ES>IT direction, outputs generated with Italian prompts were majority-preferred in seven out of ten texts (7/10), while the English-prompt version was majority-preferred in only one case (1/10). For text T04, no majority preference could be established, while for text T03 the two versions were more frequently considered to be equivalent. In the IT>ES direction, translations generated with Spanish prompts were majority-preferred in five out of ten texts (5/10), whereas English-prompt outputs were majority-preferred in two cases (2/10). Text T03 and T07 were more frequently judged to be equivalent (2/10). Text T04 did not yield a majority-preferred version, as both versions received the same number of preference indications. Notably, texts T09 and T10 represent particularly salient instances, as preferences clearly and consistently favoured the version generated using a prompt in the target language. Overall, these item-level results point to a general tendency in favour of target-language prompts, while also showing some variation across texts.

To determine whether this descriptive tendency remained evident after accounting for variability across participants and texts, statistical analyses were conducted using binary logistic mixed-

effects models, with a logit link function[5]. To account for the repeated-measures structure of the data, all models include random intercepts for study-specific participants and study-specific items. Whereas the item-level summaries reported above describe majority preferences for each text, the statistical analysis was conducted on individual preference judgments.

To assess whether participants showed a systematic preference for translations generated with target-language prompts, a binary logistic mixed-effects model to non-NP responses (No Preference) was estimated. Random intercepts for study-specific participants and study-specific items were included to account for repeated observations. The intercept-only model indicated an overall preference for translations generated from target-language prompts (TARGET), with an estimated probability of 0.646, significantly above chance ($p = 0.014$; OR = 1.82, 95% CI [1.13, 2.94]).

A second additive model examined whether this preference was associated with native language, translation direction and domain. In this model, native language is a significant predictor: Spanish native speakers were more likely than Italian native speakers to prefer TARGET ($p = 0.038$; OR = 2.18, 95% CI [1.04, 4.54]). By contrast, neither translation direction nor domain shows evidence of an effect. Estimated marginal means showed TARGET preference probabilities of 0.588 for Italian speakers (95% CI [0.460, 0.705]) and 0.756 for Spanish speakers (95% CI [0.606, 0.862]). However, the additive model did not clearly improve fit over the intercept-only model (χ^2 (3) = 4.705, $p = 0.195$), so the native-language effect should be interpreted with caution. In sum, although the percentages differ across groups, both show estimated probabilities above 50%, suggesting an overall tendency to prefer TARGET. This tendency is less pronounced among Italian native speakers, whose confidence interval includes values close to chance level.

To better understand whether the native-language effect reflected linguistic background itself or differences in the participant profiles represented in each group, an exploratory model including participant profile was estimated. Adding profile improved model fit relative to the additive model ($p = .032$). *Professor* participants were more likely than *Graduate* participants to prefer TARGET ($p = 0.007$; OR = 6.46, 95% CI [1.65, 25.38]), whereas the positive effect for *Translator* participants did not reach significance ($p = 0.076$; OR = 1.96). Importantly, after including profile, native language was no longer significant ($p = 0.601$), suggesting that the native-language effect observed in the additive model may partly reflect differences in profile composition. In other words, the difference between Italian and Spanish native speakers identified in the previous model may not be due to native language alone. Once participant profile is taken into account, the model suggests that profile-related differences may explain part of the pattern: *Professors*, in particular, showed a much stronger tendency than *Graduate* participants to prefer TARGET. *Translator* participants also appeared more likely than *Graduate* participants to prefer TARGET, but this tendency was not strong enough to be considered statistically reliable. This result, however, should be treated as exploratory rather than conclusive, and would require further research with larger groups. More broadly, this pattern suggests that the tendency to select TARGET becomes more pronounced among participants with more advanced knowledge of translation or greater professional specialisation. This may be plausibly due to a greater sensitivity to subtle textual features that are less immediately apparent to less experienced profiles. However, further research on quality assessment is needed to determine whether, and how, these preferences relate to output quality.

6 Conclusions and future research

In this study, we examined the impact of prompt language on GPT 5.1 in Italian<>Spanish advertising and biomedical translation. Our findings yield several noteworthy implications.

Firstly, the quantitative screening revealed an interesting pattern: the effect of prompt language becomes more apparent as the amount of information included in the prompt increases. This is a key point for future research investigating the role of prompt language.

[5] The authors gratefully acknowledge Davide Benussi from the Department of Statistical Sciences of the University of Padova for his valuable support with the statistical analyses reported in this study.

Secondly, the expert human preference judgments indicate that, across both the advertising and biomedical domains, translations generated from target-language prompts were more often preferred as a starting point for a subsequent post-editing process than those generated from English-language prompts. This preference was statistically significant and did not appear to vary systematically across domains or translation directions. This result is consistent with Enomoto (2025) in suggesting that the language in which a prompt is formulated can be as consequential as the prompt's content.

Finally, the statistical analysis suggests that the evaluator's profile may modulate preference judgements. Participants with more advanced knowledge of translation or greater professional experience seem to show a stronger preference for translations generated from prompts written in the target language.

The findings should also be interpreted in light of some limitations that should be addressed in future research. On the one hand, the participant sample was limited and comparatively heterogeneous. On the other hand, the research could be further developed by incorporating a quality assessment, that is, a systematic analysis of the linguistic features that may guide these preferences. This would shed light on the relationship between human preferences and translation quality.

7 References

Aldosari, Lama Abdullah, and Nasrin Altuwairesh. (2025). Assessing the effects of translation prompts on the translation quality of GPT-4 Turbo using automated and human evaluation metrics: a case study. *Perspectives*, 1–25. https://doi.org/10.1080/0907676X.2025.2464120

Armengol-Estapé, Jordi, Ona de Gibert Bonet, and Maite Melero. (2022). On the Multilingual Capabilities of Very Large-Scale English Language Models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 3056–3068, Marseille, France. European Language Resources Association.

Balashov, Yuri. (2025). Translation in the Wild. *Information, 16*(12), 1077. https://doi.org/10.3390/info16121077

Briakou, Eleftheria, Luo, Jiaming, Cherry, Colin, & Freitag, Markus. (2024). Translating Step-by-Step: Decomposing the Translation Process for Improved Translation Quality of Long-Form Texts. *Proceedings of the Ninth Conference on Machine Translation*, 1301–1317. https://doi.org/10.18653/v1/2024.wmt-1.123

Brown, Tom, 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, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. (2020). Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, Article 159, 1877–1901.

Enomoto, Taisei, Kim, Hwichan, Chen, Zhousi, & Komachi, Mamoru (2025). A fair comparison without translationese: English vs. Target-language instructions for multilingual LLMs. Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), 649–670. Presented at the Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), Albuquerque, New Mexico. doi:10.18653/v1/2025.naacl-short.55

Forcada, Mikel L. (2017). Making sense of neural machine translation. *Translation Spaces, 6*(2), 291–309. https://doi.org/10.1075/ts.6.2.06for

He, Sui. (2024). Prompting ChatGPT for Translation: A Comparative Analysis of Translation Brief and Persona Prompts. *Proceedings of the 25th Annual Conference of the European Association for Machine Translation*, *(1)*, 316–326. Sheffield, UK. European Association for Machine Translation (EAMT). https://aclanthology.org/2024.eamt-1.27/

Hendy, Andy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, Hany Hassan Awadalla. (2023). How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation. arXiv. https://doi.org/10.48550/arXiv.2302.09210

Kornacki, Michał and Paulina Pietrzak. (2025). *Hybrid workflows in translation: Integrating GenAI into translator training*. Routledge.

Levenshtein, V. I. (1966). Binary coors capable or 'correcting deletions, insertions, and reversals. In *Soviet Physics-Doklady, volume 10*.

Lin, Xi Victoria, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. (2022). Few-shot Learning with Multilingual Generative Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Reiss, Katharina and Vermeer, Hans J. (2013). *Towards a General Theory of Translational Action: Skopos Theory Explained* (C. Nord, Trans.; 1st ed.). Routledge. (Original work published in 1984).

Wu, Dingjun, Jing Zhang, and Xinmei Huang. (2023). Chain of Thought Prompting Elicits Knowledge Augmentation. In *Findings of the Association for Computational Linguistics: ACL 2023*, 6519–6534, Toronto, Canada. Association for Computational Linguistics.

Yamada, Masaru. (2023). Optimizing Machine Translation through Prompt Engineering: An Investigation into ChatGPT's Customizability. In *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track*, 195–204, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Yamada, Masaru. (2026). Teaching translation with AI: Bridging theory and practice through prompt engineering. In JC Penet, Joss Moorkens & Masaru Yamada (eds.), *Teaching translation in the age of generative AI: New paradigm, new learning?*, 87–104. Berlin: Language Science Press.

Zhang, Biao, Barry Haddow and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. In *International conference on machine learning* (pp. 41092-41110). PMLR.

Appendix A. Details of the source texts included in the corpora for the four experiments

	Advertising (IT)	**Advertising (ES)**	**Biomedical (IT)**	**Biomedical (ES)**
Text ID	**Length**	**Length**	**Length**	**Length**
T1	321	346	192	212
T2	261	318	183	247
T3	162	263	240	164
T4	145	183	209	294
T5	145	272	156	259
T6	321	152	174	156
T7	261	266	201	205
T8	162	198	176	166
T9	145	164	241	197
T10	145	274	226	239
Average length	215,3	243,6	199,8	213,9

Length indicates the length of the corresponding documents in words and *Average length* indicates the average length of the documents in words.

Appendix B. Full example of prompts

P1	Translate from Italian into European Spanish.	Traduce del italiano al español europeo.
P2	You are an expert translator specialized in marketing and advertising translations. Translate the following text from Italian into European Spanish.	Eres un traductor experto especializado en traducción publicitaria y marketing. Traduce el siguiente texto del italiano al español europeo.
P3	Translate from Italian into European Spanish the following text intended for promotional purposes for the official website of *X*, an Italian department store chain. The text is addressed to a general audience and will be published on the company's official website in January 2026. Use a persuasive, elegant, and natural style.	Traduce del italiano al español europeo el siguiente texto con finalidad promocional para la web oficial de *X*, una cadena italiana de grandes almacenes. El texto está destinado a un público general y se publicará en la web oficial de la empresa en enero de 2026. Usa un estilo persuasivo, elegante y natural.

Example of prompts used for the advertising domain in the IT>ES translation direction. The name of the sender is replaced with 'X' for anonymization purposes.

P1	Translate from Spanish into Italian.	Traduci dallo spagnolo all'italiano.
P2	You are an expert translator specialized in marketing and advertising translations. Translate the following text from Spanish into Italian.	Sei un traduttore esperto specializzato in traduzione pubblicitaria e di marketing. Traduci il seguente testo dallo spagnolo all'italiano.
P3	Translate from Spanish into Italian the following text intended for promotional purposes for the official website of *X*, a Spanish footwear brand. The text is addressed to a general audience and will be published on the company's official website in January 2026. Use a persuasive,	Traduci dallo spagnolo all'italiano il seguente testo con finalità promozionale per la pagina web ufficiale di X, un'azienda calzaturiera spagnola. Il testo è destinato a un pubblico generale e verrà pubblicato sulla pagina ufficiale dell'azienda a gennaio 2026. Usa uno stile

	elegant, and natural style.	persuasivo, elegante e naturale.

Example of prompts used for the advertising domain in the ES>IT translation direction. The name of the sender is replaced with 'X' for anonymization purposes.

P1	Translate from Italian into European Spanish.	Traduce del italiano al español europeo.
P2	You are an expert translator specialized in medical translation. Translate the following text from Italian into European Spanish.	Eres un traductor experto especializado en traducción médica. Traduce el siguiente texto del italiano al español europeo.
P3	Translate from Italian into European Spanish the following text intended for informative purposes for the official website of X, a diagnostic and healthcare center based in Italy. The text is addressed to a general audience and will be published on the center's official website in January 2026. Use a professional and clear style.	Traduce del italiano al español europeo el siguiente texto con finalidad informativa para la web oficial de *X*, un centro de diagnóstico y atención sanitaria con sede en Italia. El texto está dirigido a un público general y se publicará en la web oficial del centro en enero de 2026. Usa un estilo profesional y claro.

Example of prompts used for the biomedical domain in the IT>ES translation direction. The name of the sender is replaced with 'X' for anonymization purposes.

P1	Translate from Spanish into Italian.	Traduci dallo spagnolo all'italiano.
P2	You are an expert translator specialized in medical translation. Translate the following text from Spanish into Italian.	Sei un traduttore esperto specializzato in traduzione medica. Traduci il seguente testo dallo spagnolo all'italiano.
P3	Translate from Spanish into Italian the following text intended for informative purposes for the official website of *X*, an ophthalmology clinic based in Spain. The text is addressed to a general audience and will be published on the clinic's official website in January 2026. Use a professional and clear style.	Traduci dallo spagnolo all'italiano il seguente testo con finalità informativa per il sito ufficiale di *X*, una clinica oftalmologica con sede in Spagna. Il testo è destinato a un pubblico generale e verrà pubblicato sulla pagina ufficiale della clinica a gennaio 2026. Usa uno stile professionale e chiaro.

Example of prompts used for the biomedical domain in the ES>IT translation direction. The name of the sender is replaced with 'X' for anonymization purposes.

Appendix 3. Levenshtein scores

Lev raw is the raw Levenshtein distance. *Lev norm* is the normalized Levenshtein distance, calculated by dividing the raw Levenshtein distance by the length of the longer string, thereby making comparisons across texts of different lengths more meaningful. The resulting value of the *Lev norm* was then multiplied by 100 to express it as a percentage and facilitate interpretation. So, the closer the value is to 0, the more similar the two texts are; conversely, the higher the value, the greater the difference between them.

	P1		**P2**		**P3**	
	Lev raw	Lev norm	Lev raw	Lev norm	Lev raw	Lev norm
T1	9	0,4447	92	4,4725	49	2,3671
T2	59	3,3333	41	2,3335	94	5,3318
T3	2	0,1919	21	2,029	73	6,6728
T4	7	0,7551	38	4,1304	109	11,066
T5	61	6,0516	28	2,7264	154	15,509
T6	6	0,4882	23	1,8608	26	2,057
T7	31	2,227	24	1,7217	47	3,3764

T8	42	2,4764	0	0	64	3,7647
T9	20	1,0689	15	0,803	92	4,9356
T10	67	4,2839	57	3,6562	117	7,4144
MEAN	**30,4**	2,1321	**33,9**	2,3734	**82,5**	6,2494

Levenshtein scores in the **advertising domain (IT>ES)**

	P1		**P2**		**P3**	
	Lev raw	Lev norm	Lev raw	Lev norm	Lev raw	Lev norm
T1	65	3,0603	119	5,5272	95	4,4351
T2	48	2,444	74	3,7563	171	8,6233
T3	22	1,2035	41	2,2356	190	9,9372
T4	19	1,7415	98	9,0074	81	7,2581
T5	67	4,1333	22	1,3995	117	7,2401
T6	0	0	12	1,3086	60	6,4516
T7	49	3,495	40	2,8349	131	8,9849
T8	26	2,0553	92	7,2441	98	7,5038
T9	8	0,8611	11	1,1579	153	15,0442
T10	19	1,0759	48	2,7119	274	15,1214
MEAN	**32,3**	**2,00699**	**55,7**	**3,71834**	**137**	**9,05997**

Levenshtein scores in the **advertising domain (ES>IT)**

	P1		**P2**		**P3**	
	Lev raw	Lev norm	Lev raw	Lev norm	Lev raw	Lev norm
T1	32	2,2315	55	3,7723	95	6,4494
T2	17	1,3418	102	8,1275	104	8,2084
T3	36	2,2018	75	4,5732	42	2,5501
T4	34	2,0286	31	1,8289	124	6,9781
T5	27	2,439	34	3,0631	62	5,4965
T6	51	3,5941	32	2,2535	70	4,902
T7	18	1,1984	47	3,1271	8	0,5295
T8	110	9,0759	117	9,5355	67	5,3175
T9	17	1,0372	77	4,6053	106	6,113
T10	59	3,5056	21	1,2419	155	9,2152
MEAN	**40,1**	**2,86539**	**59,1**	**4,21283**	**83,3**	**5,57597**

Levenshtein scores in the **biomedical domain (IT>ES)**

	P1		**P2**		**P3**	
	Lev raw	Lev norm	Lev raw	Lev norm	Lev raw	Lev norm
T1	20	1,5432	49	3,6704	124	9,1445
T2	35	2,0302	129	7,5439	209	12,2365
T3	25	2,4704	92	8,7039	152	14,3939
T4	171	9,3239	170	9,0763	376	19,0669
T5	31	1,6129	63	3,2525	293	14,9185
T6	12	1,1364	46	4,3685	62	5,709
T7	54	3,5644	105	6,6582	240	14,0762
T8	21	1,5885	40	3,028	126	9,2375

T9	97	7,7045	63	5,0521	125	9,9681
T10	46	2,8949	76	4,7769	297	17,4809
MEAN	**51,2**	**3,38693**	**83,3**	**5,61307**	**200,4**	**12,6232**

Levenshtein scores in the **biomedical domain (ES>IT).**