Translating Step-by-Step: Decomposing the Translation Process for Improved Translation Quality of Long-Form Texts

Eleftheria Briakou, Jiaming Luo, Colin Cherry, Markus Freitag
Google
{ebriakou,jmluo,colincherry,freitag}@google.com

Abstract

In this paper we present a step-by-step approach to long-form text translation, drawing on established processes in translation studies. Instead of viewing machine translation as a single, monolithic task, we propose a framework that engages language models in a multi-turn interaction, encompassing pre-translation research, drafting, refining, and proofreading, resulting in progressively improved translations. Extensive automatic evaluations using Gemini 1.5 Pro across ten language pairs show that translating step-by-step yields large translation quality improvements over conventional zero-shot prompting approaches and earlier human-like baseline strategies, resulting in state-of-the-art results on WMT 2024.

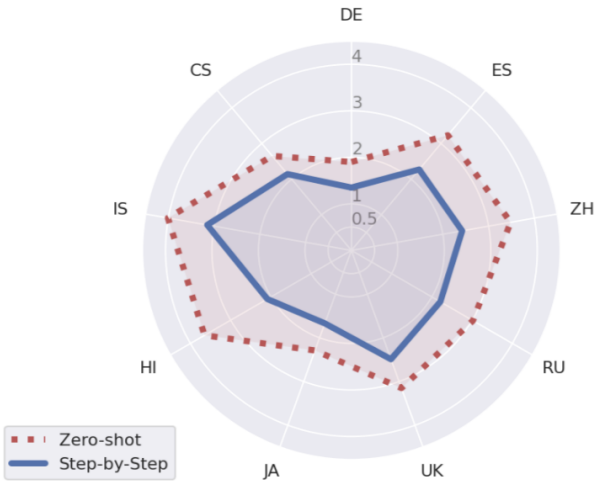

Figure 1: MetricX-23 quality improvements (where lower scores indicate better translation quality) on document-level translation on the WMT24 test set. Translate step-by-step with Gemini 1.5 Pro consistently outperforms zero-shot translation.

1 Introduction

Machine Translation (MT) has been traditionally seen as a sequence transduction task that maps a source text from one language to an equivalent translation in another language. While this simplified definition of the task served the modeling capabilities of statistical and neural machine translation systems for many years, recent advancements in large language modeling offer promise for redefining MT to align more closely with human translation processes. This shift prompts us back to a fundamental question: *what does a good translation process look like*?

Thankfully, this question has been a long-debated topic in the field of translation studies. Despite the lack of consensus around the nature of cognitive steps involved when humans translate, a common thread is apparent, i.e., translation is a multi-faceted *process* encompassing several sub-tasks that navigate a bilingual landscape. This view of translation finds a parallel in the rise of the "chain-of-thought" paradigm popularized by large language models (LLM) (Wei et al., 2022). That is, instead of attempting to generate the response to a complex task directly, LLMs are prompted to derive their final answer by decomposing the original task into several simpler sub-tasks.

But, what form would chain-of-thought take in the context of MT? While initial attempts to model the entire translation process using complex multi-stage processes has shown mixed results (Wu et al., 2024), explicitly modeling certain pre-translation or post-translation processes has led to more consistent gains in translation quality. On the pre-translation side, He et al. (2023) proposes to generate multiple translation candidates conditioned on self-generated translation-related knowledge. On the post-translation side, recent research threads prompt LLMs for refinement with (Feng et al., 2024; Xu et al., 2023b; Ki and Carpuat, 2024) or without (Chen et al., 2023) external quality estimation feedback.

Despite the promising results reported by prior work on decomposing and re-ranking MT with LLMs, it still remains unclear whether LLMs can

Proceedings of the Ninth Conference on Machine Translation, pages 1301–1317

benefit from modeling the *entire spectrum of translation processes*. In this work, drawing on literature from translation studies, we view MT as a complex and iterative task adhering to distinct steps, i.e., pre-translation research, drafting, refining, and proofreading. Based on this framework, we ask: *How well can* LLMs *translate in a step-by-step manner that draws from translation processes*?

Taking Gemini 1.5 Pro (Reid et al., 2024) as a case study, we start by designing instruction prompts for various translation subtasks. Concretely, our framework implements a ***multi-turn*** interaction with Gemini that breaks down the translation process into four distinct stages. It begins by prompting the model to conduct background research that identifies potential challenges in translating the source text (***research*** phase). The next interaction focuses on drafting an initial translation prioritizing faithfulness to the source text (***drafting*** phase). This draft is then revised in subsequent turns, ensuring a polished final translation (***refinement*** and ***proofreading*** phases).

To align better with human translation processes, we test the *translate step-by-step* framework on long-form documents derived from the general MT shared tasks for WMT 2023 (Kocmi et al., 2023) and WMT 2024. We evaluate out-of-English translation for ten languages, namely Chinese (ZH), Ukrainian (UK), Russian (RU), Japanese (JA), Hebrew (HE), Czech (CS), German (DE), Hindi (HI), Icelandic (IS), and Spanish (ES). Extensive automatic evaluation according to both reference-based and QE-based versions of MetricX-23 (Juraska et al., 2023) show that translating step-by-step yields strong translation quality improvements across all languages and test sets studied (see Figure 1).

2 Background

With the recent rise of LLMs, machine translation is going through a gradual but significant paradigm shift. While much research is focusing on how LLMs' training data are improving their MT capabilities (Xu et al., 2023a; Alves et al., 2024), there are also many opportunities to improve how existing LLMs can be best used for translation. This becomes evident in recent research that explores ways to augment and refine MT to align better with human translation processes. To navigate the diverse landscape of LLM-driven research, we summarize key studies in Table 1 along their four most distinct dimensions:

PAPER	PRE-TR.	POST-TR.	DEV.	PARAM.	STEPS
He et al. (2023)	✔	✗	✗	✗	3-4
Xu et al. (2023b)	✗	✔	✔	✗	Iterative
Feng et al. (2024)	✗	✔	✗	✗	3
Huang et al. (2024)	✗	✔	✔	✗	3
Li et al. (2024)	✔	✗	✗	✗	1
Chen et al. (2023)	✗	✔	✗	✔	Iterative
Ki and Carpuat (2024)	✗	✔	✗	✗	1
Wu et al. (2024)	✔	✔	✗	✔	Iterative
Step-by-Step (ours)	✔	✔	✔	✔	4

Table 1: List of prior work leveraging LLMs to improve translation quality by modeling either pre- or post-translation processes (PRE-TR. or POST-TR.). For each study we also note key aspects of their methodology: whether prompting strategies are developed on a separate development set (DEV.), whether the approach relies solely on the LLM's parametric knowledge (PARAM.), and the number of steps in the pipeline.

- **Temporal Focus**: This differentiating factor is based on whether an LLM is engaged in the translation process before (pre-translation) or after (post-translation) an initial translation is produced (whether by the same LLM or a different system).
- **Parametric vs. External Knowledge**: This dimension focuses on whether LLMs rely solely on their internal, learned knowledge (encoded in their parameters) or whether they use external resources, i.e., dictionaries, knowledge bases, retrieval engines or QE-based metrics (Mallen et al., 2023).
- **Reported Prompt Development**: This dimension considers whether the prompting strategies are clearly developed on separate development sets, as reported in papers.[1]
- **Number of Steps**: This dimension counts the number of distinct steps that are used in multi-turn interactions with the LLM.

Table 1 shows a clear trend: most studies focus on post-translation refinement. These approaches predominantly rely on external feedback to identify and correct errors, using either automatic met-

[1]We include this column not to cast aspersions on previous work, but to encourage a culture moving forward where prompt-based research uses and reports a development set. From personal communication, some of the works receiving an "✗" here underwent little to no prompt optimization.

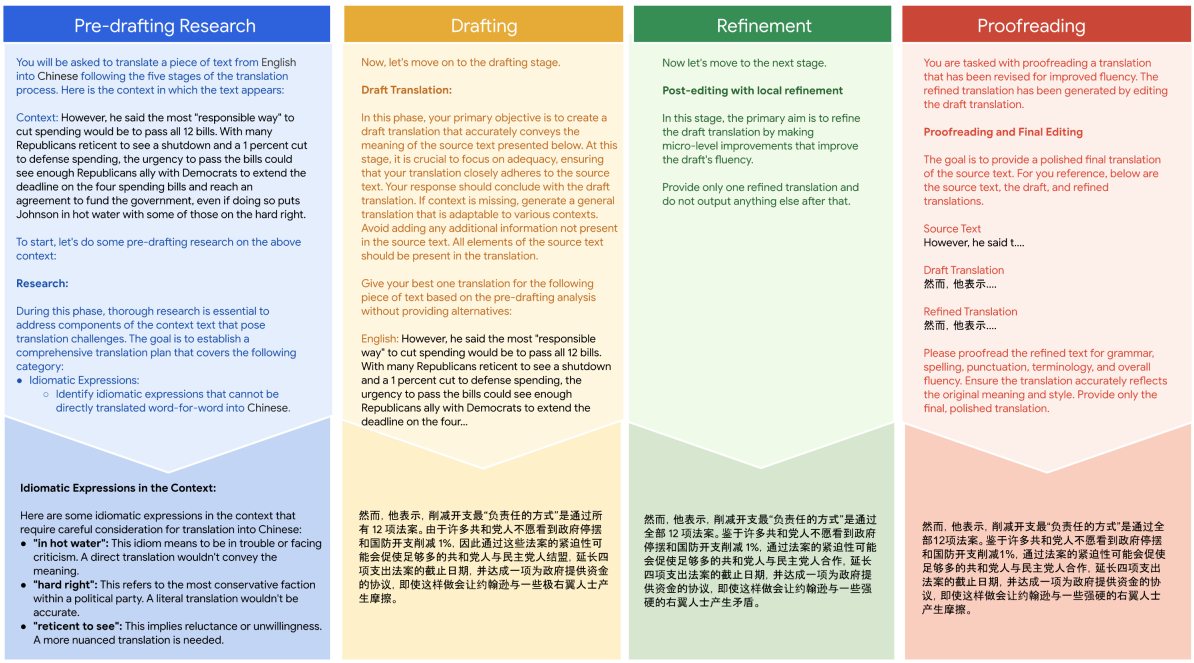

Figure 2: Translate Step-by-Step prompting framework. User prompts (top) and Gemini's responses (bottom) for the translation of an English document into Chinese. The full prompts for each step also appear in §A.3.

rics (Feng et al., 2024; Xu et al., 2023b; Huang et al., 2024) or human annotations of translation errors (Ki and Carpuat, 2024). A notable exception is the study of Chen et al. (2023), which shows that LLMs can iteratively refine their own outputs using only their parametric knowledge.

Comparatively fewer studies explore the pre-translation stage, investigating how LLMs can utilize background information to enhance their translation quality. He et al. (2023) explores this by prompting LLMs for different types of background information (similar examples, topics and keywords) related to the source text. However, they find that this knowledge alone is insufficient to improve the model's translation quality, and ultimately rely on external QE feedback for selection. In contrast, Li et al. (2024) operationalizes background research by incorporating idiom definitions retrieved from an external knowledge base.

A notable exception to the above is the recent work of Wu et al. (2024) which, similar to our approach, explores modeling the entire spectrum of translation processes. While conceptually aligned with our step-by-step approach, their framework is significantly more complex, with 30 distinct LLM roles interacting iteratively. Their use of non-standard metrics makes it difficult to gauge the method's success: the human evaluation does not give annotators source or reference texts, while the bilingual automatic evaluation collects only preference decisions using the same model family as the method being tested.

Overall, in contrast to prior work, which often relies on complex multi-stage processes and external resources, our goal is to streamline the translation process, ***unifying pre- and post-translation stages within one framework, by accessing only the LLM's parametric knowledge throughout***. We emphasize the methodological soundness of our pipeline by developing it on a separate development set, a practice not yet standardized in this area.

3 Translate Step-by-Step

Drawing on existing literature on translation studies (Borg, 2018), we design a series of staged prompts that attempt to map the translation process to instructions. This approach views translation as a multi-turn interaction with an LLM where each prompt guides the model's next action. Below, we describe those stages, along with what their function in the translation process is and how they are operationalized as instruction-following tasks. These stages are further illustrated in Figure 2.

Pre-translation Research Mirroring the human translation processes, our framework incorporates a pre-translation research stage. This stage primarily

focuses on using the source text (Mossop, 2000) to identify potential translation challenges drawing on real-world knowledge and knowledge of the target language (Dimitrova, 2005). We model this stage by prompting the LLM to identify and explain phrases of the source text that cannot be translated word-for-word into the target language.

Drafting Following the pre-translation research, the next stage aims at producing a draft translation, i.e., "the first stab at the rewriting" (Bassnett and Bush., 2016). This stage represents an initial attempt at rendering the source text into the target language. To that end, we initiate a subsequent interaction and prompt the model to focus on adequacy at this stage, ensuring the draft faithfully captures the meaning of the source.

Refinement The *post-drafting* stages are defined as editing tasks, with the goal of improving the overall quality of the draft translation. We define the first post-drafting stage as a subsequent interaction where the LLM is prompted to improve the draft's fluency such that the text works on its own (Borg, 2018).

Proofreading At the final, post-drafting stage, we task the LLM with the role of proofreading the refined translation to ensure it delivers a polished translation. We model this stage as a new conversation with the LLM, rather than a subsequent interaction, drawing inspiration from human studies suggesting that proofreading requires a new perspective after a break from revising (Shih, 2013).

3.1 Lessons During Development

While developing the above method, we found two factors to be important for the success of this approach: working at the document level and representing multi-step interactions as conversations.

Working at the Document Level Our multi-step process became more effective as we moved from the segments provided by WMT to working on multi-segment documents (see §4 for details on the setup). This had a large effect on the pre-translation research step, changing it in two ways. First, some phrases that appeared idiomatic or difficult at the segment level disappeared, as their translations became clear with context. Second, the LLM began identifying larger phrases. The refinement step also improved according to automatic metrics. We verified that our shift to the document level was either neutral or an improvement for our baselines (§5.3).

Domain	Literary	News	Social	Speech
# Docs.	40	43	48	111
Avg. Length	192	184	164	73

Table 2: Per-domain statistics for WMT 2024.

Multi-step Interactions as Conversations Modern LLMs use special markers to indicate human versus assistant turns in multi-turn interactions. When building an automated process like translate step-by-step, for each step, one has the option to either use previous outputs to build a completely new query that summarizes all previous interactions, or to continue the conversation, allowing the LLM to see all previous steps with its own outputs clearly marked. With the exception of the proofreading step, we found that continuing the conversation improved performance. Also, breaking the conversation into smaller turns helps with modularity for ablations.

4 Experimental Setting

We start by evaluating the translate step-by-step approach on the task of document-level translation. The experimental setting is described below.

Model Settings Throughout our experiments we use Gemini 1.5 Pro. All model outputs are generated with greedy decoding. All model prompts are provided in Appendix A.3. In zero-shot mode, the model is instructed to translate the source text directly, without providing any explanations.

To effectively isolate the artifacts from pre-translation research, we employ a secondary model call. This call restructures the natural language output into a JSON object, simplifying the parsing process for extracting artifacts.

Evaluation Sets We use WMT 2023 as our ***development*** set. Any prompt development and stage ablation experiments are conducted on this dataset. For our final ***test*** set, we use the WMT 2024 datasets. Each of these datasets was built by translating a set of English documents into multiple languages.

Both datasets are segmented for sentence- or paragraph-level evaluation, but our approach focuses on translating with as much context as possible. Therefore, we use meta-data to merge the original segments into larger ones. Ideally, this would result in complete documents, but current neural metrics have token-count limits beyond which they truncate their inputs. To accommodate neural evaluation, we set a maximum length of 250 (English

	Research	Draft	Refinement	Proofreading	ZH	UK	RU	JA	HE	CS	DE	AVERAGE
Ref-based												
1.	○	○	○	○	3.64	4.18	3.32	2.59	4.36	2.82	1.82	3.25
2.	○	●	○	○	3.48 ↓0.16	4.16 ↓0.02	3.32 ↓0.00	2.47 ↓0.12	4.54 ↑0.18	2.67 ↓0.15	1.92 ↑0.10	3.22
3.	○	○	●	○	2.92 ↓0.72	3.32 ↓0.86	2.43 ↓0.89	2.19 ↓0.40	3.24 ↓1.12	2.35 ↓0.47	1.31 ↓0.51	2.54
4.	○	●	●	○	2.85 ↓0.79	3.06 ↓1.12	2.54 ↓0.78	2.09 ↓0.50	3.18 ↓1.18	2.22 ↓0.60	1.37 ↓0.45	2.47
5.	●	●	○	○	3.00 ↓0.64	3.46 ↓0.72	2.56 ↓0.76	2.05 ↓0.53	3.89 ↓0.47	1.97 ↓0.85	1.56 ↓0.26	2.64
6.	●	●	●	○	2.63 ↓1.01	2.70 ↓1.47	2.13 ↓1.19	1.73 ↓0.86	2.88 ↓1.48	1.85 ↓0.96	1.17 ↓0.65	2.16
7.	●	●	●	●	2.67 ↓0.97	2.38 ↓1.80	2.16 ↓1.16	1.70 ↓0.89	2.75 ↓1.61	1.71 ↓1.10	1.07 ↓0.75	2.06
QE-based												
8.	○	○	○	○	2.64	4.87	4.16	1.73	5.55	5.39	3.96	4.04
9.	○	●	○	○	2.71 ↑0.07	4.78 ↓0.09	4.05 ↓0.11	1.65 ↓0.07	5.22 ↓0.33	5.14 ↓0.25	4.03 ↑0.08	3.94
10.	○	○	●	○	2.11 ↓0.52	4.33 ↓0.54	2.82 ↓1.34	1.30 ↓0.43	4.49 ↓1.06	4.31 ↓1.08	2.89 ↓1.07	3.18
11.	○	●	●	○	2.04 ↓0.59	4.12 ↓0.75	3.31 ↓0.85	1.19 ↓0.54	4.30 ↓1.25	4.40 ↓0.99	3.36 ↓0.60	3.25
12.	●	●	○	○	2.26 ↓0.38	4.18 ↓0.69	3.50 ↓0.66	1.54 ↓0.19	4.60 ↓0.95	4.62 ↓0.77	3.73 ↓0.23	3.49
13.	●	●	●	○	1.90 ↓0.73	3.39 ↓1.48	2.76 ↓1.40	1.23 ↓0.49	4.17 ↓1.38	4.12 ↓1.28	2.97 ↓0.99	2.93
14.	●	●	●	●	1.82 ↓0.81	3.43 ↓1.44	3.11 ↓1.05	1.25 ↓0.48	4.01 ↓1.54	3.56 ↓1.83	2.63 ↓1.33	2.83

Table 3: MetricX-23 evaluation results of translate step-by-step and its ablation variants on the WMT 2023 development datasets. We report both the reference-based and QE-based metric variants. Filled dots indicate active steps in the pipeline, while unfilled dots represent ablated steps. When all steps are ablated, the system defaults to zero-shot translation. Colored boxes highlight performance differences compared to zero-shot: blue shades indicate significant improvements at $p < 0.001$, green shades indicate significant improvements at $p < 0.05$, yellow shades indicate non-significant improvements ($p \geq 0.05$), while red shades indicate non-significant regressions ($p \geq 0.05$) against zero-shot. *Translate step-by-step surpasses zero-shot across the board, with each step incrementally improving translation quality.*

white-space separated) tokens each.[2] The resulting datasets consist of 192 documents of average token length 178 for WMT 2023 and, 243 documents of average token length 130 for WMT 2024, respectively. For WMT 2024 we also report per-domain results. Per-domain document counts and average lengths, as measured in English white-space separated tokens, are presented in Table 2.

Evaluation Metrics We evaluate our approach using MetricX-XXL-23 (Juraska et al., 2023), the metric adopted in the most recent WMT 2024 automatic evaluations. We report results on both the reference-based and the QE-based metric variants. Despite being trained at the sentence level, Deutsch et al. (2023) show that MetricX can effectively evaluate multi-sentence sequences, capped at its maximum window length. We note that MetricX is powered by mT5 (Xue et al., 2021), which minimizes the potential bias in favor of Gemini-generated translations.[3] We employ paired permutation tests to determine if the observed improvements across system pairs are statistically significant.[4]

[2]We also present results for a shorter set of documents, with a maximum length of 150 tokens in Appendix A.1.

[3]We also report ChrF (Popović, 2015) in Appendix A.2.

[4]https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.permutation_test.html

5 Quantitative Results

We start by analyzing the importance of each step in the translate step-by-step pipeline. Ablation results on the WMT 2023 development sets are presented in §5.1. Next, the generalizability of our final step-by-step recipe is evaluated on the WMT 2024 test sets in §5.2, with comparison to prior work in §5.3.

5.1 Analyzing Step Importance

Automatic evaluation results on our development sets are presented in Table 3. Overall, translation artifacts extracted through the step-by-step process yield consistently better document translations compared to the zero-shot mode according to both reference- (lines 3–7 vs. 1) and QE-based (lines 10–14 vs. 8) versions of MetricX. Ablating the various steps from the pipeline gives insights into how each step contributes to the overall quality improvements. We describe those below.

Importance of Pre-translation Research Modelling pre-translation processes is crucial for achieving higher quality translations compared to the zero-shot. Simply prompting for a draft translation without asking for pre-translation research yields only small and non-significant improvements or even regressions over the zero-shot (lines 2 vs. 1

and 9 vs. 8). This result rules out the possibility that any observed improvements are solely due to a better prompt for the draft translation, which was modified to emphasize faithfulness to the source (§3). However, combining the research and draft steps achieves consistently higher quality translations compared to zero-shot (lines 5 vs. 1 and 12 vs. 8). Importantly, those improvements are consistently statistical significant ($p < 0.0001$) across languages (measured by reference-based metrics), except for Hebrew, which shows non-significant improvements compared to zero-shot ($p \geq 0.05$).

Importance of Refinement Moving to the evaluation of the *refined* document translations, we notice an interesting trend. The refinement step consistently improves the translation quality, regardless of the initial translation it processes, i.e., the zero-shot (lines 3 vs. 1 and 10 vs. 8), the single-turn draft (lines 4 vs. 2 and 11 vs. 9), and the research-informed draft (lines 6 vs. 5 and 13 vs. 12). This demonstrates that the effectiveness of the refinement stage is not conditioned on the initial translation. However, the strongest quality improvements—reaching consistently high levels of statistical significance ($p < 0.001$) over the zero-shot translations—are observed when the refinement stage is combined with the pre-translation research (lines 6 vs. 1 and 13 vs. 8), highlighting that those stages bring complimentary benefits.

Importance of Proofreading Finally, the evaluation of the *proofreading* document translations, indicate that this stage contributes modest average improvements (lines 7 vs. 6 and 14 vs. 13). Unlike previous stages, the impact of proofreading appears to be more language dependent. Ukrainian stands out as the only language that clearly benefits from a proofreading stage, while others show only minor differences in quality compared to their refined translations.

5.2 Generalizability of Step-by-Step

Table 4 presents results on the WMT 2024 test set. Across the board, translating step-by-step exhibits the same trends noticed on our development set (as discussed in §5.1). This confirms the generalizability of our proposed approach, crucially, on a wider range of languages. Concretely, the draft translations outperform the zero-shot translations. The refined stages bring additional quality improvements across the board, with the proofreading stage contributing small improvements for most languages.

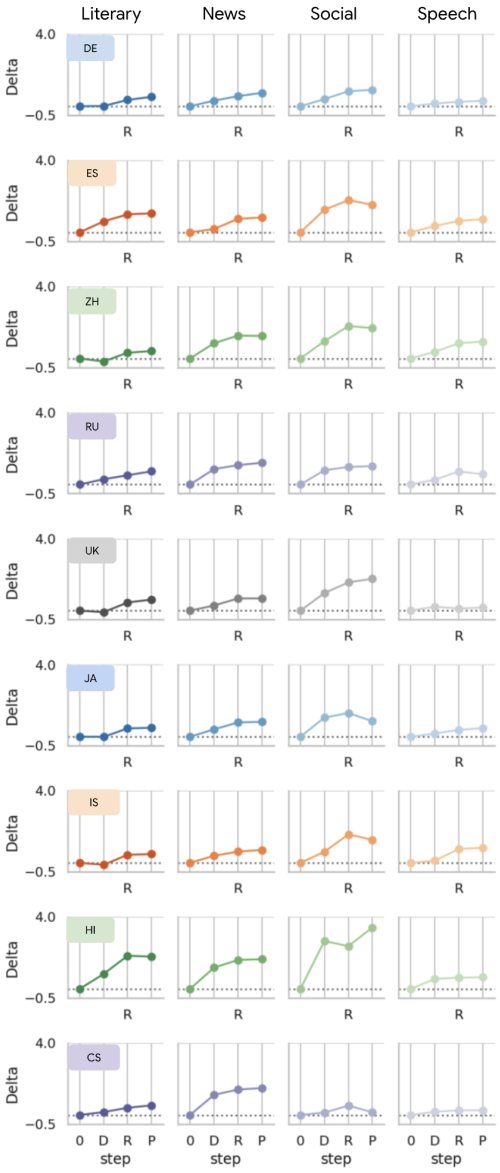

Figure 3: Domain-level comparison between zero-shot and step-by-step translations on WMT 2024 using reference-based MetricX-23. Each data point represents the delta from zero-shot (dotted horizontal line). The steps are denoted as follows: 0 (zero-shot), D (draft after research), R (refinement), and P (proofreading).

To better understand the robustness of our approach we present a per-domain analysis in Figure 3. As shown, translation quality improvements of step-by-step translations over zero-shot are observed across all domains, with speech showing the least and social the most significant gains.

5.3 Contextualizing Step-by-Step Gains

Having demonstrated how translate step-by-step improves long-form translation with LLMs over zero-shot translation, we now contextualize these gains by comparing our approach to two repre-

	DE	ES	ZH	RU	UK	JA	HI	IS	CS	AVERAGE
										Ref-based
Zero-shot	1.90	3.23	3.48	3.02	3.15	2.29	3.65	4.01	2.65	3.04
SBYS: *Research & Drafting*	1.68 ↓0.22	2.69 ↓0.54	2.99 ↓0.49	2.53 ↓0.49	2.81 ↓0.35	1.92 ↓0.37	2.52 ↓1.13	3.77 ↓0.24	2.30 ↓0.35	2.58
SBYS: *Refinement*	1.45 ↓0.45	2.29 ↓0.94	2.45 ↓1.03	2.21 ↓0.81	2.58 ↓0.57	1.64 ↓0.66	2.31 ↓1.35	3.14 ↓0.87	2.10 ↓0.55	2.24
SBYS: *Proofreading*	1.35 ↓0.54	2.27 ↓0.96	2.42 ↓1.06	2.21 ↓0.81	2.49 ↓0.66	1.67 ↓0.62	2.09 ↓1.56	3.15 ↓0.86	2.14 ↓0.51	2.20
										QE-based
Zero-shot	1.97	2.59	2.23	1.87	2.23	1.32	4.81	3.47	2.08	2.51
SBYS: *Research & Drafting*	1.72 ↓0.25	2.23 ↓0.36	2.08 ↓0.15	1.54 ↓0.33	1.81 ↓0.41	1.19 ↓0.13	4.12 ↓0.69	3.43 ↓0.04	1.97 ↓0.11	2.23
SBYS: *Refinement*	1.38 ↓0.59	1.78 ↓0.81	1.71 ↓0.52	1.21 ↓0.66	1.34 ↓0.89	0.95 ↓0.37	3.47 ↓1.34	2.79 ↓0.68	1.51 ↓0.56	1.79
SBYS: *Proofreading*	1.25 ↓0.72	1.74 ↓0.84	1.63 ↓0.60	1.14 ↓0.73	1.32 ↓0.91	0.93 ↓0.40	3.35 ↓1.46	2.65 ↓0.82	1.45 ↓0.63	1.72

Table 4: MetricX-23 results comparing step-by-step (SBYS) with zero-shot on the WMT 2024 test datasets. When all steps are ablated, the system defaults to zero-shot translation. Colored boxes highlight performance differences compared to zero-shot: blue shades indicate significant improvements at $p < 0.001$, green shades indicate significant improvements at $p < 0.05$, while yellow shades indicate non-significant improvements ($p \geq 0.05$). *Translate step-by-step surpasses zero-shot, with each step incrementally improving translation quality.*

sentative baselines: a) methods that leverage non-parametric knowledge for best translation selection, and b) segment-level baselines that translate documents using the pre-defined segmentation provided in WMT 2024 test sets.

Conditions As a representative of the first class, we compare against MAPS (He et al., 2023). This baseline employs an LLM to analyze the source text for topic, keywords, and similar examples, generating three candidate translations conditioned on each knowledge type. Then, a QE metric selects the best candidate. To create a fair comparison, we re-implement their method using Gemini 1.5 Pro, using the prompts provided in their released code. To create an even stronger baseline, we perform candidate selection with the QE variant of MetricX-23, which we know correlates well with the final reference-based MetricX-23, creating an advantage for MAPS.

For the second class of baselines, we consider two approaches: a) zero-shot translation applied to each segment individually using Gemini 1.5 Pro, both with (ZERO-SHOT IN CONTEXT) and without (ZERO-SHOT) access to the full document in the input prompt, and b) a comparison with the segment-level translations from Unbabel-Tower70B, the top-performing system of WMT 2024 based on early automatic evaluations (Kocmi et al., 2024). To get comparable document-level metrics, before evaluation, we concatenate the segment-level translation back into the mini-documents, as described in §4.

We focus our comparisons on EN-DE, EN-JA, EN-ZH, as MAPS requires in-context demonstrations that were made available only for those languages by the original authors. For a fair comparison with Unbabel-Tower70B, we exclude the speech domain from our comparison, as WMT 2024 submissions were given ASR transcripts instead of human-sourced transcripts in this domain.

METHOD	DOC.	EN-DE	EN-ZH	EN-JA
UNBABEL-TOWER70B	✗	1 1.42	1 2.77	2 2.16
ZERO-SHOT	✗	2 1.98	3 3.65	3 2.60
ZERO-SHOT IN CONTEXT	✗	2 1.86	2 3.33	2 2.19
ZERO-SHOT	✔	3 2.02	3 3.91	3 2.47
MAPS	✔	2 1.91	2 3.25	2 2.19
SBYS: *Research & Drafting*	✔	2 1.75	2 3.32	2 1.94
SBYS: *Refinement*	✔	1 1.41	1 2.73	1 1.58
SBYS: *Proofreading*	✔	1 1.27	1 2.75	1 1.73

Table 5: Comparison of step-by-step (SBYS) with representative baselines (lower scores are better) on WMT 2024 according to Metric-X (reference-based). The second column indicates whether translation is performed on the entire document or by merging segment-level translations. Numbered squares represent significance clusters (Freitag et al., 2023) at $p = 0.05$. *Translate step-by-step matches or exceeds all compared baselines, crucially, without accessing external resources.*

Results Table 5 compares step-by-step against various baselines. Notably, even the initial stage, where the draft translation is conditioned on cross-lingual research (SBYS: *Research & Drafting*) demonstrates competitive performance against MAPS, falling within the same statistical significance cluster. This highlights the effectiveness of our pre-translation strategy compared to the background information used by MAPS. Comparing the final, proofreading stage of step-by-step (SBYS: *Proofreading*) with MAPS reveals significant translation quality gains: 0.64 improvement for DE, 0.50 for ZH, and 0.46 for JA. Notably, these improvements are achieved even though MAPS uses the same QE model family as MetricX for final candidate selection, giving it an inherent advantage. In contrast, SBYS relies solely on the model's internal, parametric knowledge throughout the entire translation process.

Comparing the final, proofreading stage of step-

by-step with the segment-level baselines helps put the improvements in perspective. Concretely, the segment-level zero-shot baselines (second and third lines in Table 5) fall significantly behind the step-by-step final translations (SBYS: *Proofreading*) across all languages by more than 0.7 and 0.4 MetricX points when compared to the out-of- and in-context variants, respectively. This demonstrates that simply translating documents at a finer granularity is not sufficient for boosting the LLM's translation quality.

Finally, comparing the final, proofreading stage of our approach with the merged translations from Unbabel-Tower70B, reveals that our approach achieves statistically comparable performance for Chinese and German (0.02 and 0.15 improvements respectively) and significantly better performance for Japanese (0.43 improvement). These improvements over the top-performing WMT 2024 system demonstrate the competitiveness of the step-by-step approach, especially given that the competing system relies on external QE metrics and computationally expensive decoding strategies to improve translation quality.

6 Qualitative Analysis

We conduct a qualitative analysis on a small subset of model outputs from all stages to understand the strengths and weaknesses of our step-by-step approach. To this end, we first compute the score difference between the final translation and the zero-shot output on WMT 2024 English to Chinese, and then randomly sample up to 5 examples from either end (i.e., examples for which the final translation quality either substantially improves or degrades over the zero-shot baseline).[5] One of the authors (native speaker of Chinese) manually inspected the sampled outputs and took notes on the salient properties of the pre-translation artifacts and the incremental changes from the different stages of the step-by-step process.

Pre-drafting For pre-drafting research, we observe that the LLM is highly capable of understanding the source in a wide variety of contexts. As showcased in Table 6, the LLM is able to correctly interpret slang (example 1: *cheeked up* in the context of making miniatures), recognize figurative usage (example 2: *the weather didn't cooperate* in the context of flying a plane), and detect humorous expressions (example 3). This strength is especially pronounced when even the references show clear signs of human translators misinterpreting the source (see the next subsection for full examples).

On the other hand, the LLM is also prone to over-generate and seems too eager to confirm with the given instruction to find instances of indirect translation. This resulted in false positives where a direct and literal translation is already adequate (example 4: *a bit dazed* can be directly translated into Chinese), and in some cases bizarre cultural commentaries (example 5 for asking to contextualize the texture of bubble gum).

Translations The observed understanding of the source texts seems to directly contribute to more fluent and context-appropriate translations. Table 13 in §A.4 shows several interesting examples. There are quite a few instances where the step-by-step approach produces the correct translation even when the reference fails to recognize the context the phrase appears in. For example, our method correctly interprets the meaning of *cheeked up* in the first example to be "having a full cheek" when the zero-shot translates it to *blushing* and the reference translates it to *talking nonsense*. Similarly for the second example, the term *threading* is correctly understood as a thread of posts on a social media platform by all step-by-step translations, whereas the zero-shot interpretes it as a computing terminology (as in *multithreading*) and the reference interprets it as *study*.

It is also evident that the refinement improves the fluency significantly. The third example shows that both the zero-shot and the draft translates the source literally. This entails preserving the original source structure and translating the source word *ridiculous* directly. The result is somewhat awkward and sounds like translationese, which is particularly jarring considering the social media domain of the source text. This issue is rectified by both the refined translation and the reference. On the other hand, the refinement process is not perfect and does regress to less fluent outputs at times (example 4). Another prominent failure mode typical of all LLM outputs is hallucinations. In the fifth example, the refinement adds "touching the bruise", which is not present in the source.

[5] The exact sample ranges of the score difference are [-6, -2] and [1, 6]. Examples from beyond these ranges typically demonstrate clear signs of model degeneration and are therefore excluded from this analysis.

	SOURCE SNIPPET	PRE-TRANSLATION RESEARCH	NOTES
1. ✔	trying my hand at miniatures redid the head! *[content omitted]* @user26 cheeked up. I love HER*	**"cheeked up"**: This slang term likely refers to making the miniature's cheeks fuller or more prominent. A direct translation wouldn't be accurate.	*Understanding source correctly in the context of making miniatures.*
2. ✔	#TeaganAir #pilot #flying #pa28 #PilotTraining *[content omitted]* Last week the weather didn't cooperate so we're back at it today.	**"The weather didn't cooperate"**: This is a figurative way of saying the weather was not suitable for flying.	*Good example of indirect translation.*
3. ✔	It is nice to have art and entertainment to help us make sense of our life. @user32 I could perform a liver transplant thanks to Hospital Playlist.	**"I could perform a liver transplant thanks to Hospital Playlist"**: This is a humorous exaggeration, implying that the show provided a lot of information about liver transplants.	*Good humor detection.*
4. ✗	He was covered in cuts and scratches and seemed a bit dazed, but he didn't seem too badly injured.	**"a bit dazed"**: This describes a state of confusion or disorientation. A literal translation might not convey the same meaning.	*Literal translation would do just fine.*
5. ✗	been really enjoying kneadatite (green stuff), it's like sculpting with bubble gum but you get used to that and it's tough and flexible when it cures	**"like sculpting with bubble gum"**: : This is a simile comparing the texture of Kneadatite to bubble gum. It needs to be translated in a way that makes sense to a Chinese audience who might not be familiar with the specific texture of bubble gum.	*Strange commentary on contextualizing the texture of bubble gum to a Chinese audience.*

Table 6: Samples of pre-translation research outputs along with notes from the author-annotator.

7 Conclusion

We introduce a step-by-step approach to long-form text translation using LLMs. Inspired by literature on translation studies, we decompose the translation process into distinct stages, modeling pre-translation research, drafting, refinement, and proofreading though a multi-turn interaction with Gemini 1.5 Pro. Extensive automatic evaluations on WMT 2023 and WMT 2024 tasks in ten languages demonstrate that our approach improves translation quality over directly translating the entire document with a single prompt.

Furthermore, comparison with competitive baselines, including similar human-like LLM-driven approaches and top-performing systems that employ segment-by-segment translation of a document, reveals the strong performance of our approach. Our findings highlight the potential of LLMs to progressively improve their translations, moving beyond the traditional view of machine translation as a monolithic sequence mapping task.

Limitations

While our study reveals promising step-by-step improvements across various languages and domains, we acknowledge the limitations of solely relying on automatic metrics for evaluation. While metric improvements give us a consistent signal, human evaluation is needed to further validate the effectiveness of the approach and reveal a more nuanced understanding of the translation properties introduced at each step. We also acknowledge that our analysis is based solely on one family of metrics, due to context window limitations of other neural metrics in evaluating longer texts.

Finally, our pipeline is developed and tested solely on Gemini. Since different LLMs might exhibit different instruction-following capabilities across languages, the generalizability of this approach to other LLMs requires further investigation.

Ethics Statement

This paper explores the use of LLMs to improve translation quality. In doing so, our approach starts from an initial translation that prioritizes faithfulness to the source text. Subsequent stages focus on improving fluency which, as they deviate more from the source, increase the risk of hallucinations (Guerreiro et al., 2023)—a critical issue in machine translation, potentially leading to misleading translations.

Moreover, the increasing fluency of machine translations presents new challenges when prioritized over adequacy (Martindale and Carpuat, 2018), as users might trust their outputs blindly, even when incorrect. This highlights the need for careful adoption of those translation systems and the developing of strategies that help users calibrate their trust appropriately.

References

Duarte M. Alves, José P. Pombal, Nuno M. Guerreiro, Pedro H. Martins, Joao Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, Jos'e G. C. de Souza, and André Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. *ArXiv*, abs/2402.17733.

Susan Bassnett and Peter R. Bush. 2016. The translator as writer. london: Continuum.

Claudine Borg. 2018. The phases of the translation process: are they always three?

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2023. Iterative translation refinement with large language models. *ArXiv*, abs/2306.03856.

Daniel Deutsch, Juraj Juraska, Mara Finkelstein, and Markus Freitag. 2023. Training and meta-evaluating machine translation evaluation metrics at the paragraph level. In *Proceedings of the Eighth Conference on Machine Translation*, pages 996–1013, Singapore. Association for Computational Linguistics.

Birgitta Englund Dimitrova. 2005. Expertise and explicitation in the translation process.

Zhaopeng Feng, Yan Zhang, Hao Li, Wenqiang Liu, Jun Lang, Yang Feng, Jian Wu, and Zuozhu Liu. 2024. Improving llm-based machine translation with systematic self-correction.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.

Nuno M. Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2023. Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*, 12:229–246.

Yi-Chong Huang, Xiaocheng Feng, Baohang Li, Chengpeng Fu, Wenshuai Huo, Ting Liu, and Bing Qin. 2024. Aligning translation-specific understanding to general understanding in large language models. *ArXiv*, abs/2401.05072.

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. *ArXiv*, abs/2404.07851.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. 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, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinth'or Steingr'imsson, and Vil'em Zouhar. 2024. Preliminary wmt24 ranking of general mt systems and llms.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):18554–18563.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Marianna Martindale and Marine Carpuat. 2018. Fluency over adequacy: A pilot study in measuring user trust in imperfect MT. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 13–25, Boston, MA. Association for Machine Translation in the Americas.

Brian Mossop. 2000. The workplace procedures of professional translators.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the*

Tenth Workshop on Statistical Machine Translation, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem W. Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, Luke Vilnis, Oscar Chang, Nobuyuki Morioka, George Tucker, Ce Zheng, Oliver Woodman, Nithya Attaluri, Tomás Kociský, Evgenii Eltyshev, Xi Chen, Timothy Chung, Vittorio Selo, Siddhartha Brahma, Petko Georgiev, Ambrose Slone, Zhenkai Zhu, James Lottes, Siyuan Qiao, Ben Caine, Sebastian Riedel, Alex Tomala, Martin Chadwick, J Christopher Love, Peter Choy, Sid Mittal, Neil Houlsby, Yunhao Tang, Matthew Lamm, Libin Bai, Qiao Zhang, Luheng He, Yong Cheng, Peter Humphreys, Yujia Li, Sergey Brin, Albin Cassirer, Ying-Qi Miao, Lukás Zilka, Taylor Tobin, Kelvin Xu, Lev Proleev, Daniel Sohn, Alberto Magni, Lisa Anne Hendricks, Isabel Gao, Santiago Ontan'on, Oskar Bunyan, Nathan Byrd, Abhanshu Sharma, Biao Zhang, Mario Pinto, Rishika Sinha, Harsh Mehta, Dawei Jia, Sergi Caelles, Albert Webson, Alex Morris, Becca Roelofs, Yifan Ding, Robin Strudel, Xuehan Xiong, Marvin Ritter, Mostafa Dehghani, Rahma Chaabouni, Abhijit Karmarkar, Guangda Lai, Fabian Mentzer, Bibo Xu, YaGuang Li, Yujing Zhang, Tom Le Paine, Alex Goldin, Behnam Neyshabur, Kate Baumli, Anselm Levskaya, Michael Laskin, Wenhao Jia, Jack W. Rae, Kefan Xiao, Antoine He, Skye Giordano, Lakshman Yagati, Jean-Baptiste Lespiau, Paul Natsev, Sanjay Ganapathy, Fangyu Liu, Danilo Martins, Nanxin Chen, Yunhan Xu, Megan Barnes, Rhys May, Arpi Vezer, Junhyuk Oh, Ken Franko, Sophie Bridgers, Ruizhe Zhao, Boxi Wu, Basil Mustafa, Sean Sechrist, Emilio Parisotto, Thanumalayan Sankaranarayana Pillai, Chris Larkin, Chenjie Gu, Christina Sorokin, Maxim Krikun, Alexey Guseynov, Jessica Landon, Romina Datta, Alexander Pritzel, Phoebe Thacker, Fan Yang, Kevin Hui, A.E. Hauth, Chih-Kuan Yeh, David Barker, Justin Mao-Jones, Sophia Austin, Hannah Sheahan, Parker Schuh, James Svensson, Rohan Jain, Vinay Venkatesh Ramasesh, Anton Briukhov, Da-Woon Chung, Tamara von Glehn, Christina Butterfield, Priya Jhakra, Matt Wiethoff, Justin Frye, Jordan Grimstad, Beer Changpinyo, Charline Le Lan, Anna Bortsova, Yonghui Wu, Paul Voigtlaender, Tara N. Sainath, Charlotte Smith, Will Hawkins, Kris Cao, James Besley, Srivatsan Srinivasan, Mark Omernick, Colin Gaffney, Gabriela de Castro Surita, Ryan Burnell, Bogdan Damoc, Junwhan Ahn, Andrew Brock, Mantas Pajarskas, Anastasia Petrushkina, Seb Noury, Lorenzo Blanco, Kevin Swersky, Arun Ahuja, Thi Avrahami, Vedant Misra, Raoul de Liedekerke, Mariko Iinuma, Alex Polozov, Sarah York, George van den Driessche, Paul Michel, Justin Chiu, Rory Blevins, Zach Gleicher, Adrià Recasens, Alban Rrustemi, Elena Gribovskaya, Aurko Roy, Wiktor Gworek, S'ebastien M. R. Arnold, Lisa Lee, James Lee-Thorp, Marcello Maggioni, Enrique Piqueras, Kartikeya Badola, Sharad Vikram, Lucas Gonzalez, Anirudh Baddepudi, Evan Senter, Jacob Devlin, James Qin, Michael Azzam, Maja Trebacz, Martin Polacek, Kashyap Krishnakumar, Shuo yiin Chang, Matthew Tung, Ivo Penchev, Rishabh Joshi, Kate Olszewska, Carrie Muir, Mateo Wirth, Ale Jakse Hartman, Joshua Newlan, Sheleem Kashem, Vijay Bolina, Elahe Dabir, Joost R. van Amersfoort, Zafarali Ahmed, James Cobon-Kerr, Aishwarya B Kamath, Arnar Mar Hrafnkelsson, Le Hou, Ian Mackinnon, Alexandre Frechette, Eric Noland, Xiance Si, Emanuel Taropa, Dong Li, Phil Crone, Anmol Gulati, S'ebastien Cevey, Jonas Adler, Ada Ma, David Silver, Simon Tokumine, Richard Powell, Stephan Lee, Michael B. Chang, Samer Hassan, Diana Mincu, Antoine Yang, Nir Levine, Jenny Brennan, Mingqiu Wang, Sarah Hodkinson, Jeffrey Zhao, Josh Lipschultz, Aedan Pope, Michael B. Chang, Cheng Li, Laurent El Shafey, Michela Paganini, Sholto Douglas, Bernd Bohnet, Fabio Pardo, Seth Odoom, Mihaela Rosca, Cicero Nogueira dos Santos, Kedar Soparkar, Arthur Guez, Tom Hudson, Steven Hansen, Chulayuth Asawaroengchai, Ravichandra Addanki, Tianhe Yu, Wojciech Stokowiec, Mina Khan, Justin Gilmer, Jaehoon Lee, Carrie Grimes Bostock, Keran Rong, Jonathan Caton, Pedram Pejman, Filip Pavetic, Geoff Brown, Vivek Sharma, Mario Luvci'c, Rajkumar Samuel, Josip Djolonga, Amol Mandhane, Lars Lowe Sjosund, Elena Buchatskaya, Elspeth White, Natalie Clay, Jiepu Jiang, Hyeontaek Lim, Ross Hemsley, Jane Labanowski, Nicola De Cao, David Steiner, Sayed Hadi Hashemi, Jacob Austin, Anita Gergely, Tim Blyth, Joe Stanton, Kaushik Shivakumar, Aditya Siddhant, Anders Andreassen, Carlos L. Araya, Nikhil Sethi, Rakesh Shivanna, Steven Hand, Ankur Bapna, Ali Khodaei, Antoine Miech, Garrett Tanzer, Andy Swing, Shantanu Thakoor, Zhufeng Pan, Zachary Nado, Stephanie Winkler, Dian Yu, Mohammad Saleh, Lorenzo Maggiore, Iain Barr, Minh Giang, Thais Kagohara, Ivo Danihelka, Amit Marathe, Vladimir Feinberg, Mohamed Elhawaty, Nimesh Ghelani, Dan Horgan, Helen Miller, Lexi Walker, Richard Tanburn, Mukarram Tariq, Disha Shrivastava, Fei Xia, Chung-Cheng Chiu, Zoe C. Ashwood, Khuslen Baatarsukh, Sina Samangooei, Fred Alcober, Axel Stjerngren, Paul Komarek, Katerina Tsihlas, Anudhyan Boral, Ramona Comanescu, Jeremy Chen, Ruibo

Liu, Dawn Bloxwich, Charlie Chen, Yanhua Sun, Fangxiaoyu Feng, Matthew Mauger, Xerxes Dotiwalla, Vincent Hellendoorn, Michael Sharman, Ivy Zheng, Krishna Haridasan, Gabriel Barth-Maron, Craig Swanson, Dominika Rogozi'nska, Alek Andreev, Paul Kishan Rubenstein, Ruoxin Sang, Dan Hurt, Gamaleldin Elsayed, Ren shen Wang, Dave Lacey, Anastasija Ili'c, Yao Zhao, Woohyun Han, Lora Aroyo, Chimezie Iwuanyanwu, Vitaly Nikolaev, Balaji Lakshminarayanan, Sadegh Jazayeri, Raphael Lopez Kaufman, Mani Varadarajan, Chetan Tekur, Doug Fritz, Misha Khalman, David Reitter, Kingshuk Dasgupta, Shourya Sarcar, T. Ornduff, Javier Snaider, Fantine Huot, Johnson Jia, Rupert Kemp, Nejc Trdin, Anitha Vijayakumar, Lucy Kim, Christof Angermueller, Li Lao, Tianqi Liu, Haibin Zhang, David Engel, Somer Greene, Anais White, Jessica Austin, Lilly Taylor, Shereen Ashraf, Dangyi Liu, Maria Georgaki, Irene Cai, Yana Kulizhskaya, Sonam Goenka, Brennan Saeta, Kiran Vodrahalli, Christian Frank, Dario de Cesare, Brona Robenek, Harry Richardson, Mahmoud Alnahlawi, Christopher Yew, Priya Ponnapalli, Marco Tagliasacchi, Alex Korchemniy, Yelin Kim, Dinghua Li, Bill Rosgen, Kyle Levin, Jeremy Wiesner, Praseem Banzal, Praveen Srinivasan, Hongkun Yu, cCauglar Unlu, David Reid, Zora Tung, Daniel F. Finchelstein, Ravin Kumar, Andre Elisseeff, Jin Huang, Ming Zhang, Rui Zhu, Ricardo Aguilar, Mai Gim'enez, Jiawei Xia, Olivier Dousse, Willi Gierke, Soheil Hassas Yeganeh, Damion Yates, Komal Jalan, Lu Li, Eri Latorre-Chimoto, Duc Dung Nguyen, Ken Durden, Praveen Kallakuri, Yaxin Liu, Matthew Johnson, Tomy Tsai, Alice Talbert, Jasmine Liu, Alexander Neitz, Chen Elkind, Marco Selvi, Mimi Jasarevic, Livio Baldini Soares, Albert Cui, Pidong Wang, Alek Wenjiao Wang, Xinyu Ye, Krystal Kallarackal, Lucia Loher, Hoi Lam, Josef Broder, Daniel Niels Holtmann-Rice, Nina Martin, Bramandia Ramadhana, Daniel Toyama, Mrinal Shukla, Sujoy Basu, Abhi Mohan, Nicholas Fernando, Noah Fiedel, Kim Paterson, Hui Li, Ankush Garg, Jane Park, Donghyun Choi, Diane Wu, Sankalp Singh, Zhishuai Zhang, Amir Globerson, Lily Yu, John Carpenter, Félix de Chaumont Quitry, Carey Radebaugh, Chu-Cheng Lin, Alex Tudor, Prakash Shroff, Drew Garmon, Dayou Du, Neera Vats, Han Lu, Shariq Iqbal, Alexey Yakubovich, Nilesh Tripuraneni, James Manyika, Haroon Qureshi, Nan Hua, Christel Ngani, Maria Abi Raad, Hannah Forbes, Anna Bulanova, Jeff Stanway, Mukund Sundararajan, Victor Ungureanu, Colton Bishop, Yunjie Li, Balaji Venkatraman, Bo Li, Chloe Thornton, Salvatore Scellato, Nishesh Gupta, Yicheng Wang, Ian Tenney, Xihui Wu, Ashish Shenoy, Gabriel Carvajal, Diana Gage Wright, Ben Bariach, Zhuyun Xiao, Peter Hawkins, Sid Dalmia, Cl'ement Farabet, Pedro Valenzuela, Quan Yuan, Christoper A. Welty, Ananth Agarwal, Mianna Chen, Wooyeol Kim, Brice Hulse, Nandita Dukkipati, Adam Paszke, Andrew Bolt, Elnaz Davoodi, Kiam Choo, Jennifer Beattie, Jennifer Prendki, Harsha Vashisht, Rebeca Santamaria-Fernandez, Luis C. Cobo, Jarek Wilkiewicz, David Madras, Ali Elqursh, Grant Uy, Kevin Ramirez, Matt Harvey, Tyler Liechty, Heiga Zen, Jeff Seibert, Clara Huiyi Hu, A. Ya. Khorlin, Maigo Le, Asaf Aharoni, Megan Li, Lily Wang, Sandeep Kumar, Alejandro Lince, Norman Casagrande, Jay Hoover, Dalia El Badawy, David Soergel, Denis Vnukov, Matt Miecnikowski, Jiri Simsa, Anna Koop, Praveen Kumar, Thibault Sellam, Daniel Vlasic, Samira Daruki, Nir Shabat, John Zhang, Guolong Su, Kalpesh Krishna, Jiageng Zhang, Jeremiah Liu, Yi Sun, Evan Palmer, Alireza Ghaffarkhah, Xi Xiong, Victor Cotruta, Michael Fink, Lucas Dixon, Ashwin Sreevatsa, Adrian Goedeckemeyer, Alek Dimitriev, Mohsen Jafari, Remi Crocker, Nicholas Fitzgerald, Aviral Kumar, Sanjay Ghemawat, Ivan Philips, Frederick Liu, Yannie Liang, Rachel Sterneck, Alena Repina, Marcus Wu, Laura Knight, Marin Georgiev, Hyo Lee, Harry Askham, Abhishek Chakladar, Annie Louis, Carl Crous, Hardie Cate, Dessie Petrova, Michael Quinn, Denese Owusu-Afriyie, Achintya Singhal, Nan Wei, Solomon Kim, Damien Vincent, Milad Nasr, Christopher A. Choquette-Choo, Reiko Tojo, Shawn Lu, Diego de Las Casas, Yuchung Cheng, Tolga Bolukbasi, Katherine Lee, Saaber Fatehi, Rajagopal Ananthanarayanan, Miteyan Patel, Charbel El Kaed, Jing Li, Jakub Sygnowski, Shreyas Rammohan Belle, Zhe Chen, Jaclyn Konzelmann, Siim Poder, Roopal Garg, Vinod Koverkathu, Adam Brown, Chris Dyer, Rosanne Liu, Azade Nova, Jun Xu, Slav Petrov, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, and Oriol Vinyals. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *ArXiv*, abs/2403.05530.

Cy Shih. 2013. Translators' end-revision processing patterns and maxims: a think-aloud protocol study. *Arab World English Journal*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NeurIPS 2022, Red Hook, NY, USA. 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*, abs/2405.11804.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023a. A paradigm shift in machine translation: Boosting translation performance of large language models. *ArXiv*, abs/2309.11674.

Wenda Xu, Daniel Deutsch, Mara Finkelstein, Juraj Juraska, Biao Zhang, Zhongtao Liu, William Yang Wang, Lei Li, and Markus Freitag. 2023b. Llmrefine: Pinpointing and refining large language models via fine-grained actionable feedback.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

A Appendices

A.1 Results on Shorter Documents

Table 9 presents automatic evaluation results of step-by-step on shorter documents, where segments are grouped together such that they do not exceed a token limit of 150 white-space separated tokens. The dataset statistics are presented in Table 10. We observe the same trends with the ones reported with larger documents in §5.

A.2 Results on ChrF

Tables 7 and 8 report ChrF scores on WMT 2023 and WMT 2024, respectively. As anticipated with string-based metrics, LLM translations which prioritize fluency receive lower scores compared to those that are by construction instructed to be closer to the source text. This behavior is in line with observations of prior work that employ similar human-like translation strategies with LLMs (Wu et al., 2024).

A.3 Prompts

Tables 11 and 12 present the complete prompts we used for our translate step-by-step framework and baselines. It has come to our attention that the prompts used in the experiments contain a few typographical errors. Preliminary results using revised prompts show comparable, if not slightly improved results (in the range of $0.1 - 0.2$ MetricX-23 score points), across all steps.

A.4 More example outputs

Table 13 gives more example outputs to support the discussion in §6.

	Research	*Draft*	*Refinement*	*Proofreading*	ZH	UK	RU	JA	HE	CS	DE	AVERAGE
1.	○	○	○	○	48.04	61.85	63.55	38.75	64.03	67.62	71.81	59.38
2.	○	●	○	○	48.69 ↑ 0.65	61.81 ↓ 0.04	63.93 ↑ 0.38	39.00 ↑ 0.25	64.68 ↑ 0.65	67.63 ↑ 0.01	71.79 ↓ 0.02	59.65
3.	○	○	●	○	41.48 ↓ 6.56	59.44 ↓ 2.41	59.33 ↓ 4.22	36.19 ↓ 2.56	60.26 ↓ 3.77	63.44 ↓ 4.18	66.89 ↓ 4.92	55.29
4.	○	●	●	○	43.14 ↓ 4.90	59.58 ↓ 2.27	60.37 ↓ 3.18	37.45 ↓ 1.30	60.92 ↓ 3.11	63.04 ↓ 4.58	68.71 ↓ 3.10	56.17
5.	●	●	○	○	45.98 ↓ 2.06	61.51 ↓ 0.34	63.04 ↓ 0.51	39.30 ↑ 0.55	62.89 ↓ 1.14	67.17 ↓ 0.45	71.07 ↓ 0.74	58.71
6.	●	●	●	○	41.03 ↓ 7.01	58.72 ↓ 3.13	59.44 ↓ 4.11	37.65 ↓ 1.10	59.91 ↓ 4.12	63.02 ↓ 4.60	67.61 ↓ 4.20	55.34
7.	●	●	●	●	40.71 ↓ 7.33	58.78 ↓ 3.07	59.23 ↓ 4.32	37.51 ↓ 1.24	59.65 ↓ 4.38	63.11 ↓ 4.51	67.49 ↓ 4.32	55.21

Table 7: ChrF evaluation results of translate step-by-step and its ablation variants on the WMT 2023 development datasets. Filled dots indicate active steps in the pipeline, while unfilled dots represent ablated steps. When all steps are ablated, the system defaults to zero-shot translation

	DE	ES	ZH	RU	UK	JA	HI	IS	CS	AVERAGE
Zero-shot	65.48	72.96	44.21	55.51	59.90	39.75	55.94	53.23	60.81	56.42
Research & Drafting	64.67 ↓ 0.81	72.30 ↓ 0.66	42.73 ↓ 1.48	57.30 ↑ 1.79	60.06 ↑ 0.16	41.19 ↑ 1.44	56.16 ↑ 0.22	53.09 ↓ 0.14	60.31 ↓ 0.50	56.42
Refinement	61.72 ↓ 3.76	69.22 ↓ 3.74	38.26 ↓ 5.95	55.09 ↓ 0.42	57.25 ↓ 2.65	39.15 ↓ 0.60	52.60 ↓ 3.34	52.62 ↓ 0.61	57.29 ↓ 3.52	53.69
Proofreading	61.62 ↓ 3.86	69.04 ↓ 3.92	38.41 ↓ 5.80	54.96 ↓ 0.55	57.14 ↓ 2.76	38.87 ↓ 0.88	53.47 ↓ 2.47	52.32 ↓ 0.91	56.98 ↓ 3.83	53.65

Table 8: ChrF results comparing step-by-step with zero-shot performance on the WMT 2024 test datasets.

	DE	ES	ZH	RU	UK	JA	HI	IS	CS	AVERAGE
										Ref-based
Zero-shot	1.89	3.10	3.25	2.90	2.99	2.31	3.03	3.79	2.37	2.85
Research & Drafting	1.67 ↓ 0.23	2.61 ↓ 0.49	2.80 ↓ 0.45	2.53 ↓ 0.37	2.67 ↓ 0.32	1.91 ↓ 0.40	2.00 ↓ 1.03	3.45 ↓ 0.34	2.16 ↓ 0.21	2.42
Refinement	1.44 ↓ 0.45	2.20 ↓ 0.90	2.33 ↓ 0.92	2.17 ↓ 0.73	2.34 ↓ 0.65	1.61 ↓ 0.70	1.62 ↓ 1.41	3.02 ↓ 0.76	2.00 ↓ 0.37	2.08
Proofreading	1.36 ↓ 0.53	2.11 ↓ 0.99	2.28 ↓ 0.97	2.20 ↓ 0.69	2.27 ↓ 0.72	1.65 ↓ 0.66	1.60 ↓ 1.43	3.04 ↓ 0.75	1.98 ↓ 0.39	2.05
										QE-based
Zero-shot	1.81	2.34	2.13	1.67	1.96	1.26	1.98	3.15	1.85	2.02
Research & Drafting	1.62 ↓ 0.18	2.03 ↓ 0.31	1.85 ↓ 0.28	1.40 ↓ 0.26	1.60 ↓ 0.36	1.10 ↓ 0.15	1.40 ↓ 0.58	2.93 ↓ 0.21	1.73 ↓ 0.12	1.74
Refinement	1.24 ↓ 0.57	1.61 ↓ 0.73	1.51 ↓ 0.62	1.05 ↓ 0.62	1.17 ↓ 0.79	0.91 ↓ 0.35	0.96 ↓ 1.01	2.37 ↓ 0.78	1.31 ↓ 0.53	1.35
Proofreading	1.12 ↓ 0.68	1.54 ↓ 0.80	1.44 ↓ 0.69	0.99 ↓ 0.67	1.10 ↓ 0.86	0.88 ↓ 0.38	0.92 ↓ 1.06	2.24 ↓ 0.90	1.22 ↓ 0.62	1.27

Table 9: MetricX-23 evaluation results comparing step-by-step with zero-shot performance on the WMT 2024 test datasets, where each document has a maximum length of 150 tokens. *Translate step-by-step surpasses zero-shot, with each step incrementally improving translation quality.*

Domain	Literary	News	Social	Speech
# Docs.	66	73	75	112
Avg. Length	120	110	105	72

Table 10: Per-domain statistics for WMT 2024, when blobbing with 150 max for total of 327 docs.

PRE-TRANSLATION RESEARCH

You will be asked to translate a piece of text form English into Chinese following the five stages of the translation process. Here is the context in which the text appears:

Context: *placeholder 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 category:

* **Idiomatic Expressions:**
    * Identify idiomatic expressions that cannot be directly translated word-for-word into Chinese.

DRAFTING

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.

Give your best one translation for the following piece of text based on the pre-drafting analysis without providing alternatives:

English: *placeholder source text*

REFINEMENT

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.

PROOFREADING

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 you reference, below are the source text, the draft, and refined translations.

Source Text
placeholder source text

Draft Translation
placeholder draft translation

Refined Translation
placeholder draft refined translation

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.

Table 11: Complete prompts used by the translate step-by-step pipeline.

ZERO-SHOT

You are asked to translate the text below into Chinese. Please output only the translation of the text without any other explanation.

English: *placeholder source text*
Chinese:

ZERO-SHOT IN CONTEXT

You are asked to translate the text below into Chinese. You are also given access to the context it appears.

Context: *placeholder document context*

Please output only the translation of the text without any other explanation.

English: *placeholder source text*
Chinese:

DRAFT JSON

Analyze the previous responses and create a JSON object that organizes the linguistic information they contain. This object should have two sections: “idiomatic_expressions", and “draft_translation":

- **“idiomatic_expressions"**:
  - This section should also be a list of dictionaries.
  - Each dictionary represents a phrase and has the keys: “source_phrase", “description", “translation", and “literal_translation".
  - The “translation" key should hold a list of all provided translations for the phrase.
  - If the response doesn’t provide a literal translation, use ‘null‘ for the “literal_translation" value.
  - If the response doesn’t identify relevant idiomatic expressions use ‘null‘ for the corresponding value.
  - Don’t include transliterated versions of the translations.
- **“draft_translation"**:
  - A string containing the draft translation.
  - If a translation includes multiple alternatives for the same source text, extract the first one.
  - Don’t include transliterated versions of the translations.
  - If the translation includes alternatives separated by "/", extract the first one.

Table 12: Prompts used for baseline comparisons (zero-shot and zero-shot in context) and structuring the pre-translation research translation artifacts.

1. ✔*Good example*	
SOURCE	@user26 cheeked up. I love HER*
REFERENCE	@user26 口出狂言。我喜欢她* *Talking nonsense. I like her**
ZERO-SHOT	@user26 脸红了。我爱她* *Blushing. I love her**
RESEARCH AND DRAFT	@user26 脸颊做得很饱满。我喜欢她* *The cheeks are very plump. I like her**
REFINEMENT	@user26 脸颊做得鼓鼓的，超可爱！我喜欢她* *The cheeks are very chubby, super cute! I like her**
NOTES	*Both the reference and the zero-shot misinterpreted the phrase "cheeked up" in the context of making miniatures. Both the draft and the refined version understood it correctly, with the latter being a bit more colloquial.*
2. ✔*Good example*	
SOURCE	I donked up my threading
REFERENCE	我完成了这部分学习 *I finished this part of study.*
ZERO-SHOT	我把我的线程搞砸了 *I messed up my threads [note: a terminology in computing].*
RESEARCH AND DRAFT	我搞砸了我的帖子顺序 *I messed up the order of my posts.*
REFINEMENT	我搞砸了帖子顺序 *I messed up the post order.*
NOTES	*The reference misinterpreted both "donked up" and "threading" in this context, and the zero-shot translation mistook the latter for the computing terminology.*
2. ✔*Good example*	
SOURCE	It is a ridiculous level.
REFERENCE	这关太难了。 *This level is too hard.*
ZERO-SHOT	这是一个荒谬的关卡。 *This is a ridiculous level.*
RESEARCH AND DRAFT	这是一个荒谬的关卡。 *This is a ridiculous level.*
REFINEMENT	这个关卡真是太难了。 *This level is really too hard.*
NOTES	*Zero-shot and draft translations are direct, literal, and stylistically awkward in Chinese, especially in the social media domain.*
4. ✗ *Bad example*	
SOURCE	Ivory chuckled from the memory.
REFERENCE	艾弗里回忆到这里，笑了起来。 *Ivory thought of this and laughed*
ZERO-SHOT	艾弗里想起这件事，不禁轻笑起来。 *Ivory thought of this and couldn't help starting to chuckle.*
RESEARCH AND DRAFT	艾弗里回想起当时的情景，不禁轻笑出声 *Ivory thought of the situation at that time and couldn't help letting out a chuckle.*
REFINEMENT	艾弗里摸了摸肩膀上上次留下的瘀伤，不禁轻笑出声。 *Ivory touched the bruise on the shoulder left from the last time and couldn't help letting out a chuckle.*
NOTES	*All translations did a good job translating the phrase "chuckled from the memory" fluently into Chinese, but the refinement hallucinated by adding "touched the bruise."*
5. ✗ *Bad example*	
SOURCE	one of them ran for something and grabbed it
REFERENCE	其中一个人跑过去抓起了什么东西 *One of them ran over and grabbed something.*
ZERO-SHOT	其中一个人跑去拿了什么东西 *One of them ran and grabbed something*
RESEARCH AND DRAFT	其中一个人跑到某个东西那里，抓起它 *One of them ran to something and grabbed it.*
REFINEMENT	其中一个人跑到某个东西那里，抓起它 *One of them ran to something and grabbed it.*
NOTES	*Although all translations are adequate and capture the semantic meaning of the source correctly, the draft and the refinement keep the original source structure and the resulting translation sounds like translationese.*

Table 13: Sample of step-by-step and zero-shot outputs along with notes.