Please Translate Again: Two Simple Experiments on Whether Human-Like Reasoning Helps Translation

Anonymous ACL submission

Abstract

Large Language Models (LLMs) demonstrate strong reasoning capabilities for many tasks, often by explicitly decomposing the task via Chain-of-Thought (CoT) reasoning. Recent work on LLM-based translation designs hand-crafted prompts to decompose translation, or trains models to incorporate intermediate steps. *Translating Step-by-step* (Briakou et al., 2024), for instance, introduces a multi-step prompt with decomposition and refinement of translation with LLMs, which achieved state-of-the-art results on WMT24. In this work, we scrutinise this strategy's effectiveness. Empirically, we find no clear evidence that performance gains stem from explicitly decomposing the translation process, at least for the models on test; and we show that simply prompting LLMs to "translate again" yields even better results than human-like step-by-step prompting. Our analysis does not rule out the role of reasoning, but instead invites future work exploring the factors for CoT's effectiveness in the context of translation.

1 Introduction

Large Language Models (LLMs) exhibit strong reasoning capabilities, often characterized by a lengthy, step-by-step decomposition of the question prior to generating the answer—known as Chain-of-Thought (CoT) (Wei et al., 2022)—along with possible attempts and revisions of the answer, referred to as self-refinement (Madaan et al., 2023; Chen et al., 2023; Pan et al., 2024). Both processes resemble human behaviour when tackling complex problems such as coding and mathematics, and have been shown to be effective for such tasks.

Driven by recent advancements in LLMs' reasoning capabilities, a trend has developed in improving translation quality through a human-like *decomposition–translation–refinement* paradigm. Here, the source text is decomposed into different aspects including meanings, topics, idiomatic expressions etc., followed by translation drafting and refinement based on these aspects, before generating the final translation.

Some recent works explore pre-translation decomposition, focusing on keywords (He et al., 2024) or idiomatic expressions (Li et al., 2024), aided by external resources. Others address post-translation refinement, guided by external translation quality assessment (Huang et al., 2024; Ki and Carpuat, 2024) or explicit self-evaluation (Feng et al., 2025). Refinement can be applied iteratively (Chen et al., 2023; Xu et al., 2024), and is particularly effective for long document-level translation (Wu et al., 2024). A key work by Briakou et al. (2024) combines pre- and post-translation processes via a fixed 4-step prompting strategy—decomposition (or research), drafting, refinement, and proofreading. Their method shows progressive improvements for long-form translation, achieving state-of-the-art results on WMT24.

While these studies show performance gains over direct translation in some settings, the generalizability of human-like multi-pass prompting across models and input types remains unclear. Further, most lack an explicit examination or quantitative analysis of the underlying mechanisms behind these gains. To address these points, we design two simple experiments comparing against the current best practice (Briakou et al., 2024) to answer:

- Does decomposition positively impact translation quality, across models and input types?
- How faithful are translations to their decomposition, and does faithfulness lead to improved translation quality?

[0] We plan to release the 223k segment and document translations in 8 language pairs from 3 models across 4 steps at `[anonymised]` to enable further research into the effects of decomposition on translation quality.

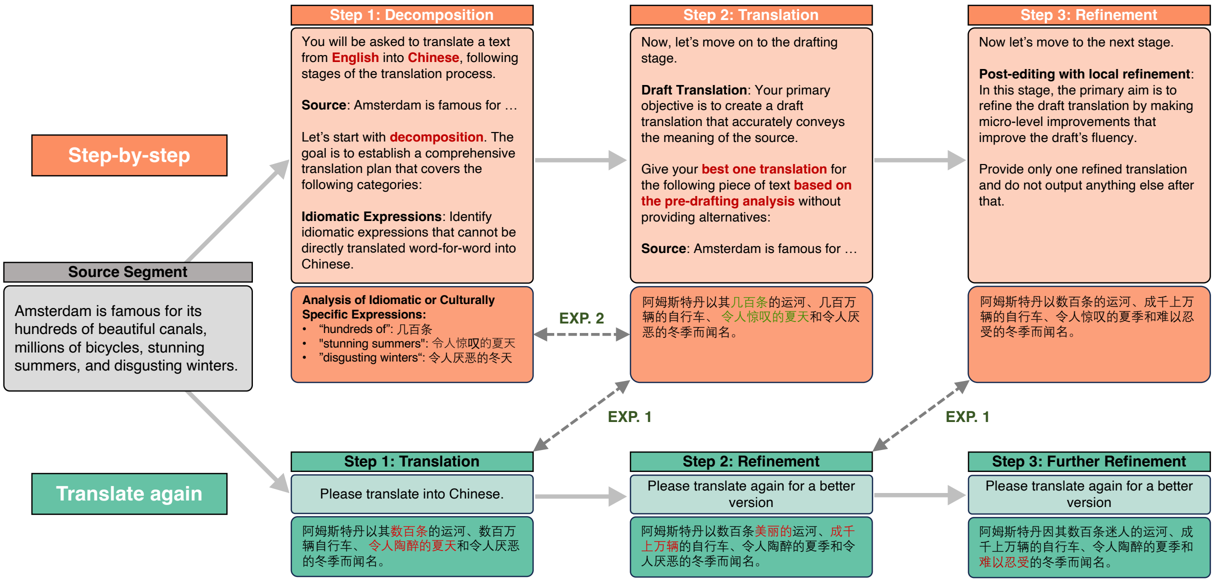


Figure 1: Schematic of prompting frameworks for Step-by-step translation with decomposition (above) and multi-pass Translate again (our method, below), with user prompts and model outputs shown for each step. Experiment 1 (EXP. 1) compares translation and refinement outcomes with and without the decomposition step across metrics, input types, and models. Experiment 2 (EXP. 2) traces back evidence to assess whether accurately following decomposition improves translation. Full prompts for both settings are provided in Appendix C.

Our two simple experiments find that: (1) most gains come from self-refinement, while decomposition has limited—and sometimes negative—effects, depending largely on the LLM and input type; and (2) decomposition clearly influences translation behaviour, but strict faithfulness to the decomposition does not necessarily improve translation quality.

Given the findings, we encourage the research community to evaluate alternative explanations and reconsider the necessity of human-like decomposition when engaging the reasoning capabilities of LLMs for translation. At a minimum, future studies should consider incorporating a CoT-free refinement strategy—such as the simple 'please translate again' prompt used here—as a baseline, given its demonstrated effectiveness and efficiency.

2 Translation Decomposition and Refinement

Translation by human translators is commonly divided into three phases: pre-drafting, drafting, and post-drafting (Mossop, 2013). First, the translator familiarises themselves with the source, consisting of comprehension and planning; next, a full draft translation is written, optionally with the help of external resources; then the translator reviews and revises the draft translation. Briakou et al. (2024) partially replicate this process with their 4-step prompting process, splitting the final step into refinement and proofreading.

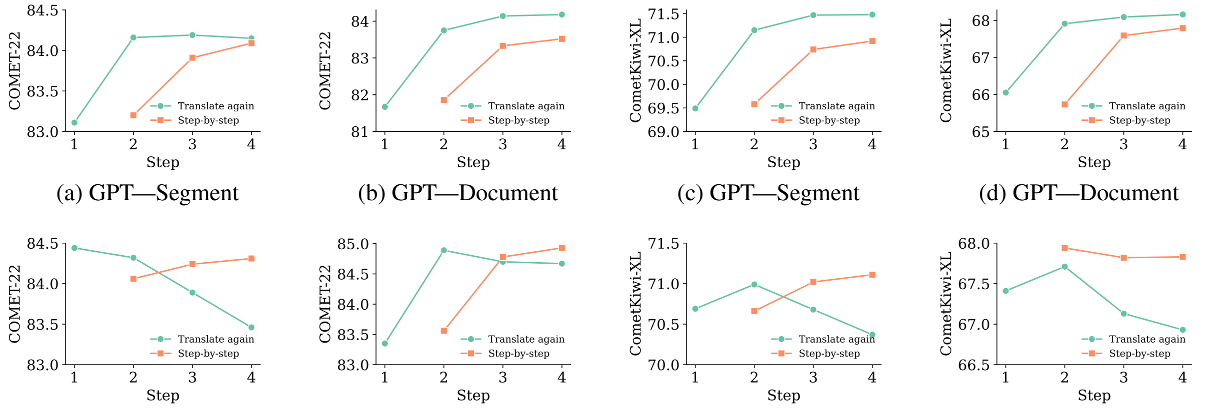
Formally, given a language model p_θ and a source text x to be translated, the output can be viewed as a sample $O \sim p_\theta(\cdot \mid I(x))$, where $I(x)$ is a prompt that may include x as a component. Multi-step prompting for translation is a sequential process in which the outputs of previous steps are fed into the next prompt. For instance, the *decomposition–translation–refinement* workflow (Figure 1, top) can be formalized as:

$$\begin{aligned} O_d &\sim p_\theta(\cdot \mid I_d(x)), \\ O_t &\sim p_\theta(\cdot \mid I_d(x), O_d, I_t(x)), \\ O_f &\sim p_\theta(\cdot \mid I_d(x), O_d, I_t(x), O_t, I_f(x)). \end{aligned}$$

Here, I_d, I_t, and I_f denote the prompts for the decomposition, translation, and refinement steps, respectively, and O_d, O_t, and O_f are their corresponding outputs. This study investigates the impact of human-like decomposition (O_d) on translation quality (O_t) and final refinement output (O_f) under varying conditions: (i) model differences (θ), (ii) sentence- vs. document-level source inputs x, and (iii) the presence or absence of O_d (see Section 4). We also explicitly verify whether faithfully following the decomposition generally leads to improved translation (see Section 5).

3 Experimental Setup

Models. We use `GPT-4o-mini` (OpenAI, 2024) and `Gemini-2.0-Flash` (Google, 2024), as perfor-



(a) GPT—Segment (b) GPT—Document (c) GPT—Segment (d) GPT—Document

(e) Gemini—Segment (f) Gemini—Document (g) Gemini—Segment (h) Gemini—Document

Figure 2: *Step-by-step* vs. *Translate again* results in COMET-22 and CometKiwi-XL for GPT-4o-mini (top) and Gemini-2.0-Flash (bottom), for segment and document-level translation. Note Steps 2–4 iteratively call the LLM to translate again, and Step 4 is a proofreading step; see Fig. 1 for an illustration and Appendix C for full prompts.

mant and cost-effective closed-source LLMs that demonstrate strong reasoning capabilities.

Data. We use the WMT24++ (Deutsch et al., 2025) dataset as our test set, because a) the dataset was released later than the LLMs we used here, ensuring no data leakage issues, and b) the translation references in WMT24++ are human-written and subsequently post-edited by professional translators, ensuring the highest possible data quality. In this study, we use the post-edited version. We select 8 language pairs (`en→cs,de,fr,he,ja,ru,uk,zh`) to cover varying writing scripts and families. Each direction shares the same 960 English source samples. For document-level tests, we combine the segments based on meta-data to give 221 documents (limited to 150 space-separated tokens).

Evaluation. Following best practice (Kocmi et al., 2024), we use both reference-based and reference-free neural metrics, using COMET$_{22}^{DA}$ (Rei et al., 2022) and COMETKiwi-XL$_{23}^{DA}$ (Rei et al., 2023), respectively. We also report results in XCOMET-XL (Guerreiro et al., 2024) and MetricX-23-XL (Juraska et al., 2023) in Appendix A.

Baselines. We replicate the step-by-step prompt introduced by Briakou et al. (2024) as our baseline, and report prompts in full in Appendix C. Briakou et al. (2024) focus primarily on long-form text using Gemini, whereas we conduct comprehensive experiments on both short- and long-form text and demonstrate generalizability across LLMs.

Proposed method. We introduce a maximally simple multi-pass prompting method in which the model is asked to produce a translation, then asked to translate again 3 more times, given the conversation history, mirroring the step-by-step prompt above. This method involves no explicit pre-drafting step, but expands the number of post-drafting steps arbitrarily, see Figure 1 (bottom).

4 Experiment 1: Decomposition's Impact on Translation

We investigate the effect of decomposition on translation by testing the baseline method (*Step-by-step*) against our simple multi-pass prompting method (*Translate again*). Figure 2 presents mean stepwise results; see App. Figure 5 for detailed results across languages. Our findings are as follows:

Decomposition. Comparing Step 2 against Step 1 shows, at best, a marginally positive effect at the document-level, particularly with Gemini (cf. (f), (h)). This suggests decomposition is not a generally effective strategy for LLM-based translation.

Self-refinement. Results after a single step of self-refinement show that simply prompting the model to *translate again* for a better version (Step 2) *without* decomposition consistently yields improvements for GPT-4o-mini over Step-by-step prompting *with* a pre-drafting step (Step 3).

Successive refinement. Additional steps of refinement, Steps 2–3–4, produce only marginal improvements, or occasional degradation for Gemini. We attribute this to the strong performance achieved after 1 refinement step, which may already maximise the LLMs' parametric knowledge and therefore leaves little space for further gains.

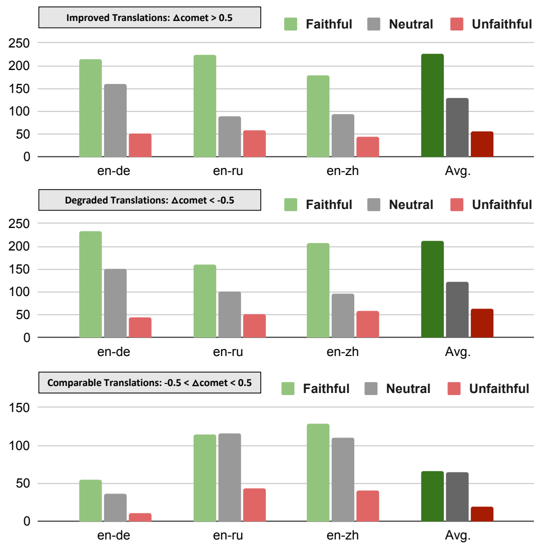


Figure 3: Counts of translations by GPT-4o-mini that are faithful, neutral, or unfaithful to the decomposition, compared to direct translation. *Avg.* denotes the average over all 8 language directions. Full results are provided in Appendix Figure 8.

Segment vs. Document-level. Generally, we observe that refinements result in larger point improvements for document-level translation over segment-level, aligning with Briakou et al. (2024).

We therefore find no compelling evidence in favour of human-like decomposition for translation, compared to direct multi-pass prompting. Unlike symbolic tasks (Sprague et al., 2025), translation only benefits weakly, if at all, from CoT prompting.

5 Experiment 2: Attribution Analysis of Decomposition

We explicitly verify via an attribution analysis whether the decomposition step substantially influences translation behaviour in the subsequent step. We also analyse whether faithfulness to the decomposition results in improved translations.

Explicit verification. Formally, for a source sentence s_i, we construct a four-tuple (s_i, d_i, t_i^1, t_i^2) by prompting an LLM (1) with decomposition d_i, resulting in t_i^1 (Step 2), and (2) without decomposition, resulting in t_i^2 (Step 1). Explicit verification with an LLM-as-a-judge proceeds as follows:

- **Differentiation:** LLM annotators are asked to identify the main differences, $\{v_1, v_2, ..., v_k\}$, between translations t_i^1 and t_i^2.
- **Attribution:** LLM annotators are asked how many of the differences for t_i^1 and t_i^2 can be attributed to the decomposition step d_i, giving trace-back counts c_i^1 and c_i^2 respectively; n.b. t_i^2 is generated without d_i thus serves as a baseline.
- **Assessment:** We measure the influence of decomposition d_i on translation t_i^1 by comparing c_i^1 and c_i^2, where $c_i^1 > c_i^2$ indicates a translation which is *faithful* to the decomposition; $c_i^1 = c_i^2$ indicates a *neutral* translation which is neither faithful nor unfaithful; and $c_i^1 < c_i^2$ indicates an *unfaithful* translation.

We categorise all WMT24-derived four-tuples into *Improved*, *Comparable*, and *Degraded Translation* based on the COMET scores of t^1 vs. t^2. For each group, we conduct explicit verification using `GPT-4o` as a judge. Figure 3 shows verification results across groups and directions. We find that:

Decomposition influences translation. Across all categories and languages, translations conditioned on decompositions contain substantially more differences that can be clearly attributed to the decomposition context (Faithful vs. Unfaithful), compared to direct translations.

Faithfulness does not improve translation. Degraded translations show a comparable number of translations influenced by the context, compared to that of improved translations, suggesting the overall effect of decomposition is neutral.

Our analysis shows that while decomposition clearly influences translation behaviour, the *positive* impact of decomposition on translation is minimal. We conservatively attribute this to the fact that, alongside useful information, the decomposition step may contain errors, which can propagate to the downstream translation task.

6 Conclusion

Our results suggest a divergence between the optimal translation strategies for humans and LLMs: while human translators benefit from decomposing the task, LLMs may rely on a different form of reasoning, and imposing human biases may lead to suboptimal outcomes. Further, faithfulness to a generated decomposition does not always yield positive effects. In fact, our maximally simple prompt, *please translate again*, achieves performance comparable to, or even exceeding, the state-of-the-art multi-pass prompting method (Briakou et al., 2024). This corroborates findings, from the related task of translation from a grammar book (Aycock et al., 2025), that for translation, LLMs exhibit different reasoning tendencies to humans.

7 Limitations

Due to limited resources, our investigation primarily focuses on two state-of-the-art LLM families: GPT-4o and Gemini. For Experiment 2, while we observe that GPT-4o is a competent judge in our explicit verification experiments, we note that incorporating judgments from different model families would strengthen the reliability of our results.

References

Seth Aycock, David Stap, Di Wu, Christof Monz, and Khalil Sima'an. 2025. Can LLMs Really Learn to Translate a Low-Resource Language from One Grammar Book? In *The Thirteenth International Conference on Learning Representations*.

Eleftheria Briakou, Jiaming Luo, Colin Cherry, and Markus Freitag. 2024. Translating step-by-step: Decomposing the translation process for improved translation quality of long-form texts. *arXiv preprint arXiv:2409.06790*.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 2023. Teaching large language models to self-debug. *arXiv preprint arXiv:2304.05128*.

Daniel Deutsch, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, and 1 others. 2025. Wmt24++: Expanding the language coverage of wmt24 to 55 languages & dialects. *arXiv preprint arXiv:2502.12404*.

Zhaopeng Feng, Yan Zhang, Hao Li, Bei Wu, Jiayu Liao, Wenqiang Liu, Jun Lang, Yang Feng, Jian Wu, and Zuozhu Liu. 2025. TEaR: Improving LLM-based machine translation with systematic self-refinement. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3922–3938, Albuquerque, New Mexico. Association for Computational Linguistics.

Google. 2024. Introducing Gemini 2.0: Our new AI model for the agentic era.

Nuno M Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André FT Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*, 12:229–246.

Yichong Huang, Baohang Li, Xiaocheng Feng, Wenshuai Huo, Chengpeng Fu, Ting Liu, and Bing Qin. 2024. Aligning translation-specific understanding to general understanding in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5028–5041, Miami, Florida, USA. Association for Computational Linguistics.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

Dayeon Ki and Marine Carpuat. 2024. Guiding large language models to post-edit machine translation with error annotations. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4253–4273, Mexico City, Mexico. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Shuang Li, Jiangjie Chen, Siyu Yuan, Xinyi Wu, Hao Yang, Shimin Tao, and Yanghua Xiao. 2024. Translate meanings, not just words: Idiomkb's role in optimizing idiomatic translation with language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18554–18563.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.

Brian Mossop. 2013. *Revising and Editing for Translators*, 3 edition. Routledge, London.

OpenAI. 2024. GPT-4o mini: Advancing cost-efficient intelligence.

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2024. Automatically correcting large language models: Surveying the landscape of diverse automated correction strategies. *Transactions of the Association for Computational Linguistics*, 12:484–506.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, JosÃ© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.

Zayne Rea Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2025. To CoT or not to CoT? Chain-of-thought helps mainly on math and symbolic reasoning. In *The Thirteenth International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Minghao Wu, Yulin Yuan, Gholamreza Haffari, and Longyue Wang. 2024. (perhaps) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts. *arXiv preprint arXiv:2405.11804*.

Wenda Xu, Daniel Deutsch, Mara Finkelstein, Juraj Juraska, Biao Zhang, Zhongtao Liu, William Yang Wang, Lei Li, and Markus Freitag. 2024. LLMRefine: Pinpointing and refining large language models via fine-grained actionable feedback. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1429–1445, Mexico City, Mexico. Association for Computational Linguistics.

A Full Results for Experiment 1

In this section, we provide all supplementary results for Experiment 1 (Section 4).

Tables 1–8 show translation results across languages at the segment and document-level, for COMET-22, CometKiwi-23-XL, MetricX-23-XL, and XCOMET-XL.

Figure 4 presents the mean results across languages for GPT-4o-mini and Gemini-2.0-Flash in MetricX and XCOMET-XL under both *step-by-step* and *translate again* prompting strategies.

Figure 5 presents the results of zero-shot (direct) translation and subsequent refinement under both the *Step-by-step* (Step 3) and *Translate again* (Step 2) strategies on GPT-4o-mini and Gemini-2.0-Flash. We observe across languages and metrics that: 1) Refinement consistently improves performance over direct translation for both strategies; 2) The *translate again* strategy generally outperforms the *step-by-step* strategy.

Figures 6 and 7 show COMET score trajectories for GPT-4o-mini at the segment- and document-level respectively. An increase in the y-axis represents a *relative* increase in COMET score compared to the *previous* step, while a downwards trajectory indicates a *relative* decrease in COMET score. We observe that *translate again* prompting increases many scores from step 1–2, and many documents benefit further from step 2–3. *Step-by-step* shows most segments and documents improve from step 2–3; n.b. for *Step-by-step* we discount steps 1–2 as no translation is produced at step 1. At the segment level, trajectories from step 3–4 are somewhat equally split, while at the document-level most trajectories see further relative improvements.

B Full Results for Experiment 2

We provide full verification results across groups and all eight language directions for Experiment 2 (Section 5); see Figure 8.

C All Prompt Templates

C.1 *Step-by-Step* Prompts

The templates for *step-by-step* prompting comprise the Decomposition stage (Figure 10), the Translation stage (Figure 11), the Refinement stage (Figure 12), and the Proofreading stage (Figure 13).

C.2 *Translate Again* Prompts

The templates for *translate again* prompting include the Translation stage (Figure 14) and the Refinement stage (Figure 15). The refinement prompt can be applied iteratively within a session to perform multiple steps of refinement.

C.3 *LLM-as-a-Judge* Prompts

Figure 16 provides the prompt used for LLM-as-a-Judge in Experiment 2 (Section 5). We also showcase the output of our *LLM-as-a-judge* in Figure 9, illustrating how it operates.

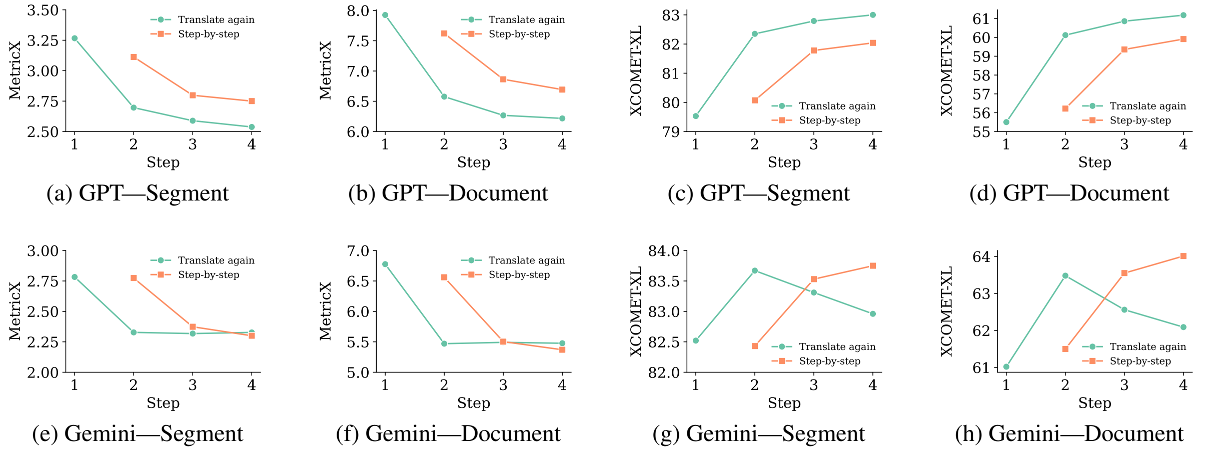


Figure 4: *Step-by-step* vs. *Translate again* results in MetricX and XCOMET-XL for GPT-4o-mini (top) and Gemini-2.0-Flash (bottom), for segment and document-level translation. For MetricX, lower scores indicate a higher translation quality. See Fig. 1 for an illustration and Appendix C for full prompts.

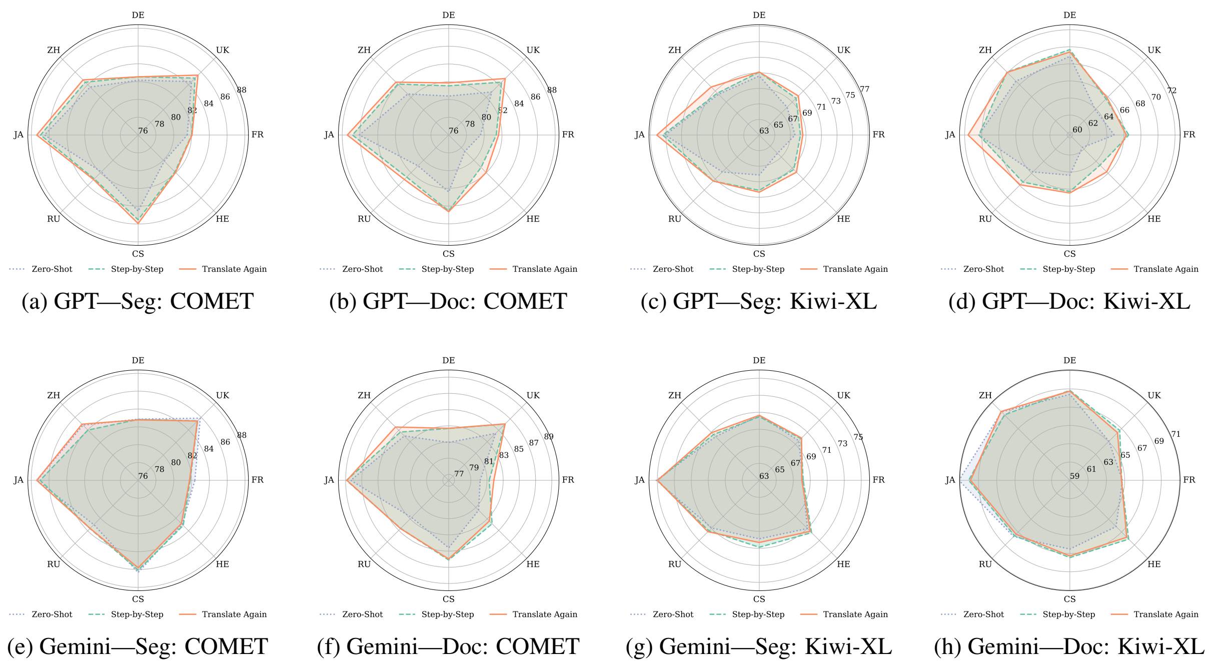


Figure 5: COMET-22 and CometKiwi-XL results per-language for *Step-by-step* (Step 3), *Translate again* (Step 2), and Zero-Shot (Step 1) prompts, for GPT-4o-mini (top) and Gemini-2.0-Flash (bottom), for segment and document-level translation.

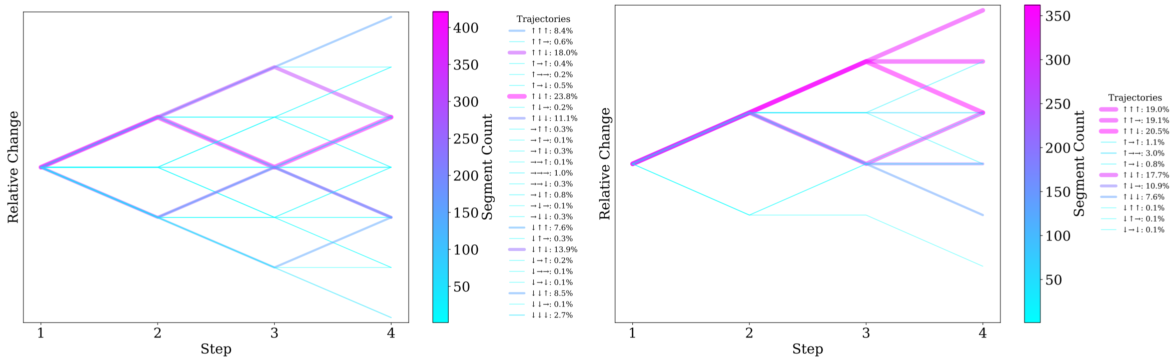


Figure 6: Segment-level COMET score trajectories for GPT-4o-mini with Translate again (left) and Step-by-step (right) prompting strategies. An increase or decrease in the y-axis indicates a *relative* COMET score improvement or degradation compared to the previous step, respectively. Trajectory proportions are shown in the legend.

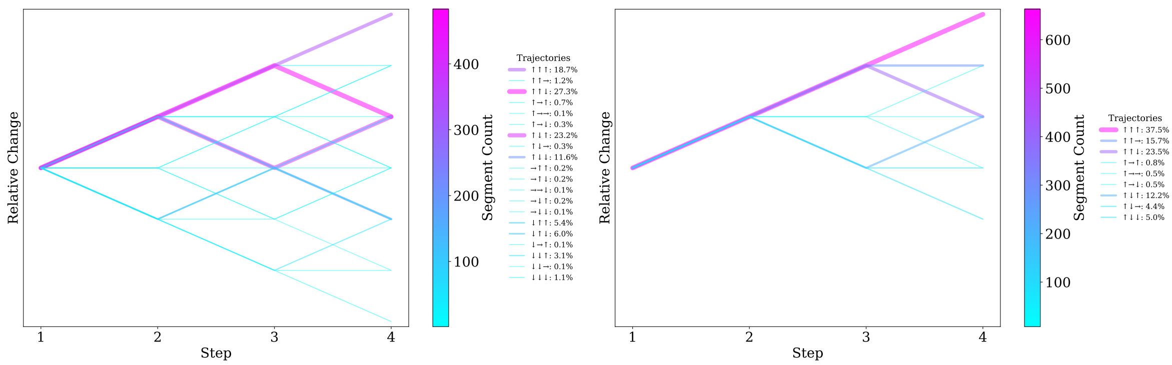


Figure 7: Document-level COMET score trajectories for GPT-4o-mini with Translate again (left) and Step-by-step (right) prompting strategies. An increase or decrease in the y-axis indicates a *relative* COMET score improvement or degradation compared to the previous step, respectively. Trajectory proportions are shown in the legend.

Model	Setup	Step	en→cs	en→de	en→fr	en→he	en→ja	en→ru	en→uk	en→zh	Avg.
GPT-4o-mini	Step-by-step	1	–	–	–	–	–	–	–	–	–
		2	84.82	82.19	81.28	80.72	86.33	82.06	84.35	83.83	83.20
		3	85.52	82.54	82.02	81.85	86.97	82.97	85.05	84.40	83.91
		4	85.77	82.74	82.17	82.16	87.06	83.01	85.29	84.48	84.09
	Translate again	1	84.55	82.16	81.52	80.15	86.53	81.85	84.49	83.62	83.11
		2	85.94	82.54	82.06	81.95	87.40	83.08	85.52	84.76	84.16
		3	86.02	82.66	81.67	82.32	87.27	83.30	85.51	84.77	84.19
		4	85.82	82.35	81.73	82.55	87.30	83.23	85.50	84.73	84.15
Gemini-2.0-Flash	Step-by-step	1	–	–	–	–	–	–	–	–	–
		2	85.92	82.74	81.92	82.86	86.83	82.95	85.61	83.68	84.06
		3	86.08	82.83	81.84	83.16	87.00	83.64	85.44	83.97	84.25
		4	86.14	82.89	81.85	83.25	87.10	83.73	85.53	84.03	84.32
	Translate again	1	86.34	82.86	82.37	83.00	87.34	83.03	85.86	84.76	84.44
		2	85.81	82.78	81.77	82.86	87.38	83.63	85.41	84.91	84.32
		3	85.39	82.20	80.96	82.61	87.04	83.13	85.14	84.63	83.89
		4	84.87	81.69	80.59	82.34	86.83	82.63	84.75	84.01	83.46

Table 1: Full COMET-22 results for segment-level translation with GPT-4o-mini and Gemini-2.0-Flash, across 8 language pairs. Step-by-step Step 1 results are not shown since the model does not generate a translation at this step. A darker green shade indicates a better score.

Model	Setup	Step	en→cs	en→de	en→fr	en→he	en→ja	en→ru	en→uk	en→zh	Avg.
GPT-4o-mini	Step-by-step	1	–	–	–	–	–	–	–	–	–
		2	68.76	70.51	67.44	67.12	74.29	70.27	68.46	69.78	69.58
		3	70.11	71.09	68.30	69.28	75.42	71.37	69.69	70.62	70.74
		4	70.46	71.17	68.41	69.81	75.40	71.55	69.71	70.87	70.92
	Translate again	1	68.14	70.59	67.60	66.32	75.10	69.72	68.06	70.41	69.49
		2	70.35	71.11	68.60	69.76	76.19	71.37	70.11	71.74	71.15
		3	70.70	71.68	68.97	70.41	76.17	71.66	70.10	72.03	71.47
		4	70.77	71.32	69.06	70.54	76.35	71.75	70.11	71.93	71.48
Gemini-2.0-Flash	Step-by-step	1	–	–	–	–	–	–	–	–	–
		2	69.82	70.76	67.94	70.78	74.81	71.26	69.80	70.07	70.65
		3	70.87	70.48	68.17	71.70	74.98	71.39	69.94	70.62	71.02
		4	71.00	70.53	68.30	71.76	74.91	71.52	70.12	70.74	71.11
	Translate again	1	69.88	70.63	68.12	71.01	75.12	70.93	69.58	70.27	70.69
		2	70.31	70.65	68.01	71.48	74.98	71.55	70.04	70.93	70.99
		3	70.47	70.55	67.72	71.35	74.39	70.70	69.78	70.46	70.68
		4	69.96	69.99	67.24	71.46	74.45	70.28	69.39	70.18	70.37

Table 2: Full CometKiwi-XL results for segment-level translation with GPT-4o-mini and Gemini-2.0-Flash, across 8 language pairs. Step-by-step Step 1 results are not shown since the model does not generate a translation at this step. A darker green shade indicates a better score.

Model	Setup	Step	en→cs	en→de	en→fr	en→he	en→ja	en→ru	en→uk	en→zh	Avg.
GPT-4o-mini	Step-by-step	1	–	–	–	–	–	–	–	–	–
		2	2.97	1.45	2.80	4.66	2.82	3.46	3.49	3.25	3.11
		3	2.68	1.35	2.42	4.12	2.57	3.14	3.08	3.02	2.80
		4	2.64	1.33	2.38	3.94	2.55	3.10	3.04	3.02	2.75
	Translate again	1	3.29	1.53	2.80	5.05	2.85	3.72	3.53	3.36	3.27
		2	2.65	1.33	2.21	4.14	2.45	3.08	2.87	2.84	2.70
		3	2.46	1.27	2.13	3.92	2.36	3.00	2.79	2.79	2.59
		4	2.42	1.26	2.14	3.74	2.31	2.95	2.73	2.74	2.54
Gemini-2.0-Flash	Step-by-step	1	–	–	–	–	–	–	–	–	–
		2	2.72	1.42	2.61	3.67	2.48	3.07	2.91	3.30	2.77
		3	2.18	1.27	2.24	3.04	2.19	2.61	2.56	2.90	2.37
		4	2.15	1.24	2.18	2.93	2.10	2.53	2.48	2.79	2.30
	Translate again	1	2.69	1.41	2.63	3.74	2.41	3.20	2.99	3.21	2.78
		2	2.25	1.28	2.19	2.98	2.05	2.58	2.59	2.70	2.33
		3	2.13	1.28	2.20	2.96	2.03	2.65	2.59	2.70	2.32
		4	2.22	1.27	2.20	2.99	2.05	2.59	2.52	2.77	2.33

Table 3: Full MetricX results for segment-level translation with GPT-4o-mini and Gemini-2.0-Flash, across 8 language pairs. Step-by-step Step 1 results are not shown since the model does not generate a translation at this step. A lower score and a darker green shade indicates better translation quality.

Model	Setup	Step	en→cs	en→de	en→fr	en→he	en→ja	en→ru	en→uk	en→zh	Avg.
GPT-4o-mini	Step-by-step	1	–	–	–	–	–	–	–	–	–
		2	81.10	90.40	81.70	73.60	77.52	81.25	79.63	75.37	80.07
		3	82.86	91.06	83.24	76.60	78.96	82.93	81.37	77.19	81.78
		4	83.28	91.20	83.35	77.33	79.06	82.94	81.75	77.39	82.04
	Translate again	1	80.46	90.51	81.31	71.41	77.61	80.30	79.22	75.39	79.53
		2	83.34	91.17	83.46	76.34	80.53	83.13	82.54	78.30	82.35
		3	83.66	91.68	83.66	77.62	80.98	83.48	82.58	78.68	82.79
		4	83.98	91.55	83.78	78.15	81.12	83.73	82.77	78.90	83.00
Gemini-2.0-Flash	Step-by-step	1	–	–	–	–	–	–	–	–	–
		2	83.51	91.09	82.73	79.21	80.08	83.46	82.74	76.64	82.43
		3	84.55	91.45	83.25	81.25	81.53	84.48	83.08	78.68	83.53
		4	84.80	91.49	83.42	81.62	81.63	84.74	83.19	79.13	83.75
	Translate again	1	83.43	90.99	83.14	79.40	80.38	82.91	82.78	77.16	82.52
		2	84.21	91.43	83.39	81.22	82.11	84.42	82.95	79.65	83.67
		3	84.36	91.28	82.54	80.98	81.36	84.03	82.67	79.26	83.31
		4	83.62	91.10	82.54	80.60	81.41	83.65	82.14	78.62	82.96

Table 4: Full XCOMET-XL results for segment-level translation with GPT-4o-mini and Gemini-2.0-Flash, across 8 language pairs. Step-by-step Step 1 results are not shown since the model does not generate a translation at this step. A darker green shade indicates a better score.

Model	Setup	Step	en→cs	en→de	en→fr	en→he	en→ja	en→ru	en→uk	en→zh	Avg.
GPT-4o-mini	Step-by-step	1	–	–	–	–	–	–	–	–	–
		2	83.14	80.41	80.49	78.15	85.92	81.32	83.12	82.31	81.86
		3	84.54	81.54	81.36	81.16	86.85	82.70	84.42	84.11	83.33
		4	84.70	81.65	81.52	81.61	86.56	83.13	84.97	84.02	83.52
	Translate again	1	82.41	80.39	79.59	78.58	86.08	80.90	82.86	82.54	81.67
		2	84.64	81.88	81.62	81.96	87.43	83.08	85.01	84.42	83.76
		3	85.08	82.13	81.92	82.34	87.77	83.67	85.36	84.86	84.14
		4	85.24	81.99	81.92	82.53	87.86	83.46	85.62	84.89	84.19
Gemini-2.0-Flash	Step-by-step	1	–	–	–	–	–	–	–	–	–
		2	84.98	81.35	80.59	82.10	87.83	83.05	85.24	83.35	83.56
		3	86.10	82.67	81.27	83.85	88.89	84.66	86.11	84.70	84.78
		4	86.21	82.79	81.47	83.99	88.80	84.84	86.19	85.17	84.93
	Translate again	1	84.60	80.95	80.15	81.49	88.36	82.75	84.47	84.02	83.35
		2	85.97	82.69	81.84	83.37	88.84	84.69	86.12	85.57	84.89
		3	85.94	82.73	81.44	83.28	88.92	84.62	85.55	85.16	84.70
		4	85.67	82.58	81.36	83.73	88.80	84.53	85.56	85.14	84.67

Table 5: Full COMET-22 results for document-level translation with GPT-4o-mini and Gemini-2.0-Flash, across 8 language pairs. Step-by-step Step 1 results are not shown since the model does not generate a translation at this step. A darker green shade indicates a better score.

Model	Setup	Step	en→cs	en→de	en→fr	en→he	en→ja	en→ru	en→uk	en→zh	Avg.
GPT-4o-mini	Step-by-step	1	–	–	–	–	–	–	–	–	–
		2	64.50	68.71	65.81	60.03	69.34	65.69	63.90	67.87	65.73
		3	66.28	69.68	66.57	64.66	70.33	67.46	65.66	70.09	67.59
		4	66.47	69.77	66.57	65.60	69.91	67.83	66.22	69.95	67.79
	Translate again	1	64.37	68.89	64.87	61.73	70.35	65.78	63.87	68.60	66.06
		2	66.46	69.40	66.28	65.77	71.60	67.91	65.84	70.05	67.91
		3	66.83	69.10	66.29	66.41	71.74	67.92	66.17	70.30	68.09
		4	67.09	69.26	65.97	66.68	71.73	67.80	66.32	70.41	68.16
Gemini-2.0-Flash	Step-by-step	1	–	–	–	–	–	–	–	–	–
		2	66.95	69.25	65.43	67.35	70.65	68.08	66.57	69.25	67.94
		3	67.45	68.84	64.67	68.10	70.07	67.57	66.72	69.13	67.82
		4	67.50	68.75	64.73	68.13	69.59	67.54	66.68	69.73	67.83
	Translate again	1	66.52	68.40	64.75	66.15	71.14	67.70	65.05	69.58	67.41
		2	67.27	68.75	64.65	67.80	69.87	67.33	66.37	69.65	67.71
		3	66.82	68.11	64.15	67.54	69.33	66.78	65.40	68.88	67.13
		4	66.42	67.99	63.62	67.90	68.98	66.24	65.61	68.66	66.93

Table 6: Full CometKiwi-XL results for document-level translation with GPT-4o-mini and Gemini-2.0-Flash, across 8 language pairs. Step-by-step Step 1 results are not shown since the model does not generate a translation at this step. A darker green shade indicates a better score.

Model	Setup	Step	en→cs	en→de	en→fr	en→he	en→ja	en→ru	en→uk	en→zh	Avg.
GPT-4o-mini	Step-by-step	1	–	–	–	–	–	–	–	–	–
		2	8.16	3.10	6.58	11.76	7.19	8.41	8.21	7.57	7.62
		3	7.18	2.81	5.85	10.20	6.61	7.75	7.51	6.99	6.86
		4	7.01	2.77	5.79	9.67	6.58	7.52	7.27	6.93	6.69
	Translate again	1	8.44	3.18	7.05	11.77	7.34	8.99	8.73	7.89	7.93
		2	7.00	2.60	5.56	9.46	6.43	7.54	7.26	6.75	6.58
		3	6.63	2.53	5.24	9.23	6.09	7.09	6.89	6.44	6.27
		4	6.46	2.50	5.15	9.07	6.08	7.20	6.86	6.42	6.22
Gemini-2.0-Flash	Step-by-step	1	–	–	–	–	–	–	–	–	–
		2	6.59	3.04	6.32	8.89	6.02	7.40	6.97	7.27	6.56
		3	5.40	2.46	5.62	7.25	5.33	6.12	5.77	6.08	5.50
		4	5.31	2.41	5.52	6.97	5.24	5.93	5.67	5.92	5.37
	Translate again	1	6.72	3.08	6.69	9.46	5.87	7.69	7.39	7.31	6.78
		2	5.67	2.46	5.63	7.26	4.92	6.10	5.89	5.82	5.47
		3	5.59	2.44	5.67	7.18	4.90	6.08	6.11	5.97	5.49
		4	5.87	2.37	5.75	6.89	4.93	6.21	5.93	5.87	5.48

Table 7: Full MetricX results for document-level translation with GPT-4o-mini and Gemini-2.0-Flash, across 8 language pairs. Step-by-step Step 1 results are not shown since the model does not generate a translation at this step. A lower score and a darker green shade indicates better translation quality.

Model	Setup	Step	en→cs	en→de	en→fr	en→he	en→ja	en→ru	en→uk	en→zh	Avg.
GPT-4o-mini	Step-by-step	1	–	–	–	–	–	–	–	–	–
		2	55.70	73.35	56.62	43.45	52.92	58.07	56.14	53.49	56.22
		3	59.16	75.02	58.65	49.22	55.03	61.86	58.75	57.19	59.36
		4	59.65	75.19	58.40	50.98	55.34	62.44	60.02	57.28	59.91
	Translate again	1	55.45	73.39	55.39	42.68	51.88	56.96	55.67	52.59	55.50
		2	59.98	75.66	58.50	50.36	56.57	62.54	60.27	57.11	60.12
		3	60.99	75.28	58.49	52.06	57.50	63.64	60.49	58.48	60.87
		4	61.22	75.72	58.95	52.38	57.46	63.74	61.42	58.65	61.19
Gemini-2.0-Flash	Step-by-step	1	–	–	–	–	–	–	–	–	–
		2	61.91	74.87	57.85	55.68	58.63	63.99	62.33	56.77	61.50
		3	63.72	76.58	57.92	59.59	60.05	66.10	63.91	60.51	63.55
		4	64.21	76.70	58.26	60.36	60.27	66.97	63.74	61.54	64.01
	Translate again	1	62.14	73.97	57.40	53.55	58.66	63.43	61.09	57.93	61.02
		2	63.90	76.10	58.29	59.06	59.68	66.15	62.69	61.94	63.48
		3	62.73	75.70	56.77	58.68	58.39	65.47	61.29	61.46	62.56
		4	62.56	75.40	54.97	59.70	57.39	64.11	61.47	61.08	62.09

Table 8: Full XCOMET-XL results for document-level translation with GPT-4o-mini and Gemini-2.0-Flash, across 8 language pairs. Step-by-step Step 1 results are not shown since the model does not generate a translation at this step. A darker green shade indicates a better score.

Figure 8: Counts of translations by GPT-4o-mini that are faithful, neutral, or unfaithful to the decomposition, compared to direct translation. *Avg.* denotes the average over all 8 language directions.

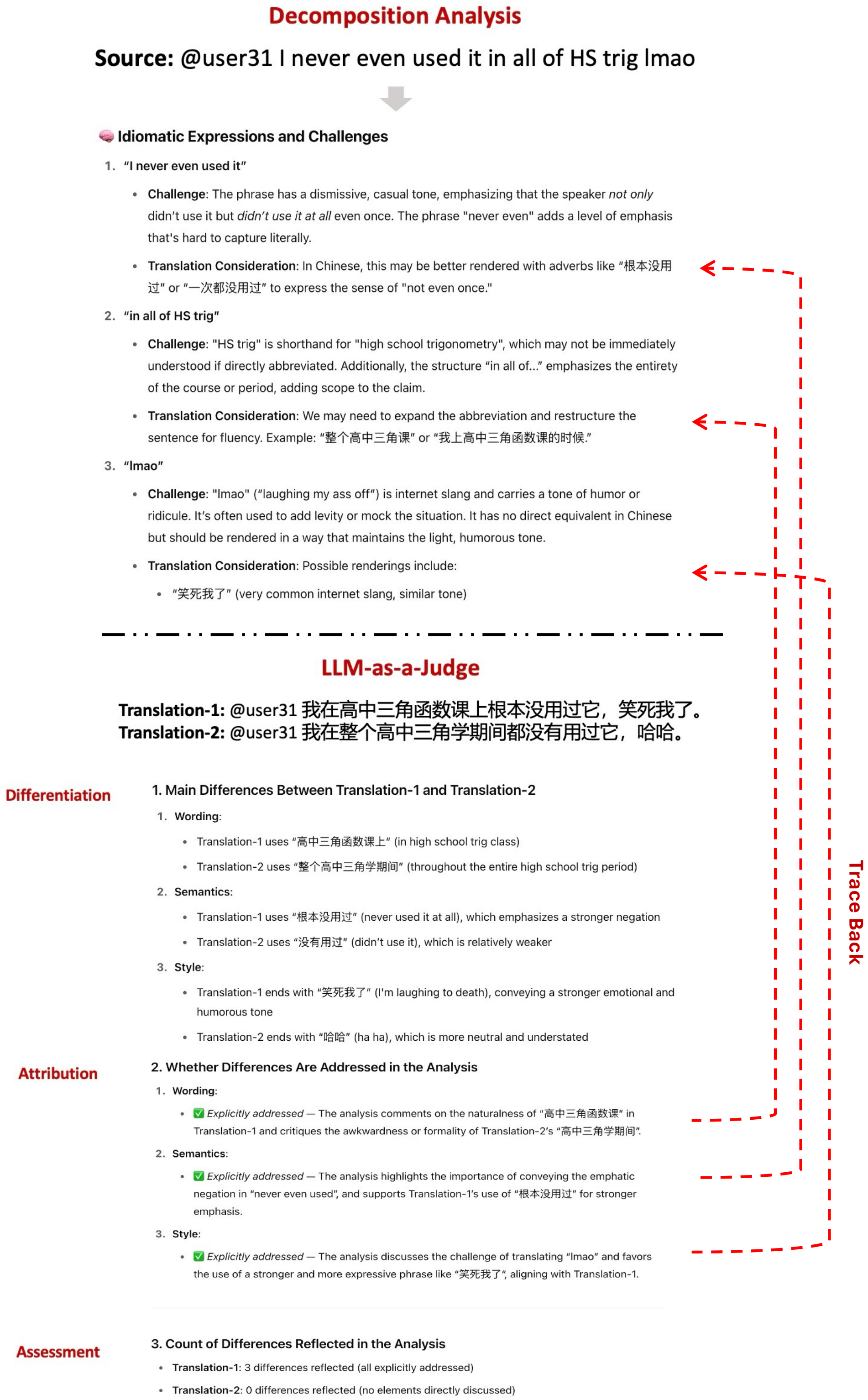


Figure 9: An illustration of using LLM-as-a-Judge to explicitly assess the impact of decomposition on translation behaviour. Given a source text and its corresponding decomposition (analysis results), GPT-4o is employed for three tasks: (1) **Differentiation** — identifying the differences between Translation 1 and Translation 2; (2) **Attribution** — mapping each translation difference back to specific elements of the decomposition; and (3) **Assessment** — evaluating the influence of the decomposition by measuring how many of the differences can be traced back to it.

System: You are a helpful assistant.

User: You will be asked to translate a piece of text from [source language] into [target language] following stages of the translation process. Here is the context in which the text appears:

Context: [source text]

To start, let's do some pre-drafting research on the above context.

Research: During this phase, thorough research is essential to address components of the context text that pose translation challenges. The goal is to establish a comprehensive translation plan that covers the following categories:

Idiomatic Expressions:

- Identify idiomatic expressions that cannot be directly translated word-for-word into [target language].

Figure 10: Prompt template used in the research (decomposition) stage of *step-by-step* translation.

System: You are a helpful assistant.

User: Now, let's move on to the drafting stage.

Draft Translation:

In this phase, your primary objective is to create a draft translation that accurately conveys the meaning of the source text presented below. At this stage, it is crucial to focus on adequacy, ensuring that your translation closely adheres to the source text. Your response should conclude with the draft translation. If context is missing, generate a general translation that is adaptable to various contexts. Avoid adding any additional information not present in the source text. All elements of the source text should be present in the translation.
Provide your single best translation of the following text, guided by the pre-drafting analysis, without adding anything further:

English: [source text]

Figure 11: Prompt used in the drafting (translation) stage of *step-by-step* translation.

System: You are a helpful assistant.

User: Now let's move to the next stage.

Post-editing with local refinement: In this stage, the primary aim is to refine the draft translation by making micro-level improvements that improve the draft's fluency.

Provide only one refined translation and do not output anything else after that.

Figure 12: Prompt used in the post-editing (refinement) stage of *step-by-step* translation.

System: You are a helpful assistant.

User: You are tasked with proofreading a translation that has been revised for improved fluency. The refined translation has been generated by editing the draft translation.

Proofreading and Final Editing: The goal is to provide a polished final translation of the source text. For your reference, below are the source text, the draft, and refined translations.

Source Text: [source text]
Draft Translation: [Step 2 output]
Refined Translation: [Step 3 output]

Please proofread the refined text for grammar, spelling, punctuation, terminology, and overall fluency. Ensure the translation accurately reflects the original meaning and style. Provide only the final, polished translation on the first line.

Figure 13: Prompt used in the proofreading stage of *step-by-step* translation.

System: You are a helpful assistant.

User: Please translate the following text from [source language] to [target language]. Provide only one translation and do not output anything else after that.

English: [source text]

Figure 14: Prompt used in the translation stage of *translate again* prompting.

System: You are a helpful assistant.

User: Please again translate the following text from [source language] to [target language] to make it better. Provide only one translation and do not output anything else after that.

English: [source text]

Figure 15: Prompt used in the refinement stage of *translate again* prompting. In this prompt, the model is provided with all previous prompts and outputs as part of a multi-turn conversation.

System: You are a helpful assistant.

User: Given the following English original text and the corresponding analysis:

English Original Text: [source text]

Analysis: [analysis]

Please analyze the differences between the following two translations in {tgt_lang}:

Translation-1: [translation 1]
Translation-2: [translation 2]

1. First, list the main differences between Translation-1 and Translation-2 in terms of wording, syntax, semantics, or style. Present the differences as a numbered list.

2. For each difference, state whether it is explicitly or implicitly addressed in the Analysis. If yes, mention the corresponding part of the analysis.

3. Count how many of the differences related to Translation-1 are reflected in the analysis, and how many related to Translation-2 are reflected.

4. Output only the following two tags on the last line:
`<trans-1-cnt>number</trans-1-cnt>` and `<trans-2-cnt>number</trans-2-cnt>`

Figure 16: Prompt used for *LLM-as-a-Judge*.