Lost in Literalism: How Supervised Training Shapes Translationese in LLMs

Anonymous ACL submission

Abstract

Large language models (LLMs) have achieved remarkable success in machine translation, demonstrating impressive performance across diverse languages. However, translationese—characterized by overly literal and unnatural translations—remains a persistent challenge in LLM-based translation systems. Despite their pre-training on vast corpora of *natural* utterances, LLMs exhibit translationese errors and generate unexpected *unnatural* translations, stemming from biases introduced during supervised fine-tuning (SFT). In this work, we systematically evaluate the prevalence of translationese in LLM-generated translations and investigate its roots during supervised training. We introduce methods to mitigate these biases, including polishing golden references and filtering unnatural training instances. Empirical evaluations demonstrate that these approaches significantly reduce translationese while improving translation naturalness, validated by human evaluations and automatic metrics. Our findings highlight the need for training-aware adjustments to optimize LLM translation outputs, paving the way for more fluent and target-language-consistent translations.

1 Introduction

Neural machine translation (NMT) has become the dominant method in machine translation (MT) research (Vaswani et al., 2017; Edunov et al., 2018; Hassan et al., 2018). Recently, advancements in large language models have further expanded the capabilities of NMT, demonstrating notable robustness and generalization across diverse text lengths, structures, and languages (Hendy et al., 2023; Jiao et al., 2023b; Kocmi and Federmann, 2023). These works show that LLMs obtain competitive performance on benchmark datasets (e.g., WMT) under automatic metrics, demonstrating strong translation adequacy. However, their translation style has been relatively less addressed. For example, limited research has been devoted to analyzing and improving the naturalness of translations (Raunak et al., 2023; Chen et al., 2024).

Existing work shows that machine translation systems can produce less natural translations, a phenomenon known as "translationese" (Burlot and Yvon, 2018; Aranberri, 2020; Dutta Chowdhury et al., 2022). Translationese occurs when source-language segments are translated too *literally* at

Sentence-level Translationese	
Source	Few-shot LLMs still lag behind vanilla fine-tuned models in the task.
LLM	少样本LLMs仍然落后于原始细化训练模型在任务中。(PPL: 151.5)
Refine	在任务中，少样本LLMs仍然落后于原始细化训练模型。(PPL: 128.8)
Source	Bei starker Hitze ließ diese Festigkeit zwar etwas nach.
LLM	However, at high temperatures this hardness did diminish somewhat. (PPL: 160.1)
Refine	However, this hardness did diminish somewhat at high temperatures. (PPL: 96.6)
Phrase-level Translationese	
Source	after a quick trip in the microwave
LLM	在微波炉的快速(quick)旅行(journey)后 (PPL: 394.3)
Refine	在微波炉中快速(quick)加热(heating)后 (PPL: 56.3)
Source	mehr Lebensqualität zu gewinnen
LLM	gain more quality of life (PPL: 620.5)
Refine	improve living standards (PPL: 368.8)

Table 1: Examples of Sentence-level and Phrase-level Translationese (English-Chinese and German-English translation). Source: source text; LLM: translations of LLMs; Refine: translations with translationese refined. Each case includes an LLM-generated translation alongside a refined version, with perplexity (PPL) values provided at the end. Blue text highlights the source segments, while red text identifies segments in the LLM translation where translationese occurs and is subsequently refined.

either the phrase or sentence level, resulting in deviations from typical target language patterns that sound unnatural to native speakers (Gellerstam, 1986; Nida and Taber, 1982). While considerable research has addressed and mitigated translationese in traditional NMT systems (Burlot and Yvon, 2018; Riley et al., 2020), there has been limited work on whether translationese exists in LLM-based translation systems.

The primary distinction of large translation models lies in the extensive prior knowledge acquired during the pre-training phase, where they learn from a vast corpus of native utterances. Consequently, LLMs should be less susceptible to translationese patterns and capable of producing natural translations due to their strong language modeling bias. However, as illustrated in Table 1, LLMs still produce "unexpected" *unnatural* translations despite their exposure to abundant *natural* language data. For instance, when translating "after a quick trip" into Chinese, the resulting sentence contains the term "旅行", which is a literal translation of "trip" but is not typically used for expressing something going into a microwave oven in Chinese.

We conduct a systematic evaluation to investigate the translationese patterns exhibited by LLMs and examine the underlying causes of these unexpected unnatural translations, engaging expert translators to meticulously analyze translationese in LLMs. Initially, we collect documents from diverse writing domains and use both translation-specialized (e.g., ALMA (Xu et al., 2024b)) and general LLMs (e.g., GPT4 (OpenAI et al., 2024)) for generating translations. For each translated document, expert translators identify specific spans exhibiting pre-defined translationese error types. We then compute the proportion of these spans, termed the Translationese Span Ratio (TSR), and average these ratios across annotators to provide a quantitative measure of translationese prevalence.

Results indicate that all LLMs exhibit significant translationese errors in both English-Chinese and German-English translations. Notably, even advanced models like GPT-4 demonstrate over 40% of their translations as exhibiting substantial translationese patterns. Interestingly, when LLMs are asked to refine their own translations, they produce more natural outputs with markedly lower TSRs. For example, in Table 1, after refining the translation, "trip" becomes "加热" (heated). This suggests that LLMs own prior knowledge and potential for generating natural translations, but may be biased during supervised training (i.e., supervised fine-tuning, SFT) for the "translation" task, placing excessive emphasis on literal semantic mapping at the expense of fluent language generation.

We validate LLMs' potential of generating natural translations by demonstrating a positive correlation between their predicted perplexities and human evaluation: higher perplexities are often associated with increased TSRs. As shown in Table 1, the perplexities of direct LLM translations are higher than those of the refined ones. This finding not only verifies our hypothesis above to some extent but also provides an automatic metric for detecting translationese. To further verify biases introduced during supervised fine-tuning (SFT), we engage expert translators to analyze translationese in sampled training instances from widely used SFT datasets. Our findings reveal that over 34% of these training instances exhibit translationese patterns, indicating that LLMs may be biased towards producing unnatural translations during SFT.

We propose two mitigation strategies to address translationese. First, LLMs' natural potential is leveraged to refine golden training references, reducing translationese patterns. Empirical evaluations on Llama-3.1-8B and Qwen-2.5-7B show that refining training instances improves translation naturalness significantly, as confirmed by both automatic and human evaluations. Second, pre-trained LLMs are used to filter unnatural translations from supervised fine-tuning (SFT) data, which also enhances translation naturalness. Extensive experiments across additional languages further demonstrate the generalizability of our method. To our knowledge, this is the first systematic study addressing translationese in LLMs. We will release our resources after the anonymous period.

2 Related Work

Translationese in Machine Translation. Translationese refers to the phenomenon in which translated texts display linguistic characteristics that diverge from the typical patterns of the target language, resulting in overly literal expressions that sound unnatural to native speakers (Gellerstam, 1986; Nida and Taber, 1982). A line of work has explored translationese and proposed dedicated mitigation strategies. Aranberri (2020) analyze the translationese by measuring various linguistic features, while Bizzoni and Lapshinova-Koltunski (2021) find that texts with translationese

elicit higher perplexities. Several studies have identified data quality issues as a contributing factor to translationese. Researchers (Toral, 2019; Zhang and Toral, 2019; Ni et al., 2022; Wang et al., 2023) study the impact of translationese on model performance, whereas another line of work (Riley et al., 2020; Jalota et al., 2023; Kuwanto et al., 2024; Doshi et al., 2024) relies on translationese to enhance data quality or achieve data augmentation. Dutta Chowdhury et al. (2022) and Wein and Schneider (2024) propose to address the translationese issue using specialized algorithms, while Kunilovskaya et al. (2024) focus on prompt-engineering to mitigate this issue. Unlike their work, we focus on the unexpected translationese in the context of powerful LLMs.

Large Language Model for Translation. Recent studies demonstrate the strong translation capabilities of LLMs like GPT-3.5 and GPT-4, particularly with in-context few-shot learning (Jiao et al., 2023b; Hendy et al., 2023; Kocmi et al., 2023; Xu et al., 2024a; Zhu et al., 2024). A line of work enhances translation performance through prompt engineering, such as dictionary-based approach (Ghazvininejad et al., 2023), knowledge extraction by self-prompting (He et al., 2024) or self-evaluation and refinement (Feng et al., 2024; Ki and Carpuat, 2024; Chen et al., 2024). From a training perspective, researchers (Ouyang et al., 2022), Jiao et al. (2023a), Zeng et al. (2023) and Mao and Yu (2024) propose instruction tuning methods to enhance model alignment with human feedback by comparing multiple translations. Yin et al. (2024) propose a dictionary-based data curation method for efficient SFT. Xu et al. (2024b) identify data quality issues in SFT as a potential contributor to suboptimal translation performance, further corroborated by findings from Gisserot-Boukhlef et al. (2024).

LLMs have excelled in producing fluent and adequate translations, effectively addressing faithfulness and accuracy. However, achieving stylistically natural translations remains a significant challenge. While Raunak et al. (2023) report a reduction in overly literal translations from LLMs, unnatural expressions still pose a significant challenge (Chen et al., 2024). In this work, we systematically analyze the origins of LLM translationese and propose training-aware mitigation methods.

3 Translationese in LLM Translation

To gain a systematic and quantitative assessment of translationese errors in LLM translation, we perform fine-grained human annotation on the outputs generated by these models based on source documents from typical writing tasks.

3.1 Data Collection

We examine four writing domains: news articles, scientific writings, Wikipedia entries, and social media comments. We consider English-Chinese (En-Zh) and German-English (De-En) translations. For the English source segments, we web-crawled 50 document-level samples from each of the following sources: CNN News[1], Arxiv[2], Wikipedia[3], and Quora forums[4]. This process results in 200 English source documents. For the German source segments, we obtained 100 document-level samples consisting of news articles from Focus[5] and comments from Quora forums.

We employ both commercial LLMs such as GPT-3.5-Turbo and GPT-4-Turbo (OpenAI et al., 2024) as well as open-source alternatives including ALMA-7B-R, ALMA-13B-R (Xu et al., 2024a,b), and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023). ALMA models are specialized translation models while the other models are general chat models[6]. All the models employ a straightforward translation prompt, with the exception of GPT models, which use two variants to mitigate translationese errors: the **specified** prompt and the **polish** prompt. While both prompts have the same requirements focused on the target language style, the polish prompt specifically requires refinement of an existing translation, which is a two-step process: first performing direct translation followed by polishing, as detailed in Appendix A.

In this way, each document is translated using nine models: ALMA-7B, ALMA-13B, Mistral-7B, GPT-3.5, GPT-3.5-Specified, GPT-3.5-Polish, GPT-4, GPT-4-Specified, and GPT-4-Polish, where "Specified" and "Polish" refer to using the respective prompts. This process yields a total of 1,800 document-level English-Chinese translations and

[1]https://www.cnn.com/
[2]https://arxiv.org/
[3]https://www.wikipedia.org/
[4]https://www.quora.com/
[5]https://www.focus.de/
[6]Model selection is based on our empirical studies of document-level translation ability.

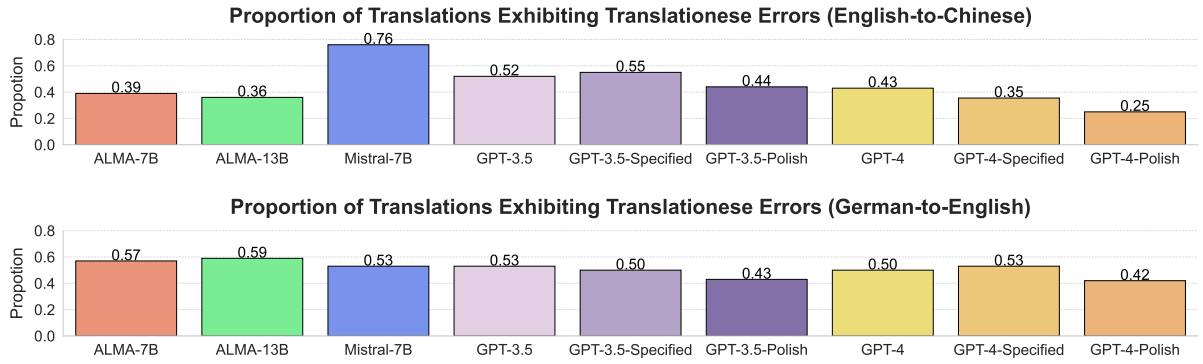

Figure 1: Proportions of translations exhibiting translationese errors. All LLMs adopt direct translation prompts, with the exception of GPT-3.5 and GPT-4, which incorporate supplementary prompts to facilitate more natural translations. Both "Specified" and "Polish" prompts have identical requirements; however, the 'Polish' prompt specifically instructs LLMs to refine their generated translations.

900 German-English translations for human annotation, as summarized in Appendix B.

3.2 Translationese Span Annotation

Using Label Studio (Tkachenko et al., 2020-2024), we develop a specialized annotation platform to help expert translators identify text spans with translationese errors. Inspired by Unbabel's annotation guidelines, we categorize translationese errors into two primary types: **unnatural sentence flow** and **unnatural phrase flow**, corresponding to sentence-level and phrase-level translationese. Unnatural sentence flow occurs when source language structures are translated directly without adequate adaptation to the target language, whereas unnatural phrase flow pertains to overly literal translations of source phrases. Recognizing that traditional translation errors (e.g., omissions and mistranslations) can also occur in LLM outputs, we include these types of errors in our annotation guidelines and platform. Based on the aforementioned translation error taxonomy, we request three expert translators to identify and annotate segments containing translation errors, specifically focusing on two types of translationese errors. The annotators, all of whom hold advanced degrees in linguistics or translation studies and possess extensive experience in professional translation, ensure a high level of accuracy and consistency in identifying nuanced translation errors. Detailed annotation guideline and platform demonstration can be found in Appendix C.

3.3 Human Evaluation Results

We gather human annotation results and calculate the length ratio of spans exhibiting translationese errors (i.e., unnatural sentence and phrase flow) for each document, termed the **translationese span ratio** (TSR). For example, a TSR of 0.2 signifies that 20% of the documents exhibit translationese. The TSRs from three translators are averaged for each document, and then aggregated across all translations for each model. To complete the fine-grained TSR metric, we evaluate the **proportion** of documents with significant translationese errors (significant errors are defined as a TSR greater than 0.2). These documents (TSR>0.2) represent translations that are notably unnatural from a native speaker's perspective. We demonstrate this document-level analysis in Figure 1. Direct TSR scores are presented in Appendix E.

Overall Results. As shown in Figure 1, all large language models display significant translationese patterns in both English-Chinese and German-English translations, with an average of 45.0% and 51.1% of document-level translations displaying translationese for English-Chinese and German-English translations, respectively. We first examine model translations under the "direct" translation prompt setting. For English-Chinese translation, larger models generate more natural translations (GPT4 v.s. GPT3.5 and ALMA-13B v.s. ALMA-7B), and specialized translation models (ALMA) generate fewer translationese errors compared to general chat models like Mistral-7B, GPT-3.5, and GPT-4. For instance, ALMA-13B produces 36.0% of documents with translationese,

whereas the lowest-performing model, Mistral-7B, exhibits a rate of 76.0%. For German-English translation, all models demonstrate minimal variati on. This discrepancy may stem from the fact that most LLMs are pre-trained on an unbalanced corpus dominated by English, with significantly varying proportions of other languages. Regarding types of translationese errors, unnatural sentence flow errors occur more frequently than unnatural phrase flow errors; averaged error annotation counts are 3549.0 versus 1690.0 for English-Chinese translations and 1655.0 versus 311.7 for German-English translations. Examples of translationese cases can be found in Appendix F.

Prompting LLMs for Reducing Translationese. We explore the effects of the two alternative prompts: "specified" and "polish" prompt. Interestingly, incorporating specific requirements (i.e., "specified") in prompts that intend to enhance naturalness does not consistently reduce the rate of translationese errors; in some cases, it may even worsen the translation quality. For instance, under specified prompts, GPT-4 exhibits an increase in translationese errors, with the proportion rising from 0.50 to 0.53. Conversely, refining translations generated by the LLM itself ("polish") effectively and steadily reduces translationese errors. In particular, GPT-4 decreases the proportion of translationese from 43% to 25% through self-polishing its own translations. This indicates that it is not style-constrained prompts that promote natural generation but rather the task formats themselves—namely "translate" and "polish". In other words, *while LLMs pre-trained on extensive native utterances can generate more natural translations, this potential is not realized within a "translation" prompt.* The subsequent sections will explore the supervised training phase, where LLMs are instructed to perform various generation tasks, to investigate the origins of "unexpected" *unnatural translations* they generate despite their exposure to massive amounts of *natural* language during pre-training.

4 Tracing Translationese in Supervised Training Data

To investigate the origins of unnatural translations produced by LLMs, we first analyze the inherent preference of LLMs for natural generations and subsequently examine potential biases introduced during supervised training. We contend that

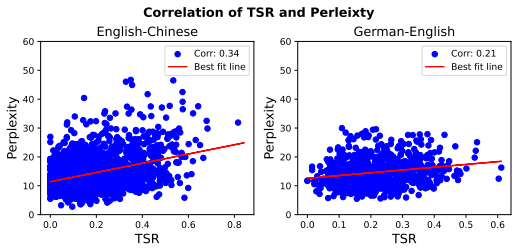

Figure 2: Correlation between the human-annotated translation span ratio (TSR) and LLM-generated perplexity.

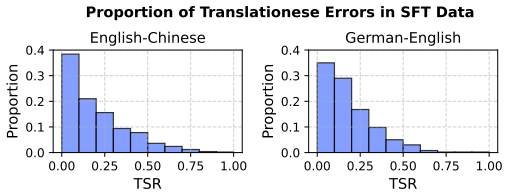

Figure 3: Proportions of supervised training instances exhibiting different levels of translationese errors (TSR).

LLMs trained on extensive corpora have the potential to distinguish unnatural generations, offering a reliable sign of generation naturalness. Previous studies (Aranberri, 2020; Bizzoni and Lapshinova-Koltunski, 2021; Jalota et al., 2023; Kuwanto et al., 2024) use target language model perplexity as a metric for translationese, where higher perplexity indicates less natural generation. However, these studies rely on language models trained on limited target-language corpora. In this work, we employ Llama-3.1-8B (Dubey et al., 2024), a large language model pre-trained on vast multilingual data that exhibits exceptional multilingual capabilities, to assess generation naturalness. Specifically, we calculate the perplexity of each translation, excluding the source text context, using Llama-3.1-8B and analyze its correlation with the human-annotated translation span ratio (TSR). As illustrated in Figure 2, despite being measured at different granularities (document-level versus span-level), these two metrics exhibit a positive correlation, particularly evident in English-Chinese translations, where higher perplexity corresponds to an increased ratio of spans identified as translationese errors.

We hypothesize that biased data in supervised training significantly contributes to translationese patterns, even though pre-trained LLMs favor natural sequences. As suggested by previous work (Xu et al., 2024a,b), supervised training data for LLM translation systems consists of test and validation

data from existing benchmark datasets (e.g., WMT and Flores (Costa-jussà et al., 2022)). However, these test datasets still exhibit translationese errors (Zhang and Toral, 2019), potentially introducing biases during supervised training. To quantify these biases, we sample 500 instances of English-Chinese and German-English translations from the ALMA training set (Xu et al., 2024a,b), asking the three expert translators to annotate the translationese spans for each instance (Details in Appendix G). Translation span ratios from the 3 annotators are computed and averaged, with results shown in Figure 3. A notable percentage of sentences contains over 20% spans identified as translationese: 40.4% for English-Chinese and 34.2% for German-English instances. The majority of errors stem from overly literal translation patterns, causing unnatural sentence- or phrase-level flows. This suggests that during supervised training, the LLM may develop a bias towards interpreting the "translation" task as a direct transformation from source to target, overemphasizing faithfulness at the expense of naturalness.

5 Mitigating Translationese from Supervised Training

In this section, we validate our hypothesis by addressing translationese biases in SFT and empirically evaluating translation naturalness.

5.1 Training Settings

We primarily adopt the training configurations from ALMA (Xu et al., 2024a) to develop LLMs for English-Chinese and German-English translation. For parallel training data, we extract instances for both translation directions (En-Zh and De-En) from the ALMA training set (WMT'17 to WMT'21 and Flores-200 (Costa-jussà et al., 2022)), resulting in a total of 31,621 parallel training instances. To construct the development set, we randomly select 10% of the training data. For evaluation, we assess models using our collected **document-level** datasets as well as **sentence-level** test sets from WMT'22. We use Llama-3.1-8B and Qwen-2.5-7B (Bai et al., 2023) as base models due to their superior multilingual capabilities. Training details are presented in Appendix H.

5.2 Evaluation Metrics

We use both automatic and human evaluation metrics to assess the translation naturalness.

Automatic Evaluation. As discussed, **perplexity** (PPL) is an effective indicator of generation naturalness (Jalota et al., 2023; Kuwanto et al., 2024). Following previous work (Aranberri, 2020; Zhang and Toral, 2019; Jalota et al., 2023; Riley et al., 2020), we consider two additional metrics: **lexical density** (Lex.) and **length variance** (Len.). Lexical density is defined as the ratio of content words to total words, as translationese typically exhibits lower lexical complexity and a reduced proportion of content words (adverbs, adjectives, nouns, and verbs) (Scarpa et al., 2006). We use Stanza (Qi et al., 2020) to extract part-of-speech tags and content words accordingly. Both machine translation (MT) systems and human translators typically refrain from restructuring the source sentence, adhering instead to prevalent sentence structures in the source language. Consequently, this practice yields translations that closely match the length of the original sentences. For each source-target pair (x, y), the length variety is calculated as: $\frac{||x|-|y||}{|x|}$. For translation quality estimation, we utilize `Unbabel/wmt22-cometkiwi-da` to compute and report COMET scores (Rei et al., 2022).

Human Evaluation. We ask the three expert translators to rank translations generated by different models in accordance with the annotation guidelines outlined in Section 3.2. Unlike previous tasks, their focus is solely on ranking translations rather than identifying fine-grained spans (Details in Appendix I).

5.3 Improving Naturalness of Training Data

As suggested in Section 3.3, using LLMs to polish existing translations can enhance translation naturalness. To mitigate translationese bias in SFT data, we use the polish prompt to let GPT-4 refine the golden references (Appendix A). Subsequently, we fine-tune LLMs with these polished translations, referred to as "**SFT-Polish**". Additionally, to ablate knowledge distillation from GPT-4, we use GPT-4 to generate direct translations of the source training instances, termed "**SFT-KD**". Table 2 compares translation naturalness between the baseline "SFT" method and other approaches.

As shown in the Table, addressing translationese bias in SFT data effectively mitigates model translationese for both base LLMs, with SFT-Polish yielding consistent improvements across all automatic metrics, i.e., higher lexical densities, increased length variability, and reduced perplexi-

Training	Document-level Translation						Sentence-level Translation					
	En-Zh			De-En			En-Zh			De-En		
	Lex.↑	Len.↑	PPL↓	Lex.↑	Len.↑	PPL↓	Lex.↑	Len.↑	PPL↓	Lex.↑	Len.↑	PPL↓
	Llama-3.1-8B											
SFT	0.509	0.639	13.8	0.421	0.079	15.0	0.500	0.377	103.3	0.415	0.150	84.2
SFT-KD	0.509	0.648	14.3	0.424	0.078	14.4	0.503	0.406	104.9	0.415	0.153	88.1
SFT-Polish	0.522	0.717	11.9	0.438	0.080	13.8	0.514	0.466	90.0	0.419	0.165	72.7
	Qwen-2.5-7B											
SFT	0.511	0.600	13.8	0.418	0.077	14.8	0.508	0.279	101.6	0.409	0.136	88.8
SFT-KD	0.513	0.651	13.9	0.424	0.068	14.7	0.505	0.272	104.2	0.415	0.129	88.4
SFT-Polish	0.523	0.687	12.1	0.436	0.073	14.3	0.518	0.317	87.3	0.419	0.139	71.1

Table 2: Automatic evaluation of translation naturalness at both sentence and document levels across different training methods, where a red background indicates the best performance and a blue one signifies the worst.

Direction	SFT	SFT-KD	SFT-Polish
En-Zh	2.3	2.2	**1.4**
De-En	2.3	2.0	**1.7**

Table 3: Average ranks for various SFT methods. Lower values indicate better performance.

Training	Llama-3.1-8B		Qwen-2.5-7B	
	En-Zh	De-En	En-Zh	De-En
SFT	80.0	80.5	73.8	74.0
SFT-KD	81.5	**81.2**	**74.7**	75.3
SFT-Polish	**81.8**	81.0	74.2	**75.6**

Table 4: Translation quality evaluation (COMET).

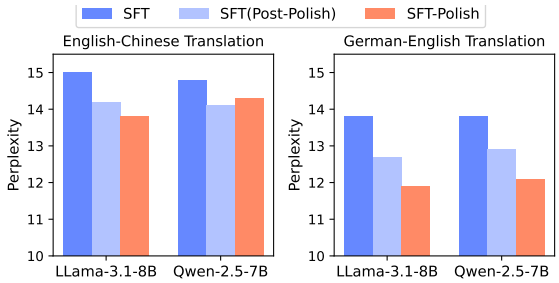

Figure 4: Comparison of naturalness between inference-time (Post-Polish) and training-time polishing (Polish).

ties. Specifically, the perplexities of translations from SFT-Polish are significantly lower than those from SFT and SFT-KD ($p < 0.01$), with average reductions of 7.8 for English-Chinese and 7.7 for German-English translations. In contrast, direct knowledge distillation from GPT-4 fails to enhance translation naturalness and may even degrade it in certain cases. This finding suggests that using LLMs such as GPT-4 to directly translate training data can not rectify existing translationese bias, as these LLMs may already be influenced by biases introduced during supervised training for translation tasks. Nevertheless, LLMs can improve naturalness through alternative task formats such as polishing.

As shown in Table 3, human evaluations of translations from models fine-tuned on Llama-3.1-8B corroborate the automatic assessments: SFT-Polish achieves the highest rankings and demonstrates strong inter-annotator agreement in both directions (details regarding inter-annotator agreement are provided in Appendix I). Translation quality estimation on the WMT test sets, as shown in Table 4, indicates that both SFT-KD and SFT-Polish significantly enhance translation quality ($p < 0.01$). Table 5 highlights the improvements achieved by SFT-Polish, such as transforming overly literal German-to-English translations like "Lots of fantasy, lots of complicated names, a dazzling look" into the more stylistically natural "Rich in fantasy, brimming with complex characters, and boasting stunning visuals" (see Appendix J for additional examples).

Additionally, we compare SFT-Polish models, which are trained on polished data, with SFT-Post-Polish models that employ GPT-4 to refine translations produced by SFT models. As shown in Figure 4, incorporating polishing during both training and inference improves translation naturalness, as indicated by reduced perplexities. Nevertheless, training on polished training instances results in more substantial improvements in translation naturalness, further supporting our hypothesis that translationese is predominantly shaped during supervised training.

	English-to-Chinese
Source	I've looked into it and I can see that your area is currently having a high volumes of order that is why they were assigning a rider for your order.
SFT	我已经调查过了，你的地区订单量非常大，才会把骑手分配给你的订单。
SFT-KD	我已经调查过了，你的地区当前订单量很大，这就是为什么他们会为你的订单安排骑手的原因。
SFT-Polish	我已经调查了情况，你的地区当前订单量很大，因此才有骑手为你配送订单。
	German-to-English
Source	Viel Fantasy, viele komplizierte Namen, eine atemberaubende Aufmachung: "Arcane", die Serie aus dem "League of Legends"-Computerspiel-Universum, ist vor kurzem auf Netflix gestartet.
SFT	Lots of fantasy, lots of complicated names, a dazzling look: "Arcane", the series from the "League of Legends" computer game universe, recently launched on Netflix.
SFT-KD	A lot of fantasy, many complicated names, a breathtaking setup: "Arcane", the series from the "League of Legends" video game universe, has recently launched on Netflix.
SFT-Polish	Rich in fantasy, brimming with complex characters, and boasting stunning visuals: "Arcane", the series set in the "League of Legends" video game universe, has recently premiered on Netflix.

Table 5: Case study of different model translations.

5.4 Filtering Unnatural Training Instances

An alternative approach to mitigate translationese bias involves filtering out unnatural training references before supervised training. We take perplexity as a measure of naturalness, allowing us to rank training instances and exclude the least natural subset. Experiments are conducted using Llama-3.1-8B. The results are illustrated in Figure 5, which displays the relationship between translation naturalness and quality on sentence-level WMT test sets relative to the proportion of filtered training instances. As shown in Figure 5, filtering up to 40% of the least natural references consistently enhances translation naturalness. Moreover, moderate filtering also improves translation quality. Specifically, a filtering proportion of 20% yields improvements in both metrics. However, excessive filtering adversely affects both naturalness and quality.

5.5 Generalization to More Languages

We extend our hypothesis to additional languages and evaluate the effectiveness of SFT-Polish. Specifically, we focus on translating from English to two high-resource languages: German (De) and Russian (Ru), as well as two moderate-resource languages: Czech (Cs) and Icelandic (Is). We use the same training and test sets from ALMA (Xu et al., 2024a). To train a multilingual translation model based on Llama-3.1-8B. We combine the additional training data with the original training set in Section 5.1. The naturalness of translations for these four languages is presented in Table 6. SFT-Polish generates translations with an average perplexity decrease of 7.6. In particular, the perplexity decreases from 56.5 to 40.0 for English-German translation. Our results demonstrate that polishing the training data consistently and significantly ($p < 0.01$) reduces translationese bias across all four languages, yielding a more natural translation.

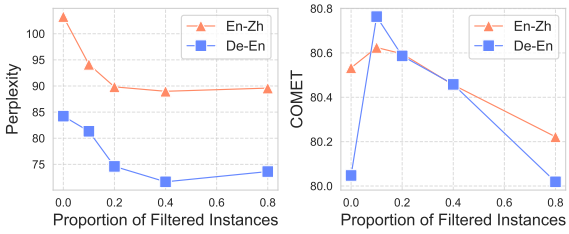

Figure 5: Translation naturalness and quality w.r.t. filtered training samples.

Training	En-Is	En-Cs	En-De	En-Ru
SFT	27.0	59.9	56.5	42.8
SFT-Polish	**24.9**	**50.9**	**44.0**	**35.9**

Table 6: Generation naturalness (perplexity) of translations from English to four additional languages.

6 Conclusion

In this work, we revealed how translationese, a long-standing issue in machine translation, persists even in state-of-the-art LLMs due to biases introduced during supervised training. Systematic analysis demonstrated the high prevalence of unnatural translations across multiple models and language pairs, attributed to training data with inherent translationese patterns. By leveraging techniques such as refining golden references and filtering unnatural instances, we achieved significant improvements in translation naturalness, confirming the potential of LLMs to align closer to native linguistic patterns. These findings underscored the importance of addressing data quality and training methodologies in developing robust and natural translation systems. Future research should extend these approaches to a broader range of language pairs and domains.

Limitations

While this study provides valuable insights into the issue of translationese in LLM-generated translations, several limitations should be acknowledged. First, due to the significant costs in time and resources required for human annotations, the evaluation primarily focuses on English-Chinese and German-English translations, which may limit the generalizability of the findings to other language pairs, especially low-resource or morphologically rich languages. Second, despite efforts to include a broad range of LLM translation systems, there are still other models and architectures that warrant further exploration. Finally, while human and automatic evaluations are employed, subjective biases in human annotations and the limitations of current automatic metrics could influence the assessment of translation naturalness. Addressing these limitations in future work could enhance the robustness and applicability of the findings.

Ethic Considerations

The data utilized in this study is web-crawled from publicly available sources, or obtained from publicly available datasets designed for academic research and contains no sensitive information. These datasets, including sources such as WMT and Flores, are freely accessible for non-commercial use, and their legality for academic purposes has been confirmed by our institution's legal advisors.

Our data construction involves human annotations to identify translationese patterns (Section C and Section G) and rank LLM translations (Section I). All annotators are tasked with reviewing translations, ensuring that no personal or sensitive information is included in the process. Three expert translators with advanced degrees in Linguistics or related fields are hired for annotation work of both translation directions. Before conducting formal annotations, they undergo a training phase that includes annotating 100 samples to ensure consistency and accuracy. Subsequently, they completed the aforementioned formal annotation tasks. Annotators are paid for both their training and formal annotation work at a rate of $16 per hour, determined based on the average annotation time for the training samples. This rate is designed to ensure fair and ethical compensation. Each annotator spends a total of 216 hours on the annotation (for English-Chinese), or 192 hours (for German-English), with compensation of $3,456 or $3,072, respectively.

No datasets are created that involve unethical content, and we make every effort to remove any data points that could potentially cause ethical concerns. We comply with the terms set by companies offering commercial LLM APIs and extend our gratitude to all collaborators for their invaluable support in utilizing these APIs. Additionally, our findings and methodologies aim to improve translation quality and do not promote harmful or biased content generation. By adhering to these standards, we ensure that this study was conducted ethically and responsibly.

References

Nora Aranberri. 2020. Can translationese features help users select an MT system for post-editing? *Proces. del Leng. Natural*, 64:93–100.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Yuri Bizzoni and Ekaterina Lapshinova-Koltunski. 2021. Measuring translationese across levels of expertise: Are professionals more surprising than students? In *Proceedings of the 23rd Nordic Conference on Computational Linguistics, NoDaLiDa 2021, Reykjavik, Iceland (Online), May 31 - June 2, 2021*, pages 53–63. Linköping University Electronic Press, Sweden.

Franck Burlot and François Yvon. 2018. Using monolingual data in neural machine translation: a systematic study. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 144–155, Brussels, Belgium. Association for Computational Linguistics.

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2024. Iterative translation refinement with large language models. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1), EAMT 2024, Sheffield, UK, June 24-27, 2024*, pages 181–190. European Association for Machine Translation (EAMT).

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Y. Sun, Skyler Wang, Guillaume

Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Meet Doshi, Raj Dabre, and Pushpak Bhattacharyya. 2024. Pretraining language models using translationese. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5843–5862, Miami, Florida, USA. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Koel Dutta Chowdhury, Rricha Jalota, Cristina España-Bonet, and Josef Genabith. 2022. Towards debiasing translation artifacts. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3983–3991, Seattle, United States. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proc. of EMNLP*, pages 489–500.

Zhaopeng Feng, Yan Zhang, Hao Li, Bei Wu, Jiayu Liao, Wenqiang Liu, Jun Lang, Yang Feng, Jian Wu, and Zuozhu Liu. 2024. Tear: Improving llm-based machine translation with systematic self-refinement. *Preprint*, arXiv:2402.16379.

Martin Gellerstam. 1986. Translationese in swedish novels translated from english.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based phrase-level prompting of large language models for machine translation. *Preprint*, arXiv:2302.07856.

Hippolyte Gisserot-Boukhlef, Ricardo Rei, Emmanuel Malherbe, Céline Hudelot, Pierre Colombo, and Nuno M. Guerreiro. 2024. Is preference alignment always the best option to enhance llm-based translation? an empirical analysis. *Preprint*, arXiv:2409.20059.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation. *CoRR*, abs/1803.05567.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Exploring Human-Like Translation Strategy with Large Language Models. *Transactions of the Association for Computational Linguistics*, 12:229–246.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *Preprint*, arXiv:2302.09210.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Rricha Jalota, Koel Dutta Chowdhury, Cristina España-Bonet, and Josef van Genabith. 2023. Translating away translationese without parallel data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 7086–7100. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Wenxiang Jiao, Jen-tse Huang, Wenxuan Wang, Zhiwei He, Tian Liang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023a. ParroT: Translating during chat using large language models tuned with human translation and feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15009–15020, Singapore. Association for Computational Linguistics.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023b. Is chatgpt a good translator? yes with gpt-4 as the engine. *Preprint*, arXiv:2301.08745.

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

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme

Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

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).

Garry Kuwanto, Eno-Abasi Urua, Priscilla Amondi Amuok, Shamsuddeen Hassan Muhammad, Aremu Anuoluwapo, Verrah Otiende, Loice Emma Nanyanga, Teresiah W. Nyoike, Aniefon D. Akpan, Nsima Ab Udouboh, Idongesit Udeme Archibong, Idara Effiong Moses, Ifeoluwatayo A. Ige, Benjamin Ajibade, Olumide Benjamin Awokoya, Idris Abdulmumin, Saminu Mohammad Aliyu, Ruqayya Nasir Iro, Ibrahim Said Ahmad, Deontae Smith, Praise-EL Michaels, David Ifeoluwa Adelani, Derry Tanti Wijaya, and Anietie Andy. 2024. Mitigating translationese in low-resource languages: The storyboard approach. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 11349–11360. ELRA and ICCL.

Zhuoyuan Mao and Yen Yu. 2024. Tuning llms with contrastive alignment instructions for machine translation in unseen, low-resource languages. *Preprint*, arXiv:2401.05811.

Jingwei Ni, Zhijing Jin, Markus Freitag, Mrinmaya Sachan, and Bernhard Schölkopf. 2022. Original or translated? a causal analysis of the impact of translationese on machine translation performance. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5303–5320, Seattle, United States. Association for Computational Linguistics.

Eugene Albert Nida and Charles Russell Taber. 1982. The theory and practice of translation.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, and et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 3505–3506. ACM.

Vikas Raunak, Arul Menezes, Matt Post, and Hany Hassan. 2023. Do gpts produce less literal translations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1041–1050. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Parker Riley, Isaac Caswell, Markus Freitag, and David Grangier. 2020. Translationese as a language in "multilingual" NMT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7737–7746. Association for Computational Linguistics.

Federica Scarpa et al. 2006. Corpus-based quality assessment of specialist translation: A study using parallel and comparable corpora in english and italian. In *Insights into specialized translation*, pages 154–172. Peter Lang.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2024. Label Studio: Data labeling software. Open source software available from https://github.com/HumanSignal/label-studio.

Antonio Toral. 2019. Post-editese: an exacerbated translationese. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 273–281, Dublin, Ireland. European Association for Machine Translation.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Neural Information Processing Systems*.

Jiaan Wang, Fandong Meng, Yunlong Liang, Tingyi Zhang, Jiarong Xu, Zhixu Li, and Jie Zhou. 2023. Understanding translationese in cross-lingual summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3837–3849, Singapore. Association for Computational Linguistics.

Shira Wein and Nathan Schneider. 2024. Lost in translationese? reducing translation effect using abstract meaning representation. *Preprint*, arXiv:2304.11501.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024a. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024b. Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Yongjing Yin, Jiali Zeng, Yafu Li, Fandong Meng, and Yue Zhang. 2024. LexMatcher: Dictionary-centric data curation for LLM-based machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14767–14779, Miami, Florida, USA. Association for Computational Linguistics.

Jiali Zeng, Fandong Meng, Yongjing Yin, and Jie Zhou. 2023. Tim: Teaching large language models to translate with comparison. In *AAAI Conference on Artificial Intelligence*.

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 73–81, Florence, Italy. Association for Computational Linguistics.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

A Translation Prompt

We employ three types of prompts for translations using large language models. As illustrated in Table 7, all models utilize the basic translation prompt; however, the well-instructed GPT models (GPT-3.5 and GPT-4) incorporate two additional prompts: the specified prompt and the polish prompt.

B Data Statistics

The data statistics of the collected source documents are presented in Table 8.

C Translationese Span Annotation

Following the definition in Unbabel's guideline[1], in this work, we define translationese as too literal translations of the source. Through preliminary research, we generally categorized the issue into three subcategories: Unnatural Sentence Flow, Unnatural Phrase Flow, and Culture-specific Reference (e.g. Source: We don't walk under ladders. Target: 我们不会在梯子下行走). Notably, the first two categories are more prevalent in LLM translation (see examples in Appendix F); therefore, this study focuses primarily on these two types.

We give our annotators a brief guideline and make detailed explanations with examples corresponding to each error category. Then, annotators

[1] https://help.unbabel.com/hc/en-us/articles/6444304419479-Annotation-Guidelines-Typology-3-0#h_01G4EYRD4K2KR9WKZ9WVT1N71K

Translation Prompt	Please translate the following {source_language} text to {target_language}. ### Source text: {source_text} ### Translation:
Specified Prompt	Please translate the following {source_language} text to {target_language}, ensuring that the translation is fluent, accurate, and conforms to typical {target_language} expressions and style. ### Source text: {source_text} ### Translation:
Polish Prompt	Please polish the corresponding {target_language} translation of an {source_language} text, ensuring that the translation is fluent, accurate, and conforms to typical {target_language} expressions and style. ### Source text: {source_text} ### Original Translation: {target_text} ### Translation:

Table 7: Three types of prompts used in large language model translation. The first one is utilized for all models whereas the other two are only used in GPT models.

Direction	Domains	Avg. Tokens	#. Docs.
En-Zh	CNN, Arixv, Wikipedia, Quora	225.6	1,800
De-En	Focus, Quora	138.1	9,00

Table 8: Data statistics of document-level translations.

are required to highlight all spans characterized as translationese errors in the document-level translation. During annotation, all translations of one given source are provided sequentially as a batch for the convenience of comparisons among different models (note that annotators do not know which model generated each translation, and the appearance order of translated documents is shuffled). The guideline for span annotation is shown as follows (see also Table 11):

You will assess model translations of a source document, where each document may contain one or more sentences. Each target-language document is aligned with its corresponding source-language document, and both are displayed simultaneously on the annotation platform. For each model translation, identify and annotate spans with the specified error types. Annotate documents sequentially, as if reading them naturally. You may revisit and revise previously annotated documents as needed.

1. The key issues in this task are style errors and unnatural expressions (so-called translationese). You can label one expression as long as it seems to be strange from the perspective of the contemporary target language. To identify an error, highlight the relevant span of text, and select a category from the available options.

2. When identifying errors, please identify all errors within each translated document and be as fine-grained as possible. For example, if there are two separate unnatural phrases in one sentence, please annotate two phrases respectively instead of selecting the whole sentence.

3. Besides the three categories of style errors we provided, there are also some categories of translation errors for mistranslation situations. If it is not possible to reliably identify distinct errors because the translation is too badly garbled or is unrelated to the source, then mark a single Nontranslation error that spans the entire document.

D Annotation Implementation

Based on the above guideline, we develop a specialized annotation platform using Label Studio (Tkachenko et al., 2020-2024), as demonstrated in Figure 6.

The annotation tasks are conducted in batches, with each batch containing 180 translated documents corresponding to 20 source texts. As mentioned above, translations generated by different models from the same source text are presented simultaneously, but in a randomized order. Given the potential subjectivity in annotators' judgments on translationese, the results of annotation are subsequently reviewed by a senior annotator. This process aims to prevent significant disparities in annotating standards. Each batch of annotations takes approximately 16 hours for English-Chinese direction and 24 hours for German-English. The total

English-Chinese Translation			
Judge	**A-1**	**A-2**	**A-3**
A-1	-	0.592	0.742
A-2	0.592	-	0.603
A-3	0.742	0.603	-
German-English Translation			
Judge	**A-1**	**A-2**	**A-3**
A-1	-	0.753	0.587
A-2	0.753	-	0.553
A-3	0.587	0.553	-

Table 9: Inter-annotator agreement (Kendall's Tau scores) on naturalness voting.

time cost is 160 hours and 120 hours, respectively.

E TSR Scores

The evaluation of the translationese span ratio for all models under both translation directions is presented in Table 10.

F Case Study of Translationese

We demonstrate several real translation cases of both translationese errors in Table 12 (English-Chinese) and Table 13 (German-English).

G Sentence-level Annotation

Annotators are assigned another translation assessment task at the sentence level. They are required to follow the same guideline shown in Appendix C as well. Similarly, each sentence is aligned with a corresponding source sentence. Annotators are asked to read in sequential order, with permission to revise previous sentences. The total time cost is 16 hours (English-Chinese) and 24 hours (German-English), respectively.

H Training Details

All models are fine-tuned using LoRA (Hu et al., 2021) with a rank of 16, employing a batch size of 16 on an A100 GPU. The learning rate is set to 1×10^{-4} with a warmup ratio of 0.1. Training is conducted for three epochs, selecting the model that achieves the lowest validation loss. We perform training using Llama-Factory (Zheng et al., 2024) and leverage Deepspeed (Rasley et al., 2020) to accelerate training.

I Human Ranking

In the voting task, annotators are given a file in which each source document is aligned with three distinctive translations. They are required to rank the severity of translationese issues in each translation. A higher rank indicates less translationese and more natural language flow. When making judgments about translationese. Annotators still follow the guideline we provided for span annotation, but we do not provide a specific breakdown of the ranking scheme. The total time cost is 24 hours (English-Chinese) and 32 hours (German-English), respectively. The inter-annotator agreement evaluation is presented in Table 9.

J Case Study of SFT Methods

Cases of translations from SFT, SFT-KD and STF-Polish are also demonstrated in Table 14 (English-Chinese) and Table 15 (German-English).

Direction	ALMA-7B	ALMA-13B	Mistral-7B	GPT-3.5			GPT-4		
				Direct	Specified	Polish	Direct	Specified	Polish
En-Zh	0.19	0.18	0.32	0.22	0.23	0.20	0.20	0.17	**0.14**
De-En	0.23	0.23	0.22	0.21	0.22	0.20	0.21	0.21	**0.19**

Table 10: Translationese span ratios of different LLMs in English-Chinese and German-English translations.

Error Category	Description
Unnatural Sentence Flow	A sentence-level translation issue where the structure of the sentence is considered unnatural in the target language. This often occurs when complex sentence structures from the source language are directly translated, resulting in sentences that are difficult to read in the target language.
Unnatural Phrase Flow	A portion of text, larger than a single word or multiword expression, is a too literal translation of the source. The meaning of the source comes through in the target, but the overall feeling of the translation is unnatural.
Culture-specific Reference	The target text contains a culture-specific reference that's not appropriate or understandable to the intended target audience. An example of this is the use of jargon related to sports or other culture-specific features that are not necessarily understood in the environment of the target language.
Sensitive Content	The presence of sensitive information in the translation or source text, such as references to violence, war, etc.
Mistranslation	Minor errors including mistranslations, omissions, or over-translations.
Terminology	Errors related to the incorrect use of domain-specific terms or technical jargon.
Non-translation	Impossible to reliably characterize distinct errors (or the model repeatedly outputs meaningless contents)
Others	Errors that affect the readability and naturalness of the text but do not fit neatly into the other defined categories. Annotators should provide specific comments on these errors.

Table 11: Annotation Guideline in the present study

Error Category		Example
Unnatural Sentence Flow	**Source**	Our benchmarking findings can serve future research aiming to improve the generic capability of LMs on semantic phrase comprehension.
	Translation	我们的评测结果将为未来研究，旨在提升语言模型在语义表达理解任务中的普适能力，提供有价值的参考。
	Source	An analysis of a core cohort comprising 380 articles from multiple disciplines captures the most recent advancements in responsible AI.
	Translation	通过一个包括来自多个学科的380篇文章的核心队列的分析，捕捉了负责任AI的最新进展。
	Source	They both contribute to the development of a unified model that is highly generalizable, versatile, and comprehensible for time series analysis.
	Translation	二者共同促进了高度通用、多功能且易于理解的统一模型的发展，用于时间序列分析。
Unnatural Phrase Flow	**Source**	demonstrated remarkable improvements
	Translation	展示了显著的改进
	Source	demonstrating promising performance
	Translation	展示了有希望的性能
	Source	credit risk management is particularly core
	Translation	信用风险管理尤为核心

Table 12: Samples of translationese errors in large language model translation (English-Chinese).

Error Category		Example
Unnatural Sentence Flow	**Source**	So geht es nicht, findet die italienische Regierung und ließ Dutzende von elektrischen Fiat Topolinos beschlagnahmen.
	Translation	This is not acceptable, finds the Italian government and seized dozens of electric Fiat Topolinos.
	Source	Das zweite Gruppenspiel bestreitet die DFB-Elf fünf Tage später am 19. Juni in Stuttgart gegen Ungarn.
	Translation	The second group game will be played five days later on 19 June in Stuttgart against Hungary.
	Source	Nach meinem Wissen sind wir die Ersten in Deutschland, die das angewendet haben, sogar in Europa
	Translation	To the best of my knowledge, we are the pioneers in Germany in using it, even in Europe,
Unnatural Phrase Flow	**Source**	schufen aber einen rockigeren sound
	Translation	crafted a grittier sound
	Source	sie sich stark mit anderen Arten vermischt
	Translation	it mixes strongly with other species
	Source	sind wir in der USA in Urlaub gewesen
	Translation	we vacationed in the US

Table 13: Samples of translationese errors in large language model translation (German-English).

	Example
Source	50 countries had purchased rights to the telecast, and when the numbers were added up, it was estimated that 300 million people worldwide had watched the fight, including 27.5 million people in the UK, about half the national population at the time. To put into perspective how big the fight was just in the UK, more people watched the fight than the closing ceremony of the 2012 Olympics games (which was in London) and the full coverage of the 2011 Royal wedding of Prince William and Kate (a British royal wedding), despite neither Frazier or Ali originating from the UK, and the fight taking place in America.
SFT	50 个国家购买了转播权，当这些数字相加时，估计有3 亿人次在全世界观看了这场比赛，包括2750 万英国观众，这几乎是当时英国人口的一半。为了将这场比赛在英国的规模置于合适的位置，我们可以看到，观众人数比2012 年伦敦奥运会闭幕式和2011 年威廉王子和凯特的婚礼（一场英国皇室婚礼）更高，尽管弗雷泽和阿里都不是英国人，比赛也在美国举行。
SFT-KD	50个国家购买了广播权，当数字相加时，全球估计有3亿人观看了这场比赛，其中包括英国的2750万人，这约占当时国民总数的一半。为了让大家对英国本地观看人数有一个感性认识，观看这场比赛的人数超过了2012年伦敦奥运会闭幕式和2011年威廉王子与凯特的皇家婚礼的全覆盖（尽管弗雷泽和阿里都不是来自英国，而且这场比赛是在美国举行的）。
SFT-Polish	50个国家已经购买了电视转播权。加总后，全球观众估计超过了3亿人次，其中包括英国的2750万人，约占当时英国国民总数的一半。为了更准确地理解这一拳赛在英国的影响力，尽管弗雷泽和阿里均非英国人，并且该比赛是在美国举行的，观众人数仍然超过了2012年伦敦奥运会闭幕式和2011年威廉王子与凯特的联合王国皇室婚礼的观众人数。
Source	Rice's biggest success was her first novel, "Interview with the Vampire," which was published in 1976 and introduced the character of the vampire Lestat, who would be the central character in the 13-book Chronicles series, the most recent of which was published in 2018.
SFT	米勒的最大成就在于她的第一部小说《吸血鬼采访》，该小说于1976 年出版，引入了吸血鬼莱斯特这个角色，这位角色将成为13 部系列小说中的主角，这些系列小说中的最新一部于2018 年出版。
SFT-KD	米尔斯最大的成功是她的第一部小说《吸血鬼的采访》，该小说于1976年出版，引入了吸血鬼莱斯特的角色，这位角色将成为13部书籍系列《纪事》的核心人物，该系列的最新作品于2018年出版。
SFT-Polish	米尔斯最大的成就是她的首部小说《吸血鬼之访谈》，该书于1976年出版，首次引入了吸血鬼莱斯特特的角色，这位角色后来成为了《吸血鬼纪事》系列中的核心人物，该系列共13部，最新的一部在2018年出版。

Table 14: Samples of translations from SFT, SFT-KD and SFT-Polish (English-Chinese).

Example	
Source	Trockener Rasen und gekieste Wege mögen für andere heimische Nacktschnecken ein Problem sein, nicht aber für Arion vulgaris. Sie vermehrt sich Experten zufolge schneller, frisst mehr und setzt sich notfalls zum Fressen in die pralle Sonne, ohne Schaden zu nehmen. Zudem zeigen Erbgutanalysen, dass sie sich stark mit anderen Arten vermischt - und sich auf diese Weise womöglich immer neue günstige Eigenarten für die jeweilige Umgebung aneignet. Und als wäre das alles nicht genug: An einer ausgewachsenen Arion vulgaris haben - von Indischen Laufenten abgesehen - kaum Fressfeinde Interesse. Auch manche Laufkäfer können junge Wegschnecken oder deren Eier fressen.
SFT	Dry lawn and cobbled paths may be a problem for other native slug species, but not for Arion vulgaris. Experts say that it multiplies faster, eats more, and if need be, sets itself down in the full sun to eat, without suffering any damage. In addition, DNA analyses show that it mixes strongly with other species - and thus acquires new favorable characteristics for the respective environment. And if that were not enough: Only Indian runners have any interest in eating an adult Arion vulgaris. Even some beetle larvae can eat young slugs or their eggs.
SFT-KD	Dry lawns and graveled paths may be a problem for other native slugs, but not for Arion vulgaris. According to experts, it reproduces faster, eats more, and if necessary, basks in the full sun to feed, without any harm. Moreover, genetic analyses show that it strongly hybridizes with other species - and in this way, possibly acquires ever new advantageous characteristics for its environment. And if all that were not enough: Among fully grown Arion vulgaris, hardly any predators have an interest - except for Indian leaf turtles. Some ground beetles can also eat young slugs or their eggs.
SFT-Polish	Dry lawns and gravel paths may pose challenges for other native slugs, but not for Arion vulgaris. Experts note that it reproduces more rapidly, consumes more, and, if necessary, basks in the full sun without suffering any harm. Additionally, genetic analyses reveal that it has extensively interbred with other species, potentially acquiring new adaptive traits for its specific environments. Moreover, apart from the Indian land tortoise, few predators show interest in an adult Arion vulgaris. Some ground beetles, however, can consume young slugs or their eggs.
Source	"Tatsächlich haben meine Frau und ich 3 Dosen des Biontech-Pfizer-Impfstoffs erhalten", schrieb Şahin am Freitag in einem Beitrag im Netzwerk "LinkedIn"
SFT	"In fact, my wife and I have received 3 doses of the Biontech/Pfizer vaccine", Şahin wrote on Friday in a contribution to the LinkedIn network.
SFT-KD	Indeed, my wife and I have received 3 doses of the BioNTech-Pfizer vaccine", Şahin wrote on Friday in a post on the "LinkedIn" network.
SFT-Polish	"Indeed, my wife and I have received three doses of the BioNTech-Pfizer vaccine", Şahin wrote in a Friday post on the LinkedIn network.

Table 15: Samples of translations from SFT, SFT-KD and SFT-Polish (German-English).

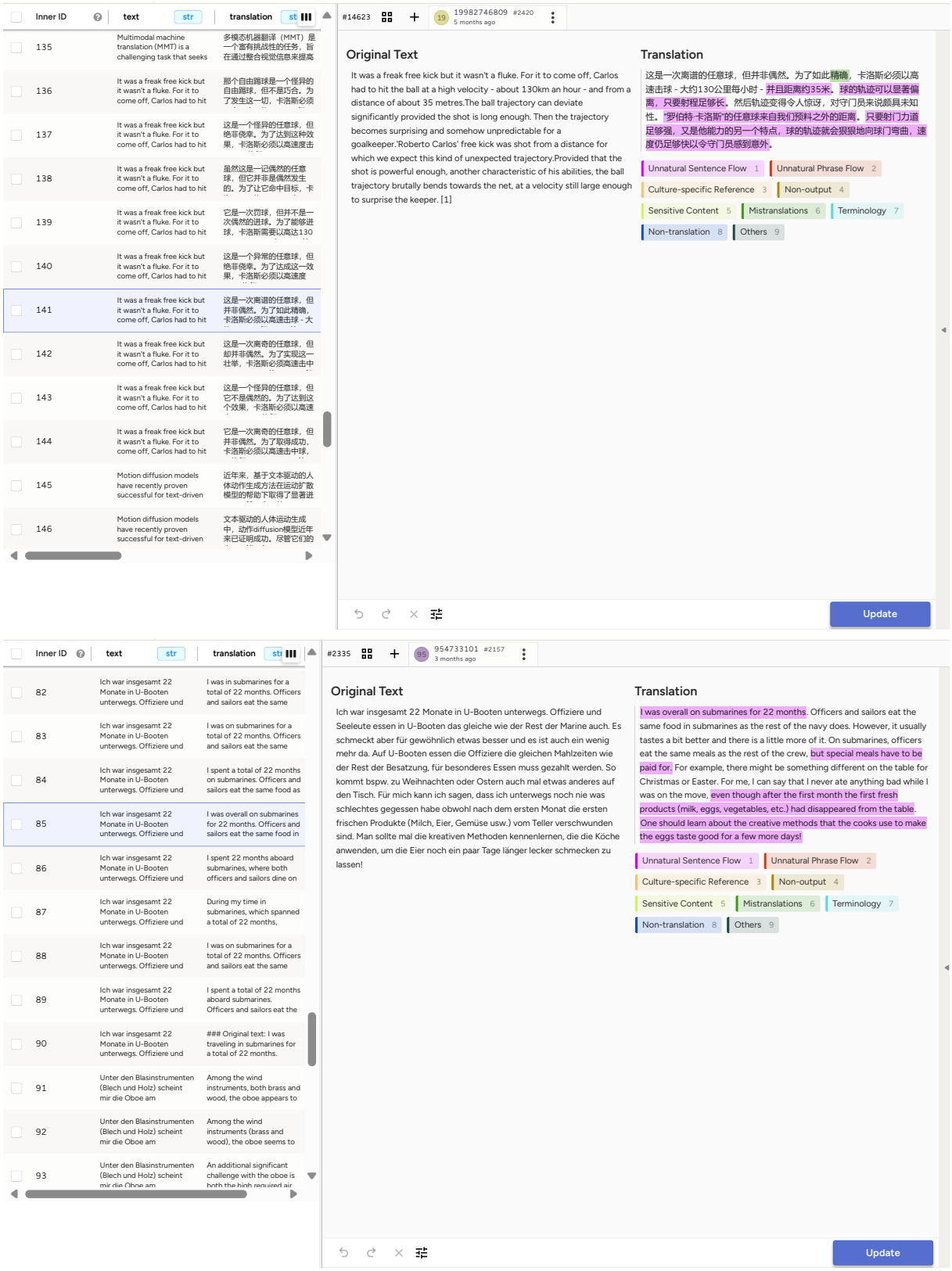

Figure 6: Annotation platform demonstration (English-Chinese and German-English).