Automatic Input Rewriting Improves Translation with Large Language Models

Dayeon Ki
University of Maryland
dayeonki@umd.edu

Marine Carpuat
University of Maryland
marine@cs.umd.edu

Abstract

Can we improve machine translation (MT) with LLMs by rewriting their inputs automatically? Users commonly rely on the intuition that well-written text is easier to translate when using off-the-shelf MT systems. LLMs can rewrite text in many ways but in the context of MT, these capabilities have been primarily exploited to rewrite outputs via post-editing. We present an empirical study of 21 input rewriting methods with 3 open-weight LLMs for translating from English into 6 target languages. We show that text simplification is the most effective MT-agnostic rewrite strategy and that it can be improved further when using quality estimation to assess translatability. Human evaluation further confirms that simplified rewrites and their MT outputs both largely preserve the original meaning of the source and MT. These results suggest LLM-assisted input rewriting as a promising direction for improving translations.[1]

1 Introduction

Machine translation (MT) users and developers have long exploited the idea that some texts are easier to translate than others. For instance, guiding people to edit their inputs so that they are well formed is a cornerstone of MT literacy courses (Bowker, 2021; Steigerwald et al., 2022), and adopting plain language has been shown to improve the readability of translated health content (Rossetti, 2019). In MT research, a wealth of studies have considered pre-processing strategies to rewrite inputs, particularly for statistical MT (Xia and McCord, 2004; Callison-Burch et al., 2006; Štajner and Popovic, 2016).

The growing use of Large Language Models (LLMs) for translation leads us to revisit the impact of rewriting inputs on MT. On the one hand, rewriting inputs for LLM translation aligns with the re-framing of MT as a multi-step process (Briakou et al., 2024a). LLMs have shown promise in rewriting MT outputs (Zeng et al., 2024; Ki and Carpuat, 2024; Xu et al., 2024), and can rewrite text according to various style specifications (Raheja et al., 2023; Hallinan et al., 2023; Shu et al., 2024; Krishna et al., 2024). On the other hand, current models might already be robust to input variability, since they are trained on vast amounts of heterogeneous data (Touvron et al., 2023), fine-tuned on diverse tasks (Raffel et al., 2020; Alves et al., 2024) and operate at a much higher quality level compared to the statistical MT systems used in previous pre-processing studies.

How should inputs be rewritten for MT? The assumption that well-written texts are easier to translate drives recommendations for MT literacy, as well as the use of paraphrasing (Callison-Burch et al., 2006; Mirkin et al., 2009; Marton et al., 2009; Aziz et al., 2010) and simplification (Štajner and Popovic, 2016; Štajner and Popović, 2019). However, can we more directly rewrite inputs so that they are easier to translate? Generic translatability has been defined as "a measurement of the time and effort it takes to translate a text" (Kumhyr et al., 1994). Uchimoto et al. (2005) introduced a metric to quantify MT translatability based on back-translation of MT hypotheses in the source language. Given recent progress in quality estimation (Fernandes et al., 2023; Naskar et al., 2023; Tomani et al., 2024), we propose instead to use reference-free quality estimation scores as a measure of translatability.

We thus ask the following research questions:

(1) Can we improve MT quality from LLMs by rewriting inputs for style?

(2) Do quality estimation metrics provide useful translatability signals for input rewriting?

We conduct an empirical study with 3 open-weight LLMs for a total of 21 input rewriting methods with varying levels of MT-awareness on

[1]We release our code and dataset at https://github.com/dayeonki/rewrite_mt.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 10829–10856

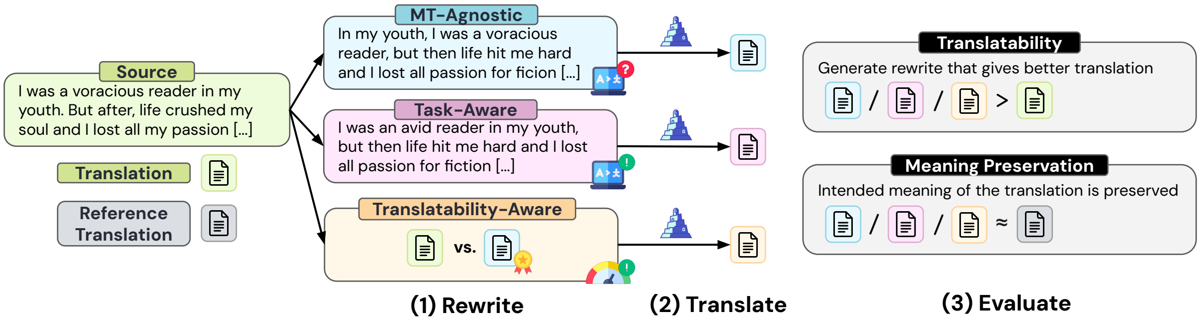

Figure 1: Overview of the rewriting pipeline. **(1) Rewrite:** Given source sentence, we generate rewrites using different rewriting methods: MT Agnostic, Task-Aware and Translatability-Aware. **(2) Translate:** We translate each rewrite using our MT system, TOWER-INSTRUCT 7B. **(3) Evaluate:** We automatically evaluate rewrites along translatability, meaning preservation, and overall translation quality.

translation from English into German, Russian and Chinese, and we further evaluate the generalizability of our best performing approach on translation from English into Czech, Hebrew and Japanese (§4.4). Our results show that simple **MT-Agnostic rewrites** obtained by prompting LLMs to simplify, paraphrase, or change the style of the input, improve translatability, and that simplification most reliably improves translation quality. Interestingly, these MT-agnostic rewrites are more effective than **Task-Aware rewrites**, where LLMs are prompted to rewrite inputs for the purpose of MT (§4.1). Finally, using quality estimation signals to assess **translatability** at the segment level and select when to use rewrites further improves MT quality, outperforming more expensive fine-tuning strategies (§4.2). Human evaluation further confirms that simplified rewrites and their MT largely preserve the original meaning of the source and MT (§5.3).

2 Input Rewriting Methods

Within the process of source rewriting, the goal of a rewrite model is to rewrite the original source sentence s into another form that is easier to translate while preserving its intended meaning. For **MT-Agnostic** rewriting methods (§2.1), which lacks translation-related knowledge, the rewrite model $\mathcal{M}_\theta$ can rewrite s into s':

$$s' = \mathcal{M}_\theta(s) \tag{1}$$

On the contrary, both **Task-Aware** (§2.2) and **Translatability-Aware** (§2.3) rewriting methods incorporate some translation signal. For Task-Aware, $\mathcal{M}_\theta$ rewrites s with the information of the end-task (MT):

$$s' = \mathcal{M}_\theta(s, \text{MT task}) \tag{2}$$

For Translatability-Aware method, it rewrites with the knowledge of segment level quality estimation scores between source and the output of a specific MT system MT(t):

$$s' = \mathcal{M}_\theta(s, \text{xCOMET}(s, \text{MT}(t))) \tag{3}$$

Figure 1 shows the overview of our proposed rewriting pipeline. To find the most effective $\mathcal{M}_\theta$, we test a total of 21 input rewriting methods.

2.1 MT-Agnostic Rewriting

MT-agnostic rewriting methods reflect various a priori assumptions on what makes text easier to translate. They do not take as input any signal of translatability or knowledge about the end-task. We consider three prompting variants here, all inspired by prior works on source rewriting (Mirkin et al., 2009, 2013; Štajner and Popovic, 2016).

Simplification. Simplification includes replacing complex words with simpler ones, rephrasing complex syntactic structures, and shortening sentences (Chandrasekar and Bangalore, 1997; Feng, 2008). Prior works show that simplified inputs are more conducive to MT, and particularly improve the fluency of MT outputs (Štajner and Popović, 2019).

Paraphrase. Paraphrases are alternative ways of expressing the same information within one language, which can help resolve unknown or complex words (Callison-Burch et al., 2006). Paraphrasing with LLMs might benefit MT by normalizing inputs using language patterns that are more frequent in LLM training data. Further, some LLMs, such as TOWER (Alves et al., 2024), are fine-tuned on both paraphrasing and MT tasks, and might thus produce paraphrases that are useful for MT.

Stylistic. We employ an off-the-shelf text editing tool COEDIT-XL (Raheja et al., 2023) to rewrite inputs according to diverse style specifications:

- **Grammar**: Fix the grammar.
- **Coherent**: Make the text more coherent.
- **Understandable**: Make it easier to understand.
- **Formal**: Rewrite the text more formally.

These operationalize the assumption that well-formed text is easier to translate. All prompt templates are shown in Appendix Table 5.

2.2 Task-Aware Rewriting

For task-aware rewriting methods, we design prompts that account for the fact that rewrites are aimed at MT. Prior work has shown that LLMs can post-edit errors in MT outputs (Ki and Carpuat, 2024; Zeng et al., 2024; Treviso et al., 2024a; Xu et al., 2024; Briakou et al., 2024b), raising the question of whether this ability can be extended to rewriting inputs to enhance translatability. Additionally, TOWER-INSTRUCT has been jointly trained on paraphrasing, grammatical error correction (GEC), and translation tasks, suggesting it may be well-suited for performing translatability rewrites in a zero-shot fashion. We consider two prompting strategies (Refer to Appendix Table 5 for exact templates):

Easy Translation. We prompt LLMs to rewrite inputs in a way that specifically facilitates translation into the target language.

Chain of Thought Rewrite+Translate. We use a Chain of Thought (Wei et al. (2023), CoT) style prompt where LLMs are prompted to handle the entire rewriting and translation process in one sequence of CoT instructions within a single model.

2.3 Translatability-Aware Rewriting

We propose to use quality estimation scores for a given input and output pair to assess the translatability of inputs at the segment level. This makes it possible to inject translatability signals at inference or training time. We introduce a lightweight inference-time selection strategy, and contrast it against a more expensive fine-tuning approach.

Inference-Time Selection. Input segments might not benefit from rewriting uniformly, since the quality of the original inputs and of their rewrites might vary. We thus propose to use translatability scores to decide whether or not to replace the original input with a rewrite at inference time. We use the state-of-the-art XCOMET quality estimation tool (Guerreiro et al., 2024) to assess how good the translation t' of a rewrite s' is: $\text{XCOMET}(s', t')$. We compare this score with the estimated quality of the translation t of the original source s, choosing to use the rewrite if $\text{XCOMET}(s', t') > \text{XCOMET}(s, t)$, and keeping the original source otherwise. This straightforward approach allows us incorporate translatability signals at inference time, with little additional cost.

Supervised Fine-tuning. The translatability-based selection process described above for inference could also be used to gather examples of good rewrites and enable instruction fine-tuning of models to rewrite text for improved translation. While designing an optimal approach for this task is out of scope for this work, we wish to compare our inference-time selection strategy with a straightforward training strategy. We construct a fine-tuning dataset of positive rewrite examples $\mathcal{D}_{pos}$, as follows: for a given input s, we generate rewrites using all MT-agnostic methods. We add to our training set the rewrites that improve translatability as measured by $\text{XCOMET}(s', t') > \text{XCOMET}(s, t)$. The base LLM is then instruction fine-tuned based to rewrite input s so that it is better translated, using s' as supervision. Detailed prompt templates are shown in Appendix A.1.

3 Experimental Setup

3.1 Model & Data

MT System. We use TOWER-INSTRUCT 7B as our MT system for all our experiments since it is specifically trained for translation-related tasks and has demonstrated superior MT performance compared to other LLMs (Alves et al., 2024).

Rewriting Models. For prompting experiments, we use 7B variant of three open-weight LLMs in zero-shot setting: LLAMA-2 (Touvron et al., 2023) – the base model for TOWER-INSTRUCT, LLAMA-3 (Grattafiori et al., 2024) – more recent multilingual model compared to LLAMA-2, and TOWER-INSTRUCT (Alves et al., 2024) – the same LLM as used for our MT system. For supervised

Language	Type	Prompt/Model	xCOMET(s,t)	xCOMET(s,t,r)	METRICX(s,t)	METRICX(t,r)
EN-DE	**Original**	-	0.893	0.898	2.038	1.534
	MT-Agnostic	Simplification (TOWER)	**0.915**	**0.907**	**1.504***	1.519
		Paraphrase (DIPPER)	**0.904**	0.838	**1.674**	2.757
	Task-Aware	Easy translate (TOWER)	**0.901**	0.903	**1.759**	2.427
		CoT (TOWER)	**0.907**	0.897	**1.892**	1.578
	Translatability-Aware	Selection	**0.921***	**0.922***	**1.734**	**1.461***
		Fine-tune (Ref)	**0.896**	0.876	2.023	2.028
EN-RU	**Original**	-	0.872	0.868	2.535	2.028
	MT-Agnostic	Simplification (TOWER)	**0.921***	**0.891**	**1.135**	**1.921**
		Paraphrase (DIPPER)	**0.904**	0.821	**1.249**	3.476
	Task-Aware	Easy translate (LLaMA-3)	**0.917**	**0.881**	**0.801***	10.401
		CoT (TOWER)	**0.903**	0.875	2.432	2.024
	Translatability-Aware	Selection	**0.914**	**0.899***	**2.096**	**1.830***
		Fine-tune (Ref)	**0.894**	0.866	**2.284**	2.012
EN-ZH	**Original**	-	0.786	0.794	3.445	2.282
	MT-Agnostic	Simplification (TOWER)	**0.821**	**0.802**	**1.521***	2.227
		Paraphrase (DIPPER)	**0.813**	0.722	**1.583**	4.009
	Task-Aware	Easy translate (LLaMA-3)	**0.793**	0.791	**1.618**	7.650
		CoT (TOWER)	**0.821**	0.771	**3.321**	2.432
	Translatability-Aware	Selection	**0.823***	**0.819***	**3.149**	**2.206***

Table 1: Results using different rewriting methods. Statistically significant average improvements (p-value < 0.05) are **bold**. Best scores for each metric is **bold** with *. xCOMET(s,t): translatability (↑); xCOMET(s,t,r): overall translation quality (↑); METRICX(s,t): quality estimation (↓); METRICX(t,r): reference-based metric (↓). We substitute s and t to s' and t' when computing scores for rewrites. For each rewriting type, we show the best and worst of each methods based on xCOMET(s,t,r). We abbreviate TOWER-INSTRUCT as TOWER and DIPPER (L80/O60) as DIPPER due to space constraints. Full results are in Appendix B.1.

fine-tuning, we draw training samples from the English-German and English-Russian subset from WMT-20, 21, and 22 General MT task datasets (Freitag et al., 2021)[2], and provide detailed parameter settings in Appendix A.2.

Test Data. We use the WMT-23 General MT task[3] from the TOWEREVAL dataset[4] to guarantee that it was held out from the various training stages. We focus on translation from English into German (EN-DE), Russian (EN-RU) and Chinese (EN-ZH) for an extensive empirical comparison, and then test whether the most promising approaches generalize to translation from English into Czech (EN-CS), Hebrew (EN-HE) and Japanese (EN-JA). See Appendix Table 7 for data statistics.

3.2 Evaluation Metrics

We use xCOMET (Guerreiro et al., 2024) and METRICX (Juraska et al., 2023) to evaluate different aspects of rewrite quality. Specifically, we use xCOMET-XL[5] and METRICX-23-XL.[6] Higher scores indicate better performance for xCOMET, while lower scores are better with METRICX.

Translatability. We quantify translatability with the quality estimation score for a specific input–output pair (xCOMET(s',t') or METRICX-QE(s',t')). A rewrite s' of the original input s is considered easier to translate if xCOMET(s',t') is higher than xCOMET(s,t).

Meaning Preservation. We do not want rewrites that are easier to translate at the expense of changing the original meaning. Our meaning preservation metric evaluates how well the rewrite maintains the intended meaning of the translation as represented by the reference (Graham et al., 2015). We use a reference-based metric as opposed to using the semantic similarity between s and s' because it abstracts the meaning away from the specific formulation of s, reducing overfitting. We compute xCOMET scores between the rewrites and reference translations (xCOMET(s',r)). The desired behavior is to minimize the deterioration in xCOMET(s',r) compared to xCOMET(s,r).

[2] We do not consider English-Chinese pair here since this language pair is not supported in the dataset.

[3] https://www2.statmt.org/wmt23/translation-task.html

[4] https://huggingface.co/datasets/Unbabel/TowerEval-Data-v0.1

[5] https://huggingface.co/Unbabel/XCOMET-XL

[6] https://huggingface.co/google/metricx-23-xl-v2p0

Translation Quality. We additionally report the combined evaluation metric, xCOMET(s', t', r) to take into account of the trade-off between the two above metrics, and METRICX(t', r) which also assesses translation quality of the rewrite but is not informed by the updated source s'.

4 Results

We first extensively compare rewrite strategies focusing on the overall translation quality achieved by MT-Agnostic rewrites (§4.1) and Translatability-Aware rewrites (§4.2). To understand how rewrites change translations, we then analyze the trade-offs between translatability and meaning preservation (§4.3). Finally, we test whether the best-performing methods identified so far generalize to new language pairs (§4.4).

4.1 Simplifying Inputs Works Best

We first compare the MT Agnostic rewriting methods: simplification, paraphrasing, and stylistic edits. Due to space limits, we show the best and worst performing variations for each input rewriting method based on the overall translation quality metric xCOMET(s, t, r) for each language pair in Table 1. Full results are available in Appendix B.1.

Results show that all rewriting strategies improve translatability, but only **simplification** also improves the overall translation quality. Even the lowest performing rewrites reach higher translatability than the original baseline. Each method surpasses the baseline by up to 0.056 and 0.027 xCOMET(s, t) average scores for EN-DE, up to 0.058 and 0.036 average scores for EN-RU, and up to 0.054 and 0.028 average scores for EN-ZH pair. Trends are consistent with METRICX(s, t). However, making inputs easier to translate often degrades quality when comparing against references r. Simplification with TOWER-INSTRUCT distinguishes itself by improving translation quality based on xCOMET(s, t, r) scores and maintaining it according to the METRICX(t, r) scores – a harder metric to improve since the reference might be biased toward the original wording of the source.

Among the three LLMs used for simplification, TOWER-INSTRUCT achieves the best translation quality, while LLAMA-3 excels in translatability at the expense of meaning preservation. Interestingly, there is no benefit to using a separate LLM, even one fine-tuned specifically on paraphrasing or style edits such as DIPPER or COEDIT. Overall, the best performing method for MT-agnostic rewrites is simplification with TOWER-INSTRUCT, the same model we use as our MT system. We attribute this to TOWER-INSTRUCT being instruction fine-tuned on translation related tasks (but not simplification) and having more domain knowledge of the WMT dataset used in our evaluation.[7]

As shown in Table 1, simplifying with TOWER-INSTRUCT still holds the top spot when compared to Task-Aware rewriting methods, as indicated by higher xCOMET(s, t, r) scores. This suggests that injecting knowledge about the end-task (MT) to LLMs is less effective than simplifying inputs to improve translation quality.

Overall, these results confirm the intuition that simpler text is easier to translate, but establish that rewrites are not uniformly helpful for translation quality, motivating the need for more selective input rewriting strategies.

4.2 Selection via Translatability Improves MT

We evaluate the impact of inference-time selection based on translatability scores (*Selection* in Table 1), and compare it further with the more expensive supervised fine-tuning strategy (*Fine-tune*).

All language pairs consistently benefit from selection. Translation quality improves significantly, with average xCOMET(s, t, r) gains of 0.024 for EN-DE, 0.031 for EN-RU, and 0.025 for EN-ZH, marking the best performance among all variants. METRICX(t, r) scores confirm this trend, showing average improvements of 0.073 for EN-DE, 0.198 for EN-RU, and 0.076 for EN-ZH. At the segment level, rewrites are preferred to original inputs in 1197/1557 cases for EN-DE, 1610/2074 cases for EN-RU, and 2163/3074 cases for EN-ZH. Fine-tuning shows smaller gains compared to MT-Agnostic or Task-Aware methods, both in terms of translatability and translation quality, despite being more resource-intensive.

In summary, the results suggest that inference-time selection of inputs based on translatability scores is a promising strategy, outperforming MT-agnostic rewrites and rewrites obtained via a more expensive fine-tuning process.

4.3 Input Rewriting Trades Off Translatability and Meaning Preservation

We observe a moderate negative correlation between translatability and meaning preservation

[7] https://huggingface.co/datasets/Unbabel/TowerBlocks-v0.1

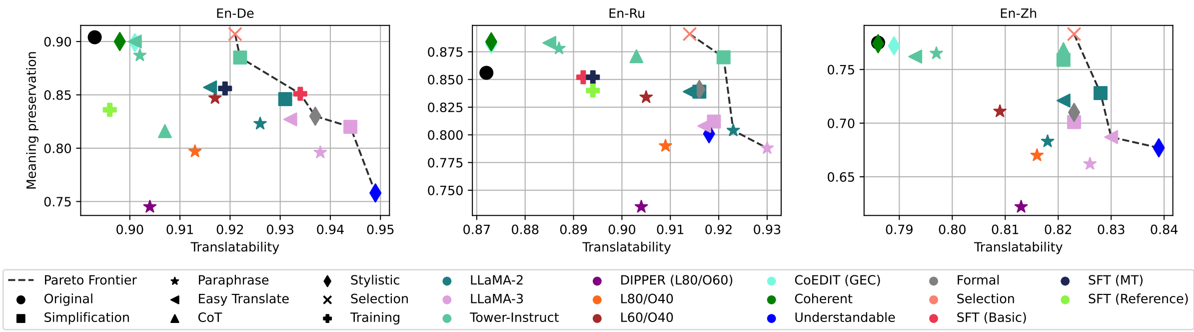

Figure 2: Pareto frontier per language pair. For each subplot, the x-axis is the translatability and y-axis is the meaning preservation scores. Pareto frontier (**dashed** line) visualizes the optimal solutions that take into account the trade-off between the two metrics. Each shape represents different rewriting methods and each color represent specific prompt or model variation.

scores, with Pearson coefficients of -0.48, -0.66, and -0.52 for EN-DE, EN-RU, and EN-ZH, respectively. This trade-off between the two metrics poses a Pareto optimization challenge: when a rewrite is easier to translate, it often results in lower meaning preservation. Therefore, we aim to find Pareto optimal solutions, which balance these trade-offs on a Pareto frontier (Huang et al., 2023).[8]

In Figure 2, we visualize our two objectives, translatability and meaning preservation, on each axis and identify the Pareto frontier. The results are consistent with the overall translation quality metric, XCOMET(s,t,r), where the scores for rewriting methods on the Pareto frontier are consistently the same as or on par with the original baseline. This also aligns with our earlier findings from comparing MT-Agnostic and Task-Aware rewrites (§4.1), where simplification with TOWER-INSTRUCT lies on the Pareto frontier for EN-DE and EN-RU. Even for EN-ZH, although this does not lie on the frontier, it has a higher XCOMET(s,t,r) score (0.802) than the original baseline (0.794). Furthermore, the best rewriting method according to XCOMET(s,t,r), translatability-based selection (§4.2), always lies on the Pareto frontier across all language pairs.

4.4 Best Input Rewriting Strategy Improves MT on Held-out Test sets

We evaluate whether the top methods that have emerged from the controlled empirical comparison conducted so far generalize to further test settings. As shown in Table 2, we test both simplification with TOWER-INSTRUCT (*Simplification*) and translatability-based input selection (*Selection*) on new test sets from the WMT-23 General MT task, English-Czech (EN-CS), English-Hebrew (EN-HE), and English-Japanese (EN-JA) to assess generalization to lower-resource target languages.

Language	Type	**X**(s,t)	**X**(s,t,r)	**M**(s,t)	**M**(t,r)
EN-CS	Original	0.646	0.655	5.376	4.493
	Simplification	0.691	0.675	4.684	4.333
	Selection	**0.736**	**0.718**	**4.152**	**3.663**
EN-HE	Original	0.327	0.320	16.66	15.48
	Simplification	0.351	0.332	15.97	15.43
	Selection	**0.389**	**0.363**	**15.39**	**14.51**
EN-JA	Original	0.746	0.718	3.514	2.688
	Simplification	0.789	0.738	2.957	2.508
	Selection	**0.826**	**0.769**	**2.781**	**2.273**

Table 2: Results of simplification and translatability-based selection for held-out test sets. We abbreviate XCOMET to **X** and METRICX to **M** due to space constraints. Best scores for each metric is **bold**.

Both simplification and translatability-based selection lead to progressive improvements in translation quality, as measured by XCOMET(s,t,r). Notably, the selection strategy tends to excel in language pairs with lower-resource target languages, showing translation quality gains of 0.064, 0.043, 0.051 scores for EN-CS, EN-HE, EN-JA, respectively, compared to increases of 0.017, 0.031, and 0.025 for EN-DE, EN-RU and EN-ZH. At the segment level, rewrites are also more preferred over original inputs, selected in 1395/2074 cases for EN-CS, 1309/2074 for EN-HE, and 1411/2074 for EN-JA. METRICX trends are consistent.

In sum, our findings generalize well to held-out test sets, further validating the effectiveness of the translatability-based selection strategy. This approach offers a practical and scalable solution for

[8]In Pareto optimization, Pareto optimal solutions are those where no single solution outperforms another in all tasks (Chen et al., 2024a). The set of Pareto optimal solutions forms the Pareto frontier.

input rewriting across a broader range of domains and language pairs, though there are many other dimensions that remain unexplored. We have conducted initial experiments with additional LLMs and source languages, shown in Appendix D.1 and D.2, which confirms our previous findings that simplification rewriting enhances translation quality. We leave a more comprehensive exploration of this direction for future work.

5 Analysis

5.1 Simplifying Inputs Improves MT Readability

Simplification as an input rewriting strategy can balance translatability and meaning preservation, leading to overall improvements in translation quality. We also examine whether this enhances the readability of both inputs and, subsequently, translation outputs. In Table 3, we present the Flesch Reading Ease score[9] and Gunning Fog index[10] to measure input readability, and the Vienna formula (WSTF) (Zowalla et al., 2023) and the Russian version of Flesch Readability test (Solnyshkina et al., 2018) to assess output readability for EN-DE and EN-RU, respectively.

As expected, input readability improves across all simplification methods, whether used in MT-Agnostic (LLAMA-2, LLAMA-3, and TOWER-INSTRUCT in Table 3) or Translatability-Aware (Selection in Table 3) manner. Interestingly, simplification not only leads to more readable input but also more readable outputs, with gains of up to 0.22 WSTF scores for EN-DE and 0.95 Flesch scores for EN-RU. We provide several qualitative examples in Appendix Tables 13 to 15 that illustrate how simplification rewrites can lead to varying degrees of readability improvements in both inputs and translation outputs.

5.2 Input Rewriting outperforms Post-editing

The symmetric task to input rewriting is post-editing, which focuses on improving and correcting errors in translation outputs. Can post-editing alone achieve the same improvements, or are both strategies *complementary*? To explore this, we compare input rewriting to post-editing by prompting

Language	Prompt/Model	Flesch	GFI	WSTF	Flesch-Ru
EN-DE	Original	60.79	10.56	1.35	-
	LLAMA-2	66.69	9.25	1.15	-
	LLAMA-3	64.00	9.98	1.24	-
	TOWER-INSTRUCT	**68.17**	**8.99**	**1.13**	-
	Selection	63.27	10.09	1.26	-
EN-RU	Original	69.93	9.91	-	65.67
	LLAMA-2	**74.73**	8.37	-	**66.62**
	LLAMA-3	72.88	9.20	-	66.36
	TOWER-INSTRUCT	74.14	**8.19**	-	65.40
	Selection	72.24	9.37	-	65.89
EN-ZH	Original	66.51	10.08	-	-
	LLAMA-2	71.64	8.74	-	-
	LLAMA-3	69.32	9.48	-	-
	TOWER-INSTRUCT	**72.22**	**8.42**	-	-
	Selection	68.41	9.68	-	-

Table 3: Input and output readability scores for simplification rewriting method. **Flesch**: Flesch Reading Ease score (↑); **GFI**: Gunning Fog Index (↓); **WSTF**: Vienna formula (↓); **Flesch-Ru**: Russian version of Flesch (↑).

Language	Type	$\mathbf{X}(s,t)$	$\mathbf{X}(s,t,r)$	$\mathbf{M}(s,t)$	$\mathbf{M}(t,r)$
EN-DE	Original	0.893	0.898	2.038	1.534
	I	**0.922**	**0.907**	**1.504**	1.519
	Owo	0.863	0.879	2.941	2.200
	Ow	0.879	0.894	2.515	1.858
	I+O	0.915	**0.907**	1.751	**1.502**
EN-RU	Original	0.861	0.854	2.535	2.028
	I	**0.921**	0.891	**1.135**	**1.921**
	Owo	0.868	0.864	2.815	2.384
	Ow	0.872	0.869	2.674	2.259
	I+O	0.917	**0.892**	1.632	2.045
EN-ZH	Original	0.786	0.794	3.445	**2.282**
	I	**0.821**	0.802	**1.521**	2.327
	Owo	0.713	0.751	5.585	4.262
	Ow	0.746	0.780	4.363	2.676
	I+O	0.818	**0.804**	3.335	2.323

Table 4: Results for input rewriting (**I**), post-editing output without source signal (**Owo**), with source signal (**Ow**), and the combination of both strategies (**I+O**). Best scores for each metric is **bold**. We use the same abbreviations for metrics as in Table 2.

TOWER-INSTRUCT[11] to simplify either inputs or outputs. As shown in Table 4, rewriting inputs (**I**) offers a notable advantage over post-editing outputs (**Owo**), even when post-editing is guided by the input sentence (**Ow**). Combining input rewriting and post-editing (**I+O**) yields the highest translation quality, though the difference compared to input rewriting alone is not statistically significant. This confirms that rewriting text for better translatability before translation plays a more decisive role than post-editing the output.[12]

[9]https://en.wikipedia.org/wiki/Flesch-Kincaid_readability_tests

[10]https://en.wikipedia.org/wiki/Gunning_fog_index

[11]We focus on TOWER-INSTRUCT as it is a multilingual LLM capable of rewriting in non-English target languages.

[12]We compare time and computational efficiency for input rewriting and output post-editing in Appendix F.

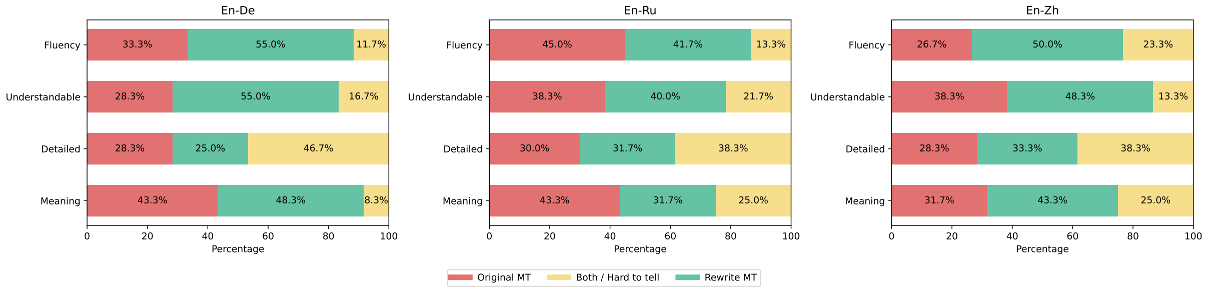

Figure 3: Win rates for human evaluation comparing Original MT vs. Rewrite MT across three language pairs (En-De, En-Ru, En-Zh) and four evaluation criteria: Fluency, Understandability, Level of detail (Detailed), and Meaning preservation relative to the reference translation.

5.3 Human Evaluation

Original MT vs. Rewrite MT. We conduct a manual evaluation to determine whether bilingual human annotators rate translations generated using our winning rewrite method (simplification with Tower-Instruct) as superior to the original translations. For each language pairs (En-De, En-Ru, En-Zh), we randomly select 20 pairs of instances, resulting in a total of 180 annotations from three annotators per pair. Inter-annotator agreement, measured by Fleiss' Kappa[13], is moderate, with values of 0.43, 0.39, and 0.51 for En-De, En-Ru, and En-Zh, respectively. For each instance, annotators are first provided with two translations and asked to evaluate on three axes: **1)** Fluency, **2)** Understandability, and **3)** Level of detail. Subsequently, we provide the reference translation, and annotators are asked to assess **4)** Meaning preservation. Annotators are also given the option to provide free form comments. Further details on the annotation set-up are available in Appendix E.1.

As illustrated in Figure 3, the human evaluation results confirm that translations from simplified inputs are rated as more fluent, understandable, and better at preserving the meaning of the reference translation. While this improvement is clear for the En-De and En-Zh pair, for En-Ru pair, annotators rate original MT as more fluent and more faithful to the original meaning.[14] Some En-Ru annotators who preferred the original MT noted that it often retained a more accurate sense of the words in the reference. In contrast, those who favored the simplified rewrite MT highlighted that translations are more contextually appropriate, easier to read, and more comprehensible than the original MT.

[13]https://en.wikipedia.org/wiki/Fleiss_kappa

[14]Note that the Fleiss's Kappa scores indicate that there is more disagreement between annotators for En-Ru pair.

Original vs. Rewrite. Our automatic meaning preservation metric evaluates the extent to which the original meaning is retained in the rewrite by comparing the rewritten source to the reference translation, rather than to the original source (Graham et al., 2015). Comparing to the original source is in the same language, but introduces a bias toward the original wording. On the other hand, comparing to the reference involves a cross-lingual comparison and is affected by unstable quality of references (Kocmi et al., 2022), but is less biased toward the original wording of the source.

To complement our automatic metric, we conduct a manual evaluation to assess how well the rewrites from simplification with Tower-Instruct preserve the meaning of the original source. We randomly sample 30 pairs of instances and collect three annotations per pair, totaling 90 annotations. Annotators are presented with both the original and rewritten sources and asked to evaluate how well the rewrite captures the meaning of the original source using a 4-point Likert scale (1: Does not capture meaning, 2: Partially, 3: Mostly, 4: Fully). Inter-annotator agreement by Fleiss' Kappa is 0.45. Of the 90 annotations, 55 were rated as 4, 27 as 3, 7 as 2, and 1 as 1, resulting an average score of 3.51. These results indicate that simplified rewrites generated by Tower-Instruct, although compared against the original source, still largely preserve the original meaning. Further details are provided in Appendix E.2.

6 Related Work

Rewriting with LLMs. Recent advances in LLMs have demonstrated impressive zero-shot capabilities in rewriting textual input based on user requirements (Shu et al., 2024). Most LLM-assisted rewriting tasks focus on query rewriting (Efthimiadis, 1996), which aims to reformulate text-based queries to enhance their representativeness and improve recall with retrieval-augmented LLMs (Mao et al., 2023; Zhu et al., 2024). Rewriting methods include prompting LLMs both as rewriters and rewrite editors (Ye et al., 2023; Kunilovskaya et al., 2024), and training LLMs as rewriters using feedback alignment learning (Ma et al., 2023; Mao et al., 2024). Another line of work focuses on style transfer, where the goal is to rewrite textual input into a specified style (Yuan et al., 2022; Hallinan et al., 2023). Our research aligns with efforts to rewrite texts with LLM assistance; however, unlike these works, we focus on rewriting source inputs to enhance MT quality.

Quality Estimation Metrics. The discrepancy between lexical-based metrics (e.g., BLEU (Papineni et al., 2002), CHRF (Popović, 2015)) and human judgments (Ma et al., 2019) has led to research in *neural* metrics. Particularly, quality estimation (QE) metrics, which compute a quality score for the translation conditioned only on the source sentence, have demonstrated benefits in improving MT quality. QE metrics are used for various purposes, including filtering out low-quality translations during training (Tomani et al., 2024), applying to post-editing workflows (Béchara et al., 2021), and providing feedback to users of MT systems (Mehandru et al., 2023). In our experiments, we use XCOMET as our main evaluation metric, as it shows the best correlation with human judgments (Agrawal et al., 2024). We primarily use XCOMET as a QE metric to compute translatability, further providing this information as knowledge to LLMs to improve MT quality.

Rewriting MT Outputs. The symmetric task of post-editing MT outputs has received significantly more attention than rewriting MT inputs. Most recent work relies on LLMs to automatically detect and correct errors in MT outputs using their internal knowledge (Raunak et al., 2023; Zeng et al., 2024; Chen et al., 2024b), with the help of external feedback (Ki and Carpuat, 2024; Xu et al., 2024) or through fine-tuning (Treviso et al., 2024b). In contrast, the task of rewriting MT inputs to make them more suitable for translation has been relatively underexplored with LLMs. While there have been some efforts in query rewriting and style transfer to improve retrieval (Mao et al., 2023; Zhu et al., 2024) and stylistic coherence (Ye et al., 2023; Hallinan et al., 2023), the specific application of LLMs to rewrite inputs for the purpose of enhancing MT quality is still emerging. Our research addresses this gap by focusing on the potential of LLM-assisted input rewriting to improve the translatability and quality of the resulting translations.

7 Conclusion

In this work, we studied the effectiveness of automatic input rewriting with LLMs in improving the quality of machine translation outputs. We explored a range of rewriting strategies with varying levels of MT-awareness: **1)** MT-Agnostic, **2)** Task-Aware (knowledge of the end-task), and **3)** Translatability-Aware rewrites (knowledge of translatability as measured with QE tools).

Our findings show that simpler texts are more translatable. However, MT-Agnostic rewrites do not uniformly help translation quality (§4.1), which motivates us to explore more selective strategies. Selecting inputs based on translatability scores during inference time further boosts translation quality (§4.2), addressing the Pareto optimization challenge by striking a balance between translatability and meaning preservation (§4.3). Analysis shows that simplifying inputs also results in more readable translation outputs (§5.1), and that input rewriting complements post-editing strategies (§5.2). Human evaluation complements our automatic metric by showing that both simplified rewrites and their corresponding MT largely preserve the original meaning of the source and MT (§5.3).

More broadly, this work suggests that LLM-assisted input rewriting is a promising direction for improving translations. The approaches introduced here represent a first step in this direction, and future work is needed to discover optimal rewriting strategies for a broader range of models. Furthermore, in line with growing research on LLM-based writing assistants (Lee et al., 2024), these results encourage future work on designing richer interactive approaches to translation with LLMs.

8 Limitations

We focus our investigation on TOWER-INSTRUCT 7B as our MT system, as it is an open-weight model. We exclude closed and larger models such as GPT-4[15] in the current experiments.

The scope of our study is also limited to out-of-English language pairs, as rewriting with LLMs has been more extensively studied in English (Ma et al.,

[15] https://openai.com/index/gpt-4/

2023; Ye et al., 2023; Shu et al., 2024; Mao et al., 2024), and using English as the source language benefits performance from its prevalence in LLM training data. One critical area of future research lies in developing rewriting tools that support a wider range of languages beyond English.

9 Acknowledgement

We thank the anonymous reviewers and the members of the CLIP lab at University of Maryland for their constructive feedback. This work was supported in part by NSF Fairness in AI Grant 2147292, by the Institute for Trustworthy AI in Law and Society (TRAILS), which is supported by the National Science Foundation under Award No. 2229885, and by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract #2022-22072200006, by NSF grant 2147292. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, NSF or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Sweta Agrawal, António Farinhas, Ricardo Rei, and Andre Martins. 2024. Can automatic metrics assess high-quality translations? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14491–14502, Miami, Florida, USA. Association for Computational Linguistics.

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. *Preprint*, arXiv:2402.17733.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress. *Preprint*, arXiv:2405.15032.

Wilker Aziz, Marc Dymetman, Lucia Specia, and Shachar Mirkin. 2010. Learning an expert from human annotations in statistical machine translation: the case of out-of-vocabulary words. In *Proceedings of the 14th Annual Conference of the European Association for Machine Translation*, Saint Raphaël, France. European Association for Machine Translation.

Lynne Bowker. 2021. Promoting Linguistic Diversity and Inclusion: Incorporating Machine Translation Literacy into Information Literacy Instruction for Undergraduate Students. *The International Journal of Information, Diversity, & Inclusion (IJIDI)*, 5(3).

Eleftheria Briakou, Jiaming Luo, Colin Cherry, and Markus Freitag. 2024a. Translating step-by-step: Decomposing the translation process for improved translation quality of long-form texts. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1301–1317, Miami, Florida, USA. Association for Computational Linguistics.

Eleftheria Briakou, Jiaming Luo, Colin Cherry, and Markus Freitag. 2024b. Translating step-by-step: Decomposing the translation process for improved translation quality of long-form texts. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1301–1317, Miami, Florida, USA. Association for Computational Linguistics.

Hannah Béchara, Constantin Orăsan, Carla Parra Escartín, Marcos Zampieri, and William Lowe. 2021. The role of machine translation quality estimation in the post-editing workflow. *Informatics*, 8(3):61.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 17–24, New York City, USA. Association for Computational Linguistics.

Raman Chandrasekar and Srinivas Bangalore. 1997. Automatic induction of rules for text simplification. *Knowl.-Based Syst.*, 10:183–190.

Liang Chen, Shuming Ma, Dongdong Zhang, Furu Wei, and Baobao Chang. 2024a. On the pareto front of multilingual neural machine translation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2024b. Iterative translation refinement with large language models. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 181–190, Sheffield, UK. European Association for Machine Translation (EAMT).

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: efficient finetuning of quantized llms. In *Proceedings of the 37th International Conference on Neural Information Processing*

Systems, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Efthimis N. Efthimiadis. 1996. Query expansion. *Annual Review of Information Science and Technology (ARIST)*, 31:121–187.

Lijun Feng. 2008. Text simplification: A survey. Technical report, CUNY.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083, Singapore. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2015. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23:3 – 30.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc

Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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

Skyler Hallinan, Faeze Brahman, Ximing Lu, Jaehun Jung, Sean Welleck, and Yejin Choi. 2023. STEER: Unified style transfer with expert reinforcement. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7546–7562, Singapore. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Yichong Huang, Xiaocheng Feng, Xinwei Geng, Baohang Li, and Bing Qin. 2023. Towards higher Pareto frontier in multilingual machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3802–3818, Toronto, Canada. 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, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2024. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

David Kumhyr, Carla Merrill, and Karin Spalink. 1994. Internationalization and translatability. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, Columbia, Maryland, USA.

Maria Kunilovskaya, Koel Dutta Chowdhury, Heike Przybyl, Cristina España-Bonet, and Josef Genabith. 2024. Mitigating translationese with GPT-4: Strategies and performance. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 411–430, Sheffield, UK. European Association for Machine Translation (EAMT).

Mina Lee, Katy Ilonka Gero, John Joon Young Chung, Simon Buckingham Shum, Vipul Raheja, Hua Shen, Subhashini Venugopalan, Thiemo Wambsganss, David Zhou, Emad A. Alghamdi, Tal August, Avinash Bhat, Madiha Zahrah Choksi, Senjuti Dutta, Jin L.C. Guo, Md Naimul Hoque, Yewon Kim, Simon Knight, Seyed Parsa Neshaei, Antonette Shibani, Disha Shrivastava, Lila Shroff, Agnia Sergeyuk, Jessi Stark, Sarah Sterman, Sitong Wang, Antoine Bosselut, Daniel Buschek, Joseph Chee Chang, Sherol Chen, Max Kreminski, Joonsuk Park, Roy Pea, Eugenia Ha Rim Rho, Zejiang Shen, and Pao Siangliulue. 2024. A design space for intelligent and interactive writing assistants. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24. ACM.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5303–5315, Singapore. Association for Computational Linguistics.

Kelong Mao, Zhicheng Dou, Bang Liu, Hongjin Qian, Fengran Mo, Xiangli Wu, Xiaohua Cheng, and Zhao Cao. 2023. Search-oriented conversational query editing. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4160–4172, Toronto, Canada. Association for Computational Linguistics.

Shengyu Mao, Yong Jiang, Boli Chen, Xiao Li, Peng Wang, Xinyu Wang, Pengjun Xie, Fei Huang, Huajun Chen, and Ningyu Zhang. 2024. RaFe: Ranking feedback improves query rewriting for RAG. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 884–901, Miami, Florida, USA. Association for Computational Linguistics.

Yuval Marton, Chris Callison-Burch, and Philip Resnik. 2009. Improved statistical machine translation using monolingually-derived paraphrases. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 381–390, Singapore. Association for Computational Linguistics.

Nikita Mehandru, Sweta Agrawal, Yimin Xiao, Ge Gao, Elaine Khoong, Marine Carpuat, and Niloufar Salehi. 2023. Physician detection of clinical harm in machine translation: Quality estimation aids in reliance and backtranslation identifies critical errors. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11633–11647, Singapore. Association for Computational Linguistics.

Shachar Mirkin, Lucia Specia, Nicola Cancedda, Ido Dagan, Marc Dymetman, and Idan Szpektor. 2009. Source-language entailment modeling for translating unknown terms. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 791–799, Suntec, Singapore. Association for Computational Linguistics.

Shachar Mirkin, Sriram Venkatapathy, Marc Dymetman, and Ioan Calapodescu. 2013. SORT: An interactive source-rewriting tool for improved translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 85–90, Sofia, Bulgaria. Association for Computational Linguistics.

Subhajit Naskar, Daniel Deutsch, and Markus Freitag. 2023. Quality estimation using minimum Bayes risk. In *Proceedings of the Eighth Conference on Machine Translation*, pages 806–811, Singapore. Association for Computational Linguistics.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. LLM evaluators recognize and favor their own generations. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):140:5485–140:5551.

Vipul Raheja, Dhruv Kumar, Ryan Koo, and Dongyeop Kang. 2023. CoEdIT: Text editing by task-specific instruction tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5274–5291, Singapore. Association for Computational Linguistics.

Vikas Raunak, Amr Sharaf, Yiren Wang, Hany Awadalla, and Arul Menezes. 2023. Leveraging GPT-4 for automatic translation post-editing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12009–12024, Singapore. Association for Computational Linguistics.

Alessandra Rossetti. 2019. *Simplifying, Reading, and Machine Translating Health Content: An Empirical Investigation of Usability*. Doctoral, Dublin City University. School of Applied Language and Intercultural Studies.

Lei Shu, Liangchen Luo, Jayakumar Hoskere, Yun Zhu, Yinxiao Liu, Simon Tong, Jindong Chen, and Lei Meng. 2024. Rewritelm: an instruction-tuned large language model for text rewriting. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

Marina Solnyshkina, Vladimir Ivanov, and Valery Solovyev. 2018. *Readability Formula for Russian Texts: A Modified Version: 17th Mexican International Conference on Artificial Intelligence, MICAI 2018, Guadalajara, Mexico, October 22–27, 2018, Proceedings, Part II*, pages 132–145.

Sanja Štajner and Maja Popovic. 2016. Can text simplification help machine translation? In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 230–242.

Sanja Štajner and Maja Popović. 2019. Automated text simplification as a preprocessing step for machine translation into an under-resourced language. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1141–1150, Varna, Bulgaria. INCOMA Ltd.

Emma Steigerwald, Valeria Ramírez-Castañeda, Débora Y C Brandt, András Báldi, Julie Teresa Shapiro, Lynne Bowker, and Rebecca D Tarvin. 2022. Overcoming Language Barriers in Academia: Machine Translation Tools and a Vision for a Multilingual Future. *BioScience*, 72(10):988–998.

Christian Tomani, David Vilar, Markus Freitag, Colin Cherry, Subhajit Naskar, Mara Finkelstein, Xavier Garcia, and Daniel Cremers. 2024. Quality-aware translation models: Efficient generation and quality estimation in a single model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15660–15679, Bangkok, Thailand. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Marcos V Treviso, Nuno M Guerreiro, Sweta Agrawal, Ricardo Rei, José Pombal, Tania Vaz, Helena Wu, Beatriz Silva, Daan Van Stigt, and Andre Martins. 2024a. xTower: A multilingual LLM for explaining and correcting translation errors. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15222–15239, Miami, Florida, USA. Association for Computational Linguistics.

Marcos V Treviso, Nuno M Guerreiro, Sweta Agrawal, Ricardo Rei, José Pombal, Tania Vaz, Helena Wu, Beatriz Silva, Daan Van Stigt, and Andre Martins. 2024b. xTower: A multilingual LLM for explaining and correcting translation errors. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15222–15239, Miami, Florida, USA. Association for Computational Linguistics.

Kiyotaka Uchimoto, Naoko Hayashida, Toru Ishida, and Hitoshi Isahara. 2005. Automatic rating of machine translatability. In *Proceedings of Machine Translation Summit X: Papers*, pages 235–242, Phuket, Thailand.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Fei Xia and Michael McCord. 2004. Improving a Statistical MT System with Automatically Learned Rewrite Patterns. In *Proceedings of COLING 2004*, pages 508–514, Geneva, Switzerland.

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.

Fanghua Ye, Meng Fang, Shenghui Li, and Emine Yilmaz. 2023. Enhancing conversational search: Large language model-aided informative query rewriting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5985–6006, Singapore. Association for Computational Linguistics.

Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: Story writing with large language models. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*, IUI '22, page 841–852, New York, NY, USA. Association for Computing Machinery.

Jiali Zeng, Fandong Meng, Yongjing Yin, and Jie Zhou. 2024. Improving machine translation with large language models: A preliminary study with cooperative decoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13275–13288, Bangkok, Thailand. Association for Computational Linguistics.

Hongyi Zhu, Jia-Hong Huang, Stevan Rudinac, and Evangelos Kanoulas. 2024. Enhancing interactive image retrieval with query rewriting using large language models and vision language models. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, ICMR '24. ACM.

R. Zowalla, D. Pfeifer, and T. Wetter. 2023. Readability and topics of the german health web: Exploratory study and text analysis. *PloS one*, 18(2):e0281582.

A Model and Experiment Details

A.1 Prompt Templates

In Tables 5 and 6, we describe the prompt templates used for prompting and fine-tuning experiments, respectively. For stylistic rewriting, we use the same prompts as those used to train the CoEdIT-XL model. During prompting, we provide the original source as the input, while for fine-tuning, we provide the positive rewrite along with the source.

A.2 Training Setup

All models are trained using one NVIDIA RTX A5000 GPU. In practice, we find that fine-tuning converges in around 3 hours. We use a 90/10 train/validation data split and adopt QLoRA (Dettmers et al., 2023), a quantized version of LoRA (Hu et al., 2022), for parameter-efficient training. We train Tower-Instruct 7B with 8-bit quantization, a LoRA rank of 16, a scaling parameter (α) of 32, and a dropout probability of 0.05 for layers. We train for 10 epochs. All unspecified hyperparameters are set to default values.

A.3 Decoding Strategy

We use greedy decoding (no sampling) when generating rewrites for prompting experiments. We fix the temperature value to 0 throughout the experiments in order to eliminate sampling variations.

A.4 Dataset Details

We provide detailed statistics of our training ($\mathcal{D}_{pos}$) and test dataset in Table 7. For $\mathcal{D}_{pos}$, we only use rewrites where the xCOMET(s', t') score is higher than the original xCOMET(s, t) score. We further conduct a two-step pre-processing procedure: **1)** Remove duplicate instances and **2)** Remove lengthy instances where the upper threshold is set as Q3 + $1.5 \times$ IQR.

B Detailed Results

B.1 Full Results

In Tables 8 to 10, we present the detailed numerical results for all tested variations. Most rewrites yield higher xCOMET(s, t) scores, indicating better translatability compared to the original baseline. For stylistic rewrites with CoEdIT, prompting to make the text easier to understand (Understandable) achieves the highest translatability score, while prompting to rewrite the text more formally (Formal) results in the highest translation quality. The Coherent prompt achieves the highest meaning preservation score but this is because most rewrites are merely copies of the original source (Appendix C.1). Overall, we demonstrate that translatability-based selection method remains the most effective method, even outperforming scores from our fine-tuned LLMs.

B.2 Impact of LLM

Among the three LLMs used for prompting, Tower-Instruct performs the best in terms of the combined metric xCOMET(s, t, r). Although it lags behind LLaMA-2 and LLaMA-3 in translatability, its meaning preservation score deteriorates the least, resulting in the highest overall score. LLaMA-3 performs the best in terms of translatability, likely due to its more multilingual training data, with over 5% of its pre-training dataset consisting of high-quality non-English data.[16] This suggests that the amount of multilingual data in the pre-training phase may enhance the model's ability to generate more translatable rewrites. However,

[16] https://ai.meta.com/blog/meta-LLaMA-3/

this advantage does not extend when comparing the LLaMA models to Tower-Instruct. Despite being inherently multilingual primarily trained on translation-related tasks, Tower-Instruct performs lower than the LLaMA models in translatability. This discrepancy can be attributed to Tower-Instruct not being specifically trained on rewriting tasks to improve MT quality, highlighting the importance of introducing translation-related knowledge for effective rewriting.

We further compare the results with off-the-shelf paraphrasing (DIPPER) and text-editing (CoEdIT-XL) tools. Despite being specifically trained for rewriting tasks, their rewrites are not as translatable as those generated by the prompted LLMs. For DIPPER, this may be due to its primary focus on paraphrasing, which has been shown to be less effective (§4.1). In the case of CoEdIT, we attribute the lower performance to the model's smaller size (3B) compared to the 7B LLMs used for prompting.

B.3 Same LLM vs. Different LLM

We distinguish whether the LLM being prompted is the same as the one used as the MT system. Initially, we expected the highest improvements when prompting Tower-Instruct, which may incur self-preference bias, where the LLM favors its own outputs due to recognition (Panickssery et al., 2024). However, our results indicate that prompting Tower-Instruct does not yield the most translatable rewrites. Instead, the LLaMA series models consistently outperform in this aspect. Interestingly, Tower-Instruct consistently produces rewrites that are more meaning-preserving compared to LLaMA-2 or LLaMA-3, resulting in higher xCOMET(s, t, r) scores overall. We conclude that prompting the same LLM used for the MT system is not helpful in generating more translatable rewrites, but these rewrites are better at preserving the intended meaning.

C Qualitative Evaluation Details

C.1 Copying Behavior

To prevent LLMs from directly copying the original source, we explicitly state in the prompt to "*avoid directly copying the source*" (Appendix A.1). However, we still observe some rewrites that are identical to the source sentence. We count the occurrences and compute the percentage per language pair in Table 11. Note that we do not consider Translatability-Aware Selection rewrite method here since this involves selecting whether to keep the original source or use the rewrite based on translatability scores. The highest occurrence appears for stylistic rewrites using the CoEdIT-XL Coherent prompt, where the source is copied most of the time (82.2%, 91.9%, 93.2% for En-De, En-Ru, and En-Zh, respectively).

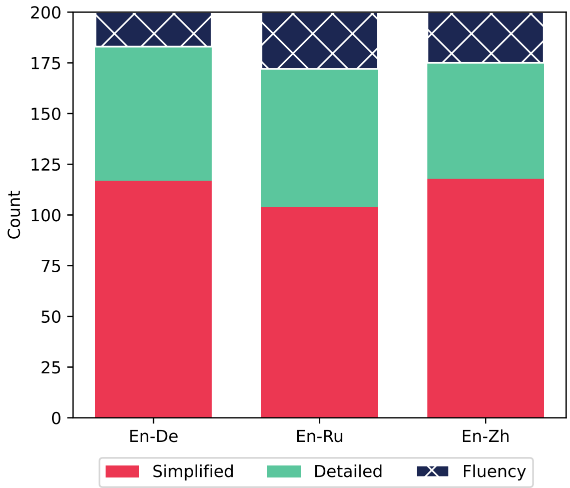

Figure 4: Distribution of properties of good rewrites.

C.2 What makes a Good Rewrite for MT?

Qualitatively examining translation outputs reveals several common patterns, which motivate us to conduct a detailed qualitative analysis. Here, we aim to identify the properties that lead to meaning-equivalent rewrites that are easier to translate. We examine 200 data instances where each rewrite is the highest performing rewrite based on the xCOMET(s, t) score. To focus on successful rewrites, we filter instances where xCOMET$(s', t') >$ xCOMET(s, t). Each rewrite is annotated with the following labels: (1) **Simplified**: Replaces complex words with simpler ones or reduces structural complexity; (2) **Detailed**: Adds information for better context; (3) **Fluency**: Restructures the sentence for better flow and readability. Examples of rewrites for each annotation label are in Table 12.

As shown in Figure 4, most successful rewrites are labeled as **Simplified**. This highlights the effectiveness of simplification, which has been consistently effective even in the context of LLMs. Notably, many simplified rewrites involve changing complex words to simpler, more conventional alternatives (e.g., "Derry City *emerged victorious* in the President's Cup as they *ran out* 2-0 *win-*

ners over Shamrock Rovers." → "Derry City *won* the President's Cup title by *defeating* Shamrock Rovers 2-0."). This finding aligns with our conclusions from MT-Agnostic rewriting methods (§2.1), where simplification emerged as the best rewrite method among the prompting variations.

D Additional Results

D.1 Additional LLM Baselines

LLMs for Rewriting. Our initial experiments consist of 21 input rewriting methods across 3 LLMs (LLAMA-2 7B, LLAMA-3 8B, and TOWER-INSTRUCT 7B). In Table 16, we present extended experiment results by applying simplification rewriting with two additional LLMs: AYA-23 8B (Aryabumi et al., 2024) and TOWER-INSTRUCT 13B (Alves et al., 2024). The results confirms that simplification rewriting improves translation quality measured by XCOMET(s,t,r) compared to the original baseline.

LLMs for MT. Furthermore, we initially relied on TOWER-INSTRUCT 7B as our MT system for all our experiments since it is specifically trained for translation-related tasks and has demonstrated superior MT performance (§2). However, we extend our analysis by comparing the original baseline and our winning strategy (simplification with TOWER-INSTRUCT 7B) using two additional LLMs as the MT system. As shown in Table 17, our method outperforms the original baseline in terms of both the translation quality (XCOMET(s,t,r)) and METRICX(s,t), regardless of the LLM used as the MT system.

D.2 Additional Language Pairs

To assess the generalizability to other source languages, we test two of our winning strategies (simplification with TOWER-INSTRUCT 7B and inference-time selection) on seven additional into-English and non-English language pairs from the WMT-23 General MT task test set.[17] As shown in Table 18, while translatability scores (XCOMET(s,t)) improve across all language pairs, translation quality (XCOMET(s,t,r)) improvements are less pronounced compared to out-of-English pairs. Notably, gains in translation quality are observed only for German-English (DE-EN) and Chinese-English (ZH-EN) pairs. These results highlight the importance of input rewrites' quality, which is currently higher for high-resource source languages. This motivates further work to strengthen input rewriting for broader range of source languages.

[17] https://www2.statmt.org/wmt23/translation-task.html

E Human Annotation Details

We use Qualtrics[18] to design our survey and Prolific[19] to recruit human annotators fluent in the tested target language.

E.1 Original MT vs. Rewrite MT Details

We randomize the order of the two sentences (original MT and rewrite MT) to mitigate position bias. Annotators evaluate which sentence is better across four dimensions: fluency, understandability, level of detail, and meaning preservation. The entire survey is estimated to take approximately 20 minutes to complete. We recruit a total of 9 annotators and provide a compensation of 5 US dollars per survey (15 US dollars/hr), totaling 45 US dollars.

E.2 Original vs. Rewrite Details

Each annotator is tasked to judge how well the rewritten sentence preserves the meaning of the original source sentence. The survey is estimated to take approximately 30 minutes to complete. We recruit a total of 3 annotators. We offer a compensation of 7.5 US dollars per survey (15 US dollars/hr), totaling 22.5 US dollars.

E.3 Annotator Instructions

In Figures 5 to 8, we present the instructions and survey content provided to annotators. For the Original MT vs. Rewrite MT evaluation, each annotator reviews 20 sets of examples. Each question consists of two parts: **1)** comparing the two sentences based on fluency, understandability, and level of detail, and **2)** selecting which sentence better preserves the meaning of the reference translation. For the Original vs. Rewrite evaluation, each annotator reviews 30 sets of examples. Additionally, a free-form text box is provided alongside each example for annotators to offer feedback or suggestions.

F Time & Computational Efficiency

We show that on average, rewriting with our winning strategy is not a resource-intensive option for downstream applications in terms of both time and

[18] https://www.qualtrics.com
[19] https://www.prolific.com

computation. For approximately 1.5K sentences, the rewrite and MT pipeline using our winning strategy (simplification with TOWER-INSTRUCT 7B takes 1 hour, compared to 30 minutes for the MT process alone. All variants of our prompting experiments are conducted using a single NVIDIA RTX958 A5000 GPU. In terms of efficiency compared to automatic post-editing (§5.2), both approaches remains equivalent in time and computational requirements since the rewriting or post-editing process only differs in its position within the pipeline. Input rewriting modifies the source before the MT system, while output post-editing adjusts the translation after the MT system.

Rewrite	Prompt
Simplification	Simplify the English sentence. Simplification may include identifying complex words and replacing with simpler or shorter words or using active voice instead of passive voice. Try to keep the meaning of the Original sentence. Original: *This is a very nice skirt. The lacy pattern is classy and soft.* Simplified:
Paraphrase	Paraphrase the English sentence. Try to not directly copy but keep the meaning of the Original sentence. Original: *This is a very nice skirt. The lacy pattern is classy and soft.* Paraphrase:
Stylistic (COEDIT)	**(GEC)** Fix the grammar: **(Coherent)** Make this text coherent: **(Understandable)** Rewrite to make this easier to understand: **(Formal)** Write this more formally:
Easy Translation	Rewrite the Original sentence to be easier for translation in target language. New sentence should be in English. Original: *This is a very nice skirt. The lacy pattern is classy and soft.* New:
CoT	**(Step 1)** Rewrite the Original English sentence to New English sentence that translates better into German. Avoid directly copying the Original sentence while keeping its meaning. New sentence should be in English. Original: *This is a very nice skirt. The lacy pattern is classy and soft.* New: **(Step 2)** Now, translate the English sentence to German. English: German:

Table 5: Exemplar prompt templates for English-German language pair used for prompting experiments. *Italic* represents the source sentence used in this example.

Basic
Instruction: Rewrite this English sentence to give a better translation.\n\n ### **English**: This is a very nice skirt. The lacy pattern is classy and soft.\n ### **English rewrite**: The lacy pattern on this skirt is elegant and soft.
MT
Instruction: Rewrite this English sentence to give a better translation in German. German sentence is the hypothesis translation that we are trying to improve.\n\n ### **English**: This is a very nice skirt. The lacy pattern is classy and soft.\n ### **German**: Das ist eine sehr schöne Röhre. Das schicke Spitzenmuster ist weich und elegant.\n ### **English rewrite**: The lacy pattern on this skirt is elegant and soft.
Reference
Instruction: Rewrite this English sentence to give a better translation in German. German sentence is the human-annotated translation that we are trying to pursue.\n\n ### **English**: This is a very nice skirt. The lacy pattern is classy and soft.\n ### **German**: Das ist ein sehr schöner Rock. Das Spitzenmuster ist stilvoll und weich.\n ### **English rewrite**: The lacy pattern on this skirt is elegant and soft.

Table 6: Exemplar prompt templates for supervised fine-tuning experiments (English-German pair). We additionally give machine translation for the **MT** prompt and reference translation for the **Reference** prompt after ### **German:**.

Split	Dataset	# Sentences
Train	Source and positive rewrite pairs for SFT (English-German, $\mathcal{D}_{pos}$)	7,016
	Source and positive rewrite pairs for SFT (English-Russian, $\mathcal{D}_{pos}$)	8,126
Test	WMT-23 General MT Task (English-German)	1,557
	WMT-23 General MT Task (English-Russian)	2,074
	WMT-23 General MT Task (English-Chinese)	3,074
	WMT-23 General MT Task (English-Czech)	2,074
	WMT-23 General MT Task (English-Hebrew)	2,074
	WMT-23 General MT Task (English-Japanese)	2,074

Table 7: Summary statistics of training and test datasets.

Language	Type	Prompt/Model	xCOMET(s,t)	xCOMET(s,r)	xCOMET(s,t,r)	MetricX(s,t)	MetricX(t,r)
EN-DE	Original	-	0.893	0.904	0.898	2.038	1.534
	MT-Agnostic	**Simplification** (LLaMA-2)	**0.931**	0.846	0.900	**1.185**	1.727
		(LLaMA-3)	**0.944**	0.820	0.903	**0.925***	1.600
		(Tower-Instruct)	**0.922**	0.885	**0.907**	1.504	1.519
		Paraphrase (LLaMA-2)	**0.926**	0.823	0.889	**1.126**	**1.480**
		(LLaMA-3)	**0.938**	0.796	0.892	**0.955**	**1.469**
		(Tower-Instruct)	0.902	0.887	0.901	**1.310**	1.534
		(DIPPER (L80/O60))	**0.904**	0.745	0.838	**1.674**	2.757
		(DIPPER (L80/O40))	**0.913**	0.797	0.863	**1.461**	2.266
		(DIPPER (L60/O40))	**0.917**	0.847	0.892	**1.555**	1.958
		Stylistic (CoEdIT GEC)	**0.901**	0.899	0.900	**1.709**	1.555
		(CoEdIT Coherent)	0.898	0.900	0.898	**1.728**	1.595
		(CoEdIT Understandable)	**0.949**	**0.758**	0.862	**0.989**	2.610
		(CoEdIT Formal)	**0.937**	0.830	0.900	**1.063**	1.879
	Task-Aware	**Easy Translation** (LLaMA-2)	**0.916**	0.857	0.893	**1.654**	2.482
		(LLaMA-3)	**0.932**	0.827	0.899	**1.151**	2.241
		(Tower-Instruct)	**0.901**	0.900	0.903	**1.759**	2.427
		CoT (Tower-Instruct)	**0.907**	0.816	0.897	**1.892**	1.578
	Translatability-Aware	**Selection**	**0.921**	0.907	**0.915***	**1.734**	**1.461***
		Fine-tune (Basic)	**0.934**	0.851	**0.909**	**1.878**	**1.499**
		(MT)	**0.919**	0.856	0.903	**1.947**	1.593
		(Reference)	0.896	0.836	0.876	2.023	2.028

Table 8: Detailed results of English-German pair using different rewrite methods. Statistically significant average improvements (p-value < 0.05) are **bold**. Best scores for each metric is **bold** with *. xCOMET(s,t): translatability (↑); xCOMET(s,r): meaning preservation (↑); xCOMET(s,t,r): overall translation quality (↑); MetricX(s,t): quality estimation (↓); MetricX(t,r): reference-based metric (↓). For DIPPER (Krishna et al., 2024) variations, L and O denote lexical and order diversity, respectively.

Language	Type	Prompt/Model	xCOMET(s,t)	xCOMET(s,r)	xCOMET(s,t,r)	METRICX(s,t)	METRICX(t,r)
EN-RU	**Original**	-	0.872	0.884	0.868	2.535	2.028
	MT-Agnostic	**Simplification** (LLAMA-2)	**0.916**	0.839	**0.882**	**0.951**	2.160
		(LLAMA-3)	**0.919**	0.812	**0.885**	**0.804**	2.039
		(TOWER-INSTRUCT)	**0.921**	0.870	**0.891**	**1.135**	**1.921**
		Paraphrase (LLAMA-2)	**0.923**	0.804	**0.881**	**0.882**	**1.853**
		(LLAMA-3)	**0.930**	0.788	**0.882**	**0.855**	**1.863**
		(TOWER-INSTRUCT)	**0.887**	0.878	**0.878**	**1.095**	**1.976**
		(DIPPER (L80/O60))	**0.904**	0.735	0.821	**1.249**	3.476
		(DIPPER (L80/O40))	**0.909**	0.790	0.853	**1.105**	2.773
		(DIPPER (L60/O40))	**0.905**	0.834	0.873	**1.119**	2.418
		Stylistic (COEDIT GEC)	0.873	0.884	0.869	**1.327**	**1.969**
		(COEDIT Coherent)	0.873	0.884	0.869	**1.368**	**1.989**
		(COEDIT Understandable)	**0.918**	0.801	0.873	**0.991**	2.726
		(COEDIT Formal)	**0.916**	0.841	**0.887**	**0.922**	2.020
	Task-Aware	**Easy Translation** (LLAMA-2)	**0.914**	0.839	**0.884**	**1.037**	10.849
		(LLAMA-3)	**0.917**	0.808	**0.881**	**0.801***	10.401
		(TOWER-INSTRUCT)	**0.885**	0.883	**0.878**	**1.277**	11.137
		CoT (TOWER-INSTRUCT)	**0.903**	0.871	0.875	**2.432**	2.024
	Translatability-Aware	**Selection**	**0.914**	**0.891***	**0.899***	**2.096**	**1.830***
		Fine-tune (Basic)	**0.912**	0.848	**0.886**	**2.123**	**1.932**
		(MT)	**0.904**	0.851	0.871	**2.119**	**1.997**
		(Reference)	**0.881**	0.812	0.859	**2.284**	2.012

Table 9: Detailed results of English-Russian pair using different rewrite methods.

Language	Type	Prompt/Model	xCOMET(s,t)	xCOMET(s,r)	xCOMET(s,t,r)	METRICX(s,t)	METRICX(t,r)
EN-ZH	**Original**	-	0.786	0.775	0.794	3.445	2.282
	MT-Agnostic	**Simplification** (LLAMA-2)	**0.828**	0.728	0.796	**1.321**	2.537
		(LLAMA-3)	**0.823**	0.701	0.795	**1.252***	2.572
		(TOWER-INSTRUCT)	**0.821**	0.759	**0.802**	**1.521**	**2.227**
		Paraphrase (LLAMA-2)	**0.818**	0.683	0.771	**1.330**	2.478
		(LLAMA-3)	**0.826**	0.662	0.766	**1.341**	2.534
		(TOWER-INSTRUCT)	**0.797**	0.765	0.798	**1.580**	2.283
		(DIPPER (L80/O60))	**0.813**	0.622	0.722	**1.583**	4.009
		(DIPPER (L80/O40))	**0.816**	0.670	0.750	**1.499**	3.196
		(DIPPER (L60/O40))	**0.809**	0.711	0.775	**1.503**	2.725
		Stylistic (COEDIT GEC)	0.789	0.772	0.795	**1.632**	2.251
		(COEDIT Coherent)	0.786	0.774	0.794	**1.658**	2.267
		(COEDIT Understandable)	**0.839***	0.677	0.774	**1.358**	3.174
		(COEDIT Formal)	**0.823**	0.730	0.798	**1.336**	2.443
	Task-Aware	**Easy Translation** (LLAMA-2)	**0.821**	0.721	0.784	**1.900**	7.732
		(LLAMA-3)	**0.830**	0.687	0.783	**1.360**	7.608
		(TOWER-INSTRUCT)	**0.793**	0.762	0.791	**1.618**	7.650
		CoT (TOWER-INSTRUCT)	**0.821**	0.769	0.771	**3.321**	2.432
	Translatability-Aware	**Selection**	**0.823**	**0.783***	**0.819***	**3.149**	**2.206***

Table 10: Detailed results of English-Chinese pair using different rewrite methods.

Type	Prompt/Model	EN-DE	EN-RU	EN-ZH
MT-Agnostic	**Simplification** (LLAMA-2)	2.06	2.37	2.37
	(LLAMA-3)	0.39	0.33	0.29
	(TOWER-INSTRUCT)	28	29.3	30.2
	Paraphrase (LLAMA-2)	0	0	0
	(LLAMA-3)	0.06	0.07	0.03
	(TOWER-INSTRUCT)	37.3	38.2	38
	(DIPPER (L80/O60))	0.19	0.94	1.04
	(DIPPER (L80/O40))	0.51	1.5	1.53
	(DIPPER (L60/O40))	1.48	2.5	2.44
	Stylistic (COEDIT GEC)	42.6	44	48.3
	(COEDIT Coherent)	82.2	91.9	93.2
	(COEDIT Understandable)	1.61	1.88	1.53
	(COEDIT Formal)	5.33	3.76	5.5
Task-Aware	**Easy Translation** (LLAMA-2)	3.04	3.55	3.63
	(LLAMA-3)	0.24	0.66	0.27
	(TOWER-INSTRUCT)	12.3	18.6	15.4
	CoT (TOWER-INSTRUCT)	0.71	1.45	1.53
Translatability-Aware	**Fine-tune** (Basic)	4.5	3.91	-
	(MT)	3.73	3.42	-
	(Reference)	6.17	7.85	-

Table 11: Percentage of occurrence (%) where the rewrite is a direct copy of the original source sentence.

Label	Original	Rewrite	Original MT	Rewrite MT	xCOMET(s,t)	xCOMET(s',t')
Simplified	When Michael “Hopper” McGrath **lobbed** a ball in, Molloy **leapt** highest before rifling a sublime goal to the roof of the net.	When Michael McGrath **threw** the ball in, Molloy **jumped** highest and scored a beautiful goal to the top of the net.	Als Michael “Hopper” McGrath einen Ball hereinwarf, sprang Molloy am höchsten und schoss einen herrlichen Treffer auf das Dach des Netzes.	Als Michael McGrath den Ball in die Luft warf, sprang Molloy am höchsten und erzielte einen wunderschönen Treffer in die obere Netzhöhe.	0.906	0.945
	Derry City **emerged victorious** in the President’s Cup as they ran out 2-0 winners over Shamrock Rovers.	Derry City **won** the President’s Cup title by defeating Shamrock Rovers 2-0.	Derry City 在总统杯赛中获胜，以2-0 的比分击败尚洛克罗弗斯。	Derry City 以2-0 的比分击败Shamrock Rovers，获得了总统杯冠军。	0.648	0.952
Detailed	The great majority of rankers never advanced beyond principalis.	The vast majority of soldiers remained in the lowest rank throughout their careers.	Die große Mehrheit der Reiter schaffte es nie über den Rang eines principalis.	Die überwiegende Mehrheit der Soldaten blieb während ihrer gesamten Karriere in der niedrigsten Ränge.	0.938	0.982
	I’ve noticed you almost need line of sight for it to work.	It appears that visibility plays a crucial role in the effectiveness of the process.	Я заметил, что для работы вам почти все время нужен прямой свет.	Похоже, что видимость играет решающую роль в эффективности процесса.	0.98	1.0
Fluency	It’s a thing I’ve never said before either.	I’ve never said that before either.	Это то, что я никогда не говорил раньше.	Я никогда этого не говорил и раньше.	0.989	1.0
	When I started in summer with those multi-source experiments.	I began a series of experiments in the summer.	我在夏天开始进行多来源实验时。	我在夏天开始了一系列的实验。	0.858	1.0

Table 12: Examples of rewrites for each annotation label (**Simplified**, **Detailed** and **Fluency**).

Prompt/Model	Original	Rewrite	Original MT	Rewrite MT	Flesch(s)	Flesch(s')	WSTF(t)	WSTF(t')
Simplification (LLAMA-3)	She **steamed via** Hawaii, Midway, Guam, and Subic Bay for Vietnam and anchored in the Saigon River on 13 September.	She **sailed from** Hawaii to Vietnam, stopping at Midway, Guam, and Subic Bay, and **arrived at** the Saigon River on September 13.	Sie fuhr über Hawaii, Midway, Guam und Subic Bay nach Vietnam und ankerte am 13. September in der Saigon-Schifffahrt.	Sie segelte von Hawaii nach Vietnam, machte Halt in Midway, Guam und Subic Bay und erreichte am 13. September in der Saigon River.	74.53	**76.56**	1.032	**0.838**
Simplification (TOWER-INSTRUCT)	The remnants of Felix continued northeastward across the Atlantic until dissipating near Shetland on August 25.	Felix's remnants continued northeastward across the Atlantic until dissipating near Shetland on August 25.	Die Überreste von Felix zogen sich über den Atlantik in nordöstlicher Richtung bis zum 25. August, als sie sich in der Nähe von Shetland auflösten.	Felix's Reste zogen sich über den Atlantik in nordöstlicher Richtung bis zum 25. August, als sie sich in der Nähe von Shetland auflösten.	31.89	**38.32**	1.193	**1.109**
	Cambrai thus **reverted**, but only briefly, to the Western Frankish Realm.	Cambrai **returned** to the Western Frankish Realm, but only briefly.	Cambrai fiel daher, aber nur kurzzeitig, wieder an das Westfrankenreich zurück.	Cambrai kehrte zum Westfrankenreich zurück, aber nur kurz.	**68.77**	54.22	0.728	**0.429**

Table 13: Examples of simplification rewrites for English-German (EN-DE) pair and their corresponding input and output readability scores. **Flesch**: Flesch Reading Ease score (↑); **WSTF**: Vienna formula (↓).

Prompt/Model	Original	Rewrite	Original MT	Rewrite MT	Flesch(s)	Flesch(s')	Flesch-Ru(t)	Flesch-Ru(t')
Simplification (LLAMA-3)	Later, Wallachia's Vornic Radu Socol traveled to Suceava, bringing Despot two steeds, a kuka hat with precious stones, and 24,000 ducats.	Radu Socol, the Vornic of Wallachia, visited Suceava and brought two horses, a hat with precious stones, and 24,000 ducats to Despot.	Позже, Ворник Раду Соколь из Валахии отправился в Сучаву, привезнув деспоту двух лошадей, кукушку с драгоценными камнями и 24 000 дукатов.	Раду Сокол, вонник Валахии, посетил Сучаву и принес деспоту два коня, шляпу с драгоценными камнями и 24 000 дукатов.	**67.08**	66.07	55.81	**64.80**
Simplification (TOWER-INSTRUCT)	**Appalled** at the thought of Emily **cavorting** with Casey, Margo **vindictively revealed** Emily's **hooker past** to Tom and Casey.	Margo was shocked that Emily was hanging out with Casey and so she told Tom and Casey about Emily's past as a prostitute.	Потрясенная мыслью о том, что Эмили развлекается с Кейси, Марго мстительно рассказала Тому и Кейси о прошлом Эмили проституткой.	Марго была потрясена тем, что Эмили общалась с Кейси, и поэтому она рассказала Тому и Кейси о прошлом Эмили как проститутке.	60.65	**81.97**	58.47	**64.40**

Table 14: Examples of simplification rewrites for English-Russian (EN-RU) pair and their corresponding input and output readability scores. **Flesch**: Flesch Reading Ease score (↑); **Flesch-Ru**: Russian version of Flesch (↑).

Prompt/Model	Original	Rewrite	Original MT	Rewrite MT	Flesch(s)	Flesch(s')
Simplification (LLAMA-3)	During the delay, the tire carcass wrapped itself around the axle, costing him several laps.	The tire wrapped around the axle, causing him to lose several laps.	延迟期间，轮胎壳破损，裹住了轮毂，让他失去了几圈的速度。	轮胎缠在轴上，让他失去了几圈。	64.71	**84.68**
Simplification (TOWER-INSTRUCT)	Japanese artillery attempted to engage them but South Dakota and the other battleships easily outranged them.	Japanese artillery tried to attack them but South Dakota and the other battleships were too far away.	日本炮兵试图与他们交战，但南达科他和其他战舰的射程远远超过他们。	日本炮兵试图袭击他们，但南达科他和其他战舰太远了。	38.32	**62.68**

Table 15: Examples of simplification rewrites for English-Chinese (EN-ZH) pair and their corresponding input readability scores. **Flesch**: Flesch Reading Ease score (↑).

Language	Prompt/Model	**xCOMET**(s,t)	**xCOMET**(s,t,r)	**MetricX**(s,t)	**MetricX**(t,r)
EN-DE	Original	0.893	0.898	2.038	1.534
	Simplification (Aya-23 8B)	0.901	0.900	1.956	1.624
	Simplification (Tower-Instruct 13B)	**0.924**	**0.912**	**1.562**	**1.445**
EN-RU	Original	0.872	0.868	2.535	2.028
	Simplification (Aya-23 8B)	0.880	**0.875**	2.428	1.938
	Simplification (Tower-Instruct 13B)	**0.901**	**0.889**	**2.137**	**1.861**

Table 16: Results with two additional LLMs for rewriting: Aya-23 8B and Tower-Instruct 13B. Statistically significant average improvements (p-value < 0.05) are **bold**. xCOMET(s,t): translatability (↑); xCOMET(s,t,r): overall translation quality (↑); MetricX(s,t): quality estimation (↓); MetricX(t,r): reference-based metric (↓).

Language	MT System	Prompt/Model	**xCOMET**(s,t)	**xCOMET**(s,t,r)	**MetricX**(s,t)	**MetricX**(t,r)
EN-DE	Tower-Instruct 7B	Original	0.893	0.898	2.038	1.534
		Simplification	**0.915**	**0.907**	1.504	1.519
	Aya-23 8B	Original	0.887	0.891	1.926	1.554
		Simplification	**0.911**	**0.902**	1.660	1.571
	Tower-Instruct 13B	Original	0.880	0.887	2.043	1.522
		Simplification	**0.900**	**0.893**	**1.778**	1.556
EN-RU	Tower-Instruct 7B	Original	0.872	0.868	2.535	2.028
		Simplification	**0.921**	**0.891**	**1.135**	**1.921**
	Aya-23 8B	Original	0.863	0.852	2.711	2.323
		Simplification	**0.892**	**0.872**	**2.300**	**2.173**
	Tower-Instruct 13B	Original	0.887	0.882	2.290	1.915
		Simplification	**0.894**	0.875	2.296	1.915
EN-ZH	Tower-Instruct 7B	Original	0.786	0.794	3.445	2.282
		Simplification	**0.821**	**0.802**	**1.521**	**2.227**
	Aya-23 8B	Original	0.769	0.779	3.758	2.572
		Simplification	**0.793**	**0.788**	**3.433**	2.530
	Tower-Instruct 13B	Original	0.755	0.764	3.421	2.341
		Simplification	**0.772**	0.767	3.236	2.413

Table 17: Results with two additional LLMs as MT system: Aya-23 8B and Tower-Instruct 13B. Simplification is done with Tower-Instruct 7B. Statistically significant average improvements (p-value < 0.05) over their respective original baselines are **bold**. xCOMET(s,t): translatability (↑); xCOMET(s,t,r): overall translation quality (↑); MetricX(s,t): quality estimation (↓); MetricX(t,r): reference-based metric (↓).

Language	Prompt/Model	xCOMET(s,t)	xCOMET(s,t,r)	METRICX(s,t)	METRICX(t,r)
CS-UK	Original	0.866	0.755	2.437	4.033
	Simplification	**0.885**	0.749	2.355	4.053
	Selection	**0.930**	0.748	3.050	4.053
DE-EN	Original	0.969	0.622	1.869	4.760
	Simplification	**0.975**	**0.632**	1.856	4.600
	Selection	**0.979**	**0.631**	1.856	4.599
HE-EN	Original	0.582	0.556	8.057	5.758
	Simplification	0.562	0.514	8.671	6.374
	Selection	**0.639**	0.514	9.192	6.541
JA-EN	Original	0.884	0.841	3.473	2.688
	Simplification	0.896	0.828	3.303	2.929
	Selection	**0.918**	0.827	3.659	2.964
RU-EN	Original	0.938	0.921	3.024	1.823
	Simplification	0.945	0.922	2.909	1.879
	Selection	**0.954**	0.923	3.079	1.912
UK-EN	Original	0.934	0.929	2.959	1.507
	Simplification	**0.951**	0.929	2.684	1.595
	Selection	**0.962**	0.929	3.055	1.656
ZH-EN	Original	0.797	0.524	5.099	5.666
	Simplification	**0.809**	**0.530**	4.849	5.582
	Selection	**0.827**	0.528	5.202	5.800

Table 18: Results with into-English and non-English language pairs. Simplification is done with TOWER-INSTRUCT 7B. Statistically significant average improvements (p-value < 0.05) over their respective original baselines are **bold**. xCOMET(s,t): translatability ($\uparrow$); xCOMET(s,t,r): overall translation quality ($\uparrow$); METRICX(s,t): quality estimation ($\downarrow$); METRICX(t,r): reference-based metric ($\downarrow$).

0% Survey Completion

In this survey, you will be asked questions about **20 pairs of sentences** written in **German**.

First, you will be asked to compare the two **sentences** to each other.

[Example]
Sentence 1: Das ist eine sehr schöne Röhre. Das schicke Spitzenmuster ist weich und elegant.
Sentence 2: Das ist eine sehr schöne Röhre. Sie ist stilvoll und weich.

Second, you will be given a **reference sentence** and asked to choose which of the two **sentences** better capture the meaning of the reference.

[Example]
Reference: Das ist ein sehr schöner Rock. Das Spitzenmuster ist stilvoll und weich.
Sentence 1: Das ist eine sehr schöne Röhre. Das schicke Spitzenmuster ist weich und elegant.
Sentence 2: Das ist eine sehr schöne Röhre. Sie ist stilvoll und weich.

We estimate that the survey will take approximately **20 minutes** to complete.

Next Page >

Figure 5: Instructions to human annotators for Original MT vs. Rewrite MT evaluation.

***Sentence 1:** Er ist mit abgewinkelten Wetterschindeln verkleidet, die schmaler sind als die des Haupthauses.
Sentence 2: Das Äußere des kleineren Cottages ist mit schlanken, horizontalen Brettern bedeckt, die schmaler sind als die des größeren Hauptcottages.

Compare the two sentences above by answering the following questions:

	Sentence 1	Sentence 2	Both / Hard to tell
Which one is more fluent?	○	○	○
Which one is easier to understand?	○	○	○
Which one contains more information?	○	○	○

Figure 6: Survey content of the first part to compare Original MT vs. Rewrite MT. To avoid position bias, we randomly shuffle the order of original translations (t) and translations of rewrites (t') for **Sentence 1** and **2**.

*Now let's compare our two sentences to a reference:

Reference: Es ist mit gespreizten Wetterbrettern verkleidet, die schmaler sind als die des Haupthauses.
Sentence 1: Er ist mit abgewinkelten Wetterschindeln verkleidet, die schmaler sind als die des Haupthauses.
Sentence 2: Das Äußere des kleineren Cottages ist mit schlanken, horizontalen Brettern bedeckt, die schmaler sind als die des größeren Hauptcottages.

	Sentence 1	Sentence 2	Both / Hard to tell
Which one best captures the meaning of the reference?	○	○	○

If you would like to further explain your answers, please do so here:

Figure 7: Survey content of the second part to compare to the **Reference translation**. An optional text box is given for each example for further comments.

Sentence 1: His tombstone represents him in armor, holding a shield with three cooking pots, marmites, on it.
Sentence 2: His tombstone shows him in armor, holding a shield with three cooking pots, marmites, on it.

To what extent does Sentence 2 capture the meaning of Sentence 1?

1: Not capture meaning

2: Partially

3: Mostly

4: Fully

If you would like to further explain your answers, please do so here:

Figure 8: Survey content to compare Original (**Sentence 1**) vs. Rewrite (**Sentence 2**).