xTower: A Multilingual LLM for Explaining and Correcting Translation Errors

Anonymous ACL submission

Abstract

While machine translation (MT) systems are achieving increasingly strong performance on benchmarks, they often produce translations with errors and anomalies. Understanding these errors can potentially help improve the translation quality and user experience. This paper introduces xTower, an open large language model (LLM) built on top of TowerBase designed to provide free-text explanations for translation errors in order to guide the generation of a corrected translation. The quality of the generated explanations by xTower are assessed via both intrinsic and extrinsic evaluation. We ask expert translators to evaluate the quality of the explanations across two dimensions: *relatedness* towards the error span being explained and *helpfulness* in error understanding and improving translation quality. Extrinsically, we test xTower across various experimental setups in generating translation corrections, demonstrating significant improvements in translation quality. Our findings highlight xTower's potential towards not only producing plausible and helpful explanations of automatic translations, but also leveraging them to suggest corrected translations.[1]

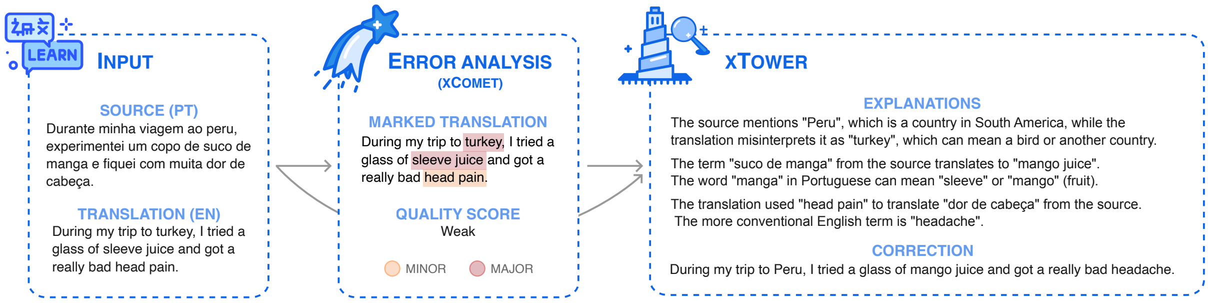
1 Introduction

Neural machine translation (MT) systems have made significant strides in recent years. However, despite their high performance on standard benchmarks, these systems often produce translations that contain errors and anomalies. Common methods for evaluating MT quality, such as BLEU (Papineni et al., 2002), and neural metrics like Comet (Rei et al., 2020) and BLEURT (Sellam et al., 2020), provide only a numerical score reflecting overall translation quality. Recent metrics like xComet (Guerreiro et al., 2023a) and AutoMQM (Fernandes et al., 2023) highlight error spans to justify their scores but do not offer explanations about the nature of these errors. InstructScore, a recent work by Xu et al. (2023), leverages large language models (LLMs) to provide a quality score conditioned on built-in error detection and explanations. However, InstructScore primarily functions as a *reference-based metric*, using explanations as a means to improve score estimates via meta-feedback/finetuning.

In this paper, we introduce xTower (Figure 1), a LLM specifically tailored to produce high-quality explanations for translation errors and to utilize these explanations to suggest corrections through chain-of-thought prompting (Wei et al., 2023). xTower is built on TowerBase 13B (Alves et al., 2024), a strong open multilingual LLM for MT-related tasks. Unlike InstructScore, xTower can operate *without the need for reference translations* while also considering information contained in the source sentence. Moreover, xTower is designed to be agnostic about the source of error spans, as they can be obtained manually via human annotation or via automatic tools. In this work, we experiment with both. For the automatic case, we leverage xComet (Guerreiro et al., 2023a). This modular approach offers flexibility to experiment with span-level error annotations from various sources, and easily incorporate future improvements in span error detection tools without requiring retraining.

We evaluate xTower's explanations both intrinsically and extrinsically. Intrinsically, we employ human evaluation to score explanations on two dimensions: *relatedness* to the error spans being explained (§4.2) and *helpfulness* in guiding towards a better translation (§4.3). Extrinsically, we assess xTower's ability to suggest translation corrections (§5), experimenting with different error span sources (human vs. predicted). We compare xTower's performance against leading closed and open LLMs, such as GPT-3.5 Turbo, Mixtral 8x7B, and TowerInstruct 13B. Our findings demon-

[1] xTower will be publicly released upon acceptance.



Durante minha viagem ao peru, experimentei um copo de suco de manga e fiquei com muita dor de cabeça.

During my trip to turkey, I tried a glass of sleeve juice and got a really bad head pain.

During my trip to turkey, I tried a glass of sleeve juice and got a really bad head pain.

The source mentions "Peru", which is a country in South America, while the translation misinterprets it as "turkey", which can mean a bird or another country.

The term "suco de manga" from the source translates to "mango juice". The word "manga" in Portuguese can mean "sleeve" or "mango" (fruit).

The translation used "head pain" to translate "dor de cabeça" from the source. The more conventional English term is "headache".

During my trip to Peru, I tried a glass of mango juice and got a really bad headache.

Figure 1: Illustration of our approach. In this example, the input consisting of a source and a translation is passed to xComet, which annotates the translation with error spans and produces a (discretized) quality score. The full input, marked translation, and quality score are passed to xTower, which, in turn, produces an explanation for each error span along with a final suggestion for a new, corrected translation.

strate that xTower improves error interpretability by providing explanations that effectively relate to the marked errors. Expert translators endorse these explanations as helpful for understanding translation errors and generally useful for improving translations, particularly for English-German. Furthermore, prompting xTower with these explanations leads to improved translation corrections. Overall, our main contributions are:

- We introduce xTower, a multilingual LLM that generates free-text explanations for translation errors and provides corrected translations.
- We conduct extensive human evaluations to assess the relatedness and helpfulness of xTower's explanations, linking their results with dedicated qualitative analyses.
- We evaluate xTower's corrected translations across multiple language pairs and experimental setups, showing significant improvements in translation quality.

2 Background

In this section, we provide an overview of the key components and concepts relevant to our work.

Tower. Alves et al. (2024) developed a suite of state-of-the-art multilingual LLMs via continued pretraining of LLaMA2 (Touvron et al., 2023) — TowerBase— and supervised finetuning for translation-related tasks — TowerInstruct. Tower is trained to handle diverse tasks such as MT, automatic post-editing, and grammatical error correction. However, it lacks support for error-annotated inputs and cannot produce high-quality, span-level explanations for translation errors. xTower addresses these limitations by extending Tower— through distillation and finetuning — enabling it to provide explanations for translation errors and generate corrected translations.

MT Evaluation. Evaluating the quality of machine translations is a critical aspect of improving MT systems. Traditional metrics like BLEU (Papineni et al., 2002) and chrF (Popović, 2015) have been widely used to measure the accuracy of translations by comparing them to reference translations. However, these lexical metrics do not correlate well with human judgments (Freitag et al., 2023). More recent neural metrics, such as BLEURT (Sellam et al., 2020) and Comet (Rei et al., 2020), offer improved performance by finetuning pretrained neural models to predict translation quality. Still, they lack the ability to explain errors in human-interpretable terms. To this end, Rei et al. (2023); Guerreiro et al. (2023a) propose methods to highlight input words relevant to the output. However, highlighting input words offers a limited view of interpretability, as the end-user often needs additional information to understand what the error consists of and how it can be fixed. Our approach with xTower aims to bridge this gap by generating free-text explanations for translation errors, thus offering more insightful and detailed quality reports.

3 xTower

In this section, we provide details on the methodology behind xTower (Figure 1), a model built on top of TowerBase via distilled supervised finetuning (Tunstall et al., 2023).

3.1 Distillation

Data. We use GPT-4 to generate explanations for samples annotated with MQM spans and to generate a final translation correction.[2] Our dataset comprises English→German (EN-DE), English→Russian (EN-RU), and Chinese→English (ZH-EN) samples from the WMT 2022 Metric shared task (Freitag et al., 2022). Each error span is annotated by humans according to the MQM framework, which includes a severity rating such as minor or major. Detailed statistics about this dataset are provided in Appendix A. Overall, our distillation dataset consists of 33,442 samples containing 63,188 human-annotated error spans.[3]

Prompt. We use an XML format to obtain an "annotated translation", which includes the demarcations of error spans as tags alongside their severity as attributes. Following Farinha et al. (2022), we discretize the MQM quality score into buckets: weak, moderate, good, excellent, best. Table 1 shows a prompt example. As output, GPT-4 generates explanations for each marked error, followed by a corrected translation in the following format:[4]

- `Explanation for error`N: the explanation given to the N^{th} error span. Explanations for each error span are separated by newlines.
- `Translation correction`: the translation refinement produced by the model. Corrections are placed in a new line after the last explanation.

Notably, this ordering acts as chain-of-thought prompting (Wei et al., 2023). We collect outputs for referenceless and reference-based evaluation (by providing a reference translation in the input).

3.2 Finetuning

We obtained xTower by finetuning TowerBase-13B on a dataset that includes the GPT-4 generated explanations described in §3.1, and machine translation data from TowerBlocks, the dataset used to train TowerInstruct.[5] We combined all available data to train a single, multilingual model, instead of training separate models for each language pair. Moreover, following (Longpre et al., 2023), we employed a mixed prompt setting (zero-shot, few-shot) during training. As a result, xTower can handle both referenceless and reference-based k-shot prompts. Our training hyperparameters and configuration follows that used to train TowerInstruct (Alves et al., 2024).

Instruction:
You are provided with a Source, Translation, Translation quality analysis, and Translation quality score (weak, moderate, good, excellent, best). The Translation quality analysis contains a translation with marked error spans with different levels of severity (minor or major). Given this information, generate an explanation for each error and a fully correct translation.

Input:
English source: All were wearing avalanche beacons.
German translation: Alle trugen Lawinenschilder.
Translation quality analysis: Alle trugen <error1 severity="major">Lawinenschilder</error1>.
Translation quality score: weak

Output:
Explanation for error1: The term "Lawinenschilder" translates to "avalanche signs", which is a significant mistranslation of "avalanche beacons". Avalanche beacons are specific devices used to locate people trapped under snow after an avalanche, not signs. This error changes the meaning of the sentence and introduces confusion about the context.
Translation correction: Alle trugen Lawinensuchgeräte.

Table 1: Example of prompt used for generating explanations and translation corrections.

4 Explaining Translation Errors

In this section, we provide a detailed human evaluation of the quality of the explanations produced by xTower, which are obtained in a more realistic setting via referenceless prompting.

4.1 Experimental Setup

Data. We evaluate our models on MQM annotations from the WMT 2023 Metrics shared task test set (Kocmi et al., 2023), spanning three language pairs: English→German (EN-DE), Hebrew→English (HE-EN), and Chinese→English (ZH-EN). This dataset contains 24,781 samples with 69,564 human-annotated error spans. To obtain a fully automatic approach, we use error spans predicted by xComet-XL (Guerreiro et al., 2023a).[6] For a consistent evaluation, we also query xComet without references for our experiments. In total, we obtain a set of 108,507 spans, indicating that xComet has a higher tendency to predict errors. Detailed statistics are shown in Appendix A.

Prompting. We use the same prompt template as the one used in our distillation experiments, shown in Table 1. We use 0-shot prompting for all experiments involving xTower in this section.

Evaluation. While recent works propose frameworks to assess free-text explanations for classification tasks (Wiegreffe et al., 2021; Ramnath et al.,

[2] We use `gpt-4-0125` available from the OpenAI API.

[3] The dataset will be released upon acceptance.

[4] We manually inspected a few outputs to ensure reliability.

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

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

	EN-DE		ZH-EN	
Level	xCOMET	Human	xCOMET	Human
Explanation	3.5±1.5	4.4±1.6	3.4±1.6	4.3±1.7
Document	3.4±1.5	4.3±1.7	3.3±1.6	4.3±1.7
Correlation	0.96	0.89	0.96	0.96

Table 2: Relatedness scores (6-Likert scale) computed at explanation and document-level, along with the Spearman correlation between the two.

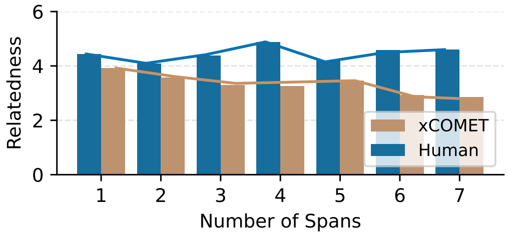


Figure 2: Relatedness according to the number of spans for xCOMET and human error spans.

2023; Joshi et al., 2023; Chen et al., 2023), applying a similar evaluation for MT is challenging due to the occurrence of multiple error spans with varied impact on translation quality. Therefore, we choose to assess our explanations through human evaluation and qualitative analysis (§4.4). The evaluation comprises the following two dimensions:

- **Relatedness**: The extent to which the explanation is related to the content of the error span.
- **Helpfulness**: The extent to which the explanation helps in understanding the nature of the error and in guiding towards a translation correction.

We present the setup and findings from both evaluations next. Human evaluation details and guidelines can be found in Appendix C.

4.2 Relatedness

A total of 6 annotators were employed for the task, evaluating samples marked with xCOMET and human-annotated error spans. 3 annotators assessed explanations for EN-DE, and other 3 for ZH-EN. For each language pair and error span source, we randomly sampled 50 translations, resulting in 200 examples in total. Inspired by the direct assessment and scalar quality metric (DA+SQM) scale used in MT evaluation (Kocmi et al., 2022), we asked annotators to rate explanations on a 6-point Likert scale: nonsense/unrelated (0), somewhat (2), mostly (4), and fully related (6).[7] Moreover, we asked annotators to rate the quality of explanations individually (**explanation-level**) and by looking at all explanations at once (**document-level**). The annotations were carried out on the Upwork platform.[8] We obtain an overall inter-annotator agreement, as measured via Spearman correlation (Pavlick and Tetreault, 2016), of 0.51 (EN-DE) and 0.40 (ZH-EN) at the explanation-level, and of 0.50 (EN-DE) and 0.37 (ZH-EN) at the document-level, suggesting a fair-to-moderate agreement among annotators, typical in explanations evaluation which is a subjective task (Wiegreffe et al., 2022; Kunz et al., 2022). Results are shown in Table 2.

Discussion. For human-annotated error spans, the overall relatedness scores range around 4.3, while for xCOMET spans the scores drop to around 3.2. This difference indicates that **the quality of error spans heavily impacts the quality of their explanations**. Nonetheless, for both cases, human ratings are in the 3-5 range, indicating that **xTower's explanations are *mostly* related to the error spans**. We also note a very high correlation between the quality of explanations assessed at the explanation and document-level, especially for human-annotated spans, indicating that the quality of explanations is consistent across granularities.

In Figure 2 we show how relatedness scores vary according to the number of error spans. We observe that, while the number of spans does not affect the relatedness of explanations produced for human-annotated error spans, they lead to a slight decrease of the relatedness scores when the spans are predicted by xCOMET. We hypothesize this is due to xCOMET overpredicting error spans (see Table 6).

4.3 Helpfulness

To quantify the idea of how helpful explanations are to the end user, we carried a new human evaluation with 4 of the same annotators from the previous task, and asked them to rate explanations based on two questions:

- Q1: How helpful is the explanation in improving the understanding of the nature of the error?
- Q2: How helpful is the explanation in guiding towards writing a better translation?

The rating is again performed on a 6-point Likert scale, ranging from less to more helpful. Moreover, we focus on studying the helpfulness of correct

[7] The full scoring rubric is provided in Appendix C.
[8] https://www.upwork.com

QUESTION	EN-DE	ZH-EN
Q1: error understanding	4.6±1.7	4.4±2.1
Q2: translation guidance	3.9±1.6	3.3±2.1
Correlation	0.85	0.72

Table 3: Helpfulness scores (6-Likert scale) for Q1 (understanding the nature of the errors) and Q2 (guiding towards better translations), along with the Spearman correlation between the two.

error spans only, in order to isolate the effect of providing accurate information towards improving error understanding. To this end, we filter out samples with an overall relatedness score lower than 4 and only use error spans labeled by humans. Table 3 shows the results.

Discussion. We find that annotators mark the explanations as being on average helpful (scores range in 4.4-4.6) in improving error understanding for both language pairs. Here, scores over 4 imply that "the explanation clearly identifies the error and provides relevant details about its nature". Furthermore, the usefulness of these explanations in guiding towards a potential correction ranges on average between 3.3-3.9, demonstrating that the explanations do hint towards a potential solution for correction, but they can be made more specific. For example, one of our expert annotators quoted:

> Many cases had a very clear explanation of the nature of the error, but in terms of helpfulness in guiding towards writing a correction, it was a bit less clear than the above-mentioned examples as they do not suggest a correction. Nonetheless, the explanation still correctly guides the editor to a post-edition.

4.4 Qualitative Analysis

Based on the annotators' feedback for the previous experiments, and by manually inspecting the annotated examples, we present a qualitative analysis of the explanations generated by XTOWER in Table 4. Our analysis reveals several interesting scenarios that highlight XTOWER's strengths and weaknesses. We categorize our findings into four main groups:

- **Correct Spans:** For error spans that correctly correspond to an error in the translation, explanations are **accurate** when they effectively detail the nature of the error, and **inaccurate** when they are unattached to the error, possibly suggesting wrong modifications.
- **Incorrect Spans**: Despite incorrect spans, explanations can still be **valuable** by pointing out that there are no errors in the translation. In other cases, they are mislead by the incorrect span and become **worthless** by being nonsensical to the error, possibly including a boilerplate suggestion for stylistic improvement.

We also estimate the prevalence and compute the average relatedness score of each category. Specifically, we consider explanations as accurate/valuable when their average relatedness score is larger or equal to 4, otherwise we consider them as inaccurate/worthless. This analysis indicates that XTOWER is not only capable of generating high-quality explanations when the error spans are correctly identified, but can also provide valuable explanations for incorrect spans, amounting to 59.1% of the cases with an average relatedness score of 5.3. However, over a quarter of all explanations (26.8%) either misidentify the nature of the error or provide generic, boilerplate suggestions. These findings suggest that while XTOWER has the potential to be a useful tool for automatic translation error analysis, there is still significant room for improvement, especially for cases where translation errors spans are incorrectly identified.

5 Correcting Translations

Having established that XTOWER can provide helpful and useful explanations, we turn to prompting it towards obtaining translation corrections.

5.1 Experimental Setup

Our setup for correcting translations follows the same design choices used for explaining translation errors in §4, including the test data (WMT23), source of error spans (human vs. XCOMET), and prompting format. In addition, we employ automatic metrics for measuring translation quality and use robust baselines, which we describe next. We focus on referenceless prompting in the main text.[9]

Evaluation. To assess the overall quality of the translation correction produced by LLMs, we rely on neural-based metrics, which obtain the strongest results in the WMT23 Metrics shared task (Freitag et al., 2023). Specifically, we report COMET (Rei et al., 2022a) as our primary metric, alongside BLEURT (Sellam et al., 2020), and COMETKIWI (Rei et al., 2022b).[10] Notably,

[9] In Appendix D.2 we carefully investigate the impact of providing reference translations to XTOWER.

[10] `wmt22-comet-da` and `wmt22-cometkiwi-da`.

Category	Example
Correct Span with Accurate Explanation Prevalence: 31.3% Relatedness: 5.3	**Source:** Indian govt withdraws appeal to hug cows on Valentine's Day after criticism **Reference:** Indische Regierung zieht Aufruf zur Umarmung von Kühen am Valentinstag nach Kritik zurück **Translation:** **Indian Govt** zieht Berufung zurück, Kühe am Valentinstag nach Kritik zu umarmen **Explanation:** The term "Indian Govt" is a direct transliteration of "Indian govt" from English to German, which is grammatically incorrect in German. The correct form should use the German spelling for "government" to match German grammatical rules, which would be "Indische Regierung."
Correct Span with Inaccurate Explanation Prevalence: 14.1% Relatedness: 1.6	**Source:** Is it so ephemeral then? Merely a "catchphrase" with no clear originator? **Reference:** Ist es also so vergänglich? Lediglich ein „Sprichwort“ ohne eindeutigen Urheber? **Translation:** Ist er also so kurzlebig? Lediglich ein **"**Schlagwort" ohne klaren Urheber? **Explanation:** The quotation marks around "Schlagwort" are correctly placed, but the analysis might be pointing out a minor stylistic preference or a typographical inconsistency. In German, the use of quotation marks can vary, but the provided translation is not incorrect in this regard.
Incorrect Span with Valuable Explanation Prevalence: 27.8% Relatedness: 5.3	**Source:** To agonize over personal failure & guilt in a world where "success" is so often bestowed on people responsible for the most reprehensible acts... **Reference:** Sich in einer Welt, in der „Erfolg“ so oft Menschen zuteilwird, die für die schlimmsten Taten verantwortlich sind, mit persönlichem Versagen und Schuldgefühlen herumzuquälen ... **Translation:** Sich über persönliches Versagen und **Schuld zu quälen** in einer Welt, in der "Erfolg" so oft Menschen zuteilwird, die für die verwerflichsten Taten verantwortlich sind... **Explanation:** The phrase "Schuld zu quälen" is a direct translation of "agonize over guilt." The error marking might suggest a preference for a different phrasing or structure to convey the depth of emotional distress. However, "Schuld zu quälen" is a correct and idiomatic way to express the idea of being tormented by guilt in German.
Incorrect Span with Worthless Explanation Prevalence: 26.83% Relatedness: 1.3	**Source:** Balenciaga boss calls holiday campaign 'stupid mistake' **Reference:** Balenciaga-Chef nennt Feiertagskampagne ‚dummen Fehler' **Translation:** Balenciaga **-Chef** nennt Weihnachtskampagne „dummer Fehler“ **Explanation:** The use of a hyphen between "Balenciaga" and "Chef" is not incorrect but could be seen as stylistically awkward in German. The hyphen here might make the name of the position seem less formal or less integrated into the sentence.

Table 4: Examples of explanations according to our qualitative analysis. Error spans are highlighted in red.

while the first metrics rely on a reference translation, CometKiwi yields an overall score without a reference. We cover lexical metrics in Appendix D.

Baselines. We adopt three models as baselines: TowerInstruct 13B (Alves et al., 2024)—Tower's translation-oriented LLM—, Mixtral 8x7B, and GPT 3.5 Turbo.[11] We use them in a **translation-only mode**: we pass a source sentence and ask for a translation in a 0-shot manner. We also prompt Mixtral 8x7B and GPT 3.5T for a translation correction given error spans and explanations. Since they are not trained to receive this information, we provide a 1-shot example for in-context learning.[12] Appendix B has more prompt details.

5.2 Results

We show results in Table 5 for each language pair.

Is xTower effective at refining translations? We observe that xTower's corrections improve the translation quality of the original translations for all language pairs. Interestingly, xTower obtains similar results with human-annotated and xComet error spans. For the latter, Comet deltas vary from 1 to 3 points, **leading to significant quality improvements for EN-DE and ZH-EN.**[13]

How does xTower compare to prompting LLMs? Comparing the best scores obtained by xTower—either from xComet or human spans—and TowerInstruct, we find that xTower outperforms TowerInstruct on HE-EN and ZH-EN, with a delta of 9 Comet points on the former.[14] Interestingly, however, xTower has a gap of only 0.2 to the original MT for HE-EN, suggesting that xTower is only slightly editing the original translation. Mixtral presents the lowest scores overall, while GPT 3.5T achieves the highest scores overall, outperforming xTower on all language pairs in terms of BLEURT and Comet. However, in contrast to xTower, we find that GPT 3.5T displays a consistent drop of performance when refining translations, suggesting that it may not utilize error spans and explanations as effectively.

Are the error spans being fixed? To assess how effectively the models address the errors highlighted in the prompt, we computed the percentage of fixed error spans with a string matching approach. Overall, xTower fixes 80% of the errors for EN-DE, 83% for HE-EN, and 84% for ZH-EN, while GPT 3.5T fixes 75% of the errors for EN-DE, 82% for HE-EN, and 80% for ZH-EN. These results indicate that both GPT-3.5T and xTower can, to some degree, leverage error spans and explanations to fix a large portion of the errors, with xTower showing a consistent edge over GPT-3.5T.

[11] We move from GPT 4 to 3.5T due to financial constraints.

[12] We experiment with 5-shot in Appendix D, but the results are on par with 1-shot, while also being more costly.

[13] As per (Kocmi et al., 2024), Comet deltas of ~1.0 denote improvements with a 90% accuracy with human judgments.

[14] Tower models were not trained to support Hebrew.

	EN-DE			HE-EN			ZH-EN		
MODEL	BLEURT	COMET	CKIWI	BLEURT	COMET	CKIWI	BLEURT	COMET	CKIWI
Original MT	48.4	78.4	75.5	59.8	77.5	75.5	55.2	78.0	76.7
Translation-only LLMs:									
Mixtral 8x7B	46.4 ↓ 2.0	80.4 ↑ 2.0	76.6 ↑ 1.1	53.9 ↓ 5.9	71.6 ↓ 5.9	69.3 ↓ 6.2	53.5 ↓ 1.7	77.7 ↓ 0.3	77.3 ↑ 0.6
GPT 3.5T	51.3 ↑ 2.9	82.7 ↑ 4.3	78.6 ↑ 3.1	65.5 ↑ 5.8	80.9 ↑ 3.4	77.8 ↑ 2.2	57.1 ↑ 1.8	79.9 ↑ 2.0	79.2 ↑ 2.5
TOWERINST 13B	50.0 ↑ 1.6	82.2 ↑ 3.8	78.7 ↑ 3.2	50.7 ↓ 9.1	68.7 ↓ 8.8	66.5 ↓ 9.0	56.5 ↑ 1.3	79.1 ↑ 1.1	78.4 ↑ 1.7
With predicted error spans:									
Mixtral 8x7B	42.9 ↓ 5.5	64.9 ↓ 13.5	58.7 ↓ 16.8	58.1 ↓ 1.6	76.4 ↓ 1.0	73.2 ↓ 2.3	51.2 ↓ 4.1	74.4 ↓ 3.6	73.4 ↓ 3.3
GPT 3.5T	53.4 ↑ 5.0	81.6 ↑ 3.2	77.5 ↑ 2.1	63.9 ↑ 4.1	80.9 ↑ 3.5	77.9 ↑ 2.4	56.2 ↑ 1.0	79.1 ↑ 1.1	77.9 ↑ 1.1
XTOWER 13B	52.7 ↑ 4.3	81.3 ↑ 2.9	77.0 ↑ 1.5	60.9 ↑ 1.1	78.5 ↑ 1.0	75.6 ↑ 0.1	56.0 ↑ 0.7	79.0 ↑ 1.0	78.4 ↑ 1.7
+ Hybrid	52.4 ↑ 4.0	82.2 ↑ 3.8	80.1 ↑ 4.6	62.4 ↑ 2.6	80.0 ↑ 2.5	78.7 ↑ 3.2	55.4 ↑ 0.2	79.1 ↑ 1.1	78.8 ↑ 2.1
With human-annotated error spans:									
Mixtral 8x7B	42.1 ↓ 6.2	66.8 ↓ 11.7	61.3 ↓ 14.2	57.7 ↓ 2.0	76.0 ↓ 1.5	73.1 ↓ 2.4	52.8 ↓ 2.5	75.7 ↓ 2.2	74.1 ↓ 2.7
GPT 3.5T	50.2 ↑ 1.8	80.6 ↑ 2.2	76.5 ↑ 1.0	62.6 ↑ 2.8	80.0 ↑ 2.5	77.4 ↑ 1.9	56.5 ↑ 1.3	79.2 ↑ 1.2	77.9 ↑ 1.2
XTOWER 13B	50.2 ↑ 1.8	81.3 ↑ 2.9	77.3 ↑ 1.8	60.0 ↑ 0.2	77.7 ↑ 0.2	75.0 ↓ 0.5	56.4 ↑ 1.2	79.4 ↑ 1.4	78.6 ↑ 1.9
+ Hybrid	52.7 ↑ 4.3	82.5 ↑ 4.1	79.9 ↑ 4.4	63.6 ↑ 3.8	80.8 ↑ 3.4	79.4 ↑ 3.9	56.2 ↑ 1.0	79.7 ↑ 1.7	79.2 ↑ 2.5

Table 5: Results for correcting translations with XCOMET-predicted or human-annotated error spans. We also show absolute differences from the original translation, where red and blue denote negative and positive deltas.

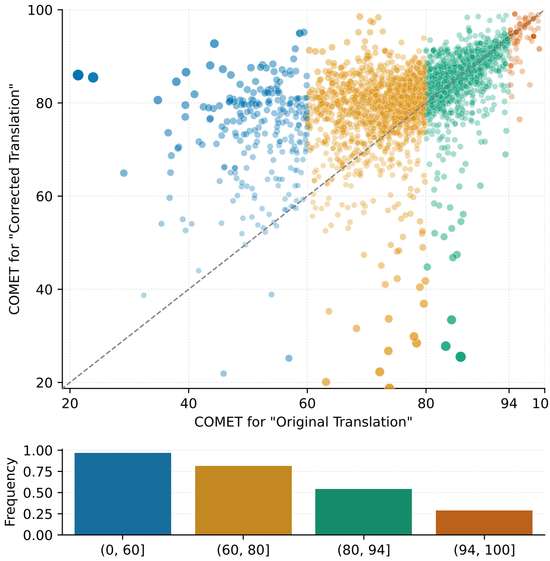


Figure 3: At the top, we show the quality of the original translation versus the corrected translation on EN-DE with human spans. At the bottom, we show how often the latter is higher than the former per quality bin.

Can we design an effective hybrid approach? We have established above that XTOWER's corrections can lead to significant improvements in translation quality. Here, we analyse where it is most effective in regards to the translation quality of the original translation. The scatter plot in Figure 3 illustrates the relationship between COMET scores for original and the corrected translations on EN-DE samples. It shows that XTOWER is most effective for low-quality original translations (COMET score $\leq$ 80), while for high-quality translations (COMET score $>$ 80) retaining the original translation may be better.[15] This is because the test dataset (WMT23) includes translations from diverse MT systems, including strong models like GPT-4 and (private) commercial systems (Freitag et al., 2023). Given these findings, we propose a **hybrid approach** that selects the best method based on the original translation's COMET score. Instead of a fixed threshold, we find the optimal threshold τ on 10% of the samples and use the following rule to obtain the final translation y:

$$y = \begin{cases} y_{\text{original}} & \text{if } m(y_{\text{original}}) > \tau \\ y_{\text{correction}} & \text{elif } m(y_{\text{correction}}) > m(y_{\text{original}}) \\ y_{\text{original}} & \text{otherwise,} \end{cases} \tag{1}$$

where m is a metric. We use COMETKIWI, a *referenceless* metric, as m. Results in Table 5 (under "Hybrid") show that this approach consistently improves translation quality across all language pairs, with boosts as high as 2 COMET points for HE-EN. These results suggest that a hybrid approach can significantly improve translation performance, especially for the more realistic scenario of using XCOMET spans, while also reducing inference costs by only querying XTOWER sporadically.[16]

How does explanation quality affect corrections? In Figure 4 we show that the largest quality gains are typically associated with explanations that have a high relatedness score (§4.2).

[15] This is consistent for all language pairs (*cf.* Figure 7).

[16] Portion of original translations kept: {46%, 41%} for EN-DE, {49%, 48%} for HE-EN, and {30%, 32%} for ZH-EN using XCOMET and human spans, respectively.

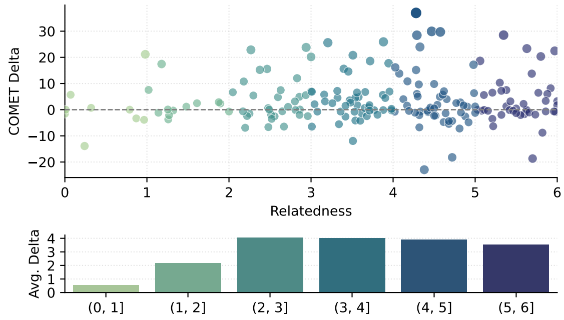


Figure 4: Delta between COMET scores for corrected and original translations according to how related explanations are to error spans.

Furthermore, we find a negative Pearson correlation ($r = -0.15$) between explanations' relatedness and original translations' COMET scores, highlighting that higher quality explanations are often associated with poorer quality original translations. This suggests that **high-quality explanations lead to significant improvements primarily for contexts where the initial translation quality is poor**, as hinted by Figure 3.

6 Related Work

Here, we discuss key related works in the domains of free-text explanations, automatic post-editing, span-level error detection, and the use of LLMs for translation and error explanation.

Free-text Explanations. Recent work has explored methods for generating free-text explanations either by utilizing human-written examples (Marasovic et al., 2022; Wiegreffe et al., 2022) or by prompting LLMs (Wei et al., 2022; Jung et al., 2022; Atanasova et al., 2023; Joshi et al., 2023). However, these explanations are typically produced to understand a model's decision rather than being constrained to justify marked spans in the input. In a similar vein, Feldhus et al. (2023) propose leveraging dense saliency maps to improve the verbalization of explanations by LLMs. In contrast, xTOWER focuses on producing explanations that are tied to specific error spans (§4.2) and helpful to humans (§4.3), within the context of MT.

Span-level Error Detection and Correction. In the context of span-level error detection, AUTOMQM (Fernandes et al., 2023), InstructScore (Xu et al., 2023), and xCOMET (Guerreiro et al., 2023a) have demonstrated the effectiveness of using neural models to identify errors in machine translations. For correcting errors in translations, a task more generally known as automatic post-editing (APE; Simard et al. 2007; Bhattacharyya et al. 2023), recent works prompt LLMs to produce suggestions for a new translation, such as TOWERAPE (Alves et al., 2024) and prompting GPT-4 (Raunak et al., 2023). We experiment with error spans annotated by humans or predicted by xCOMET for correcting translations in §5. Finally, incorporating error feedback into post-editing prompts has been concurrently explored by Ki and Carpuat (2024). While effective, their approach does not consider the addition of detailed explanations, which, as shown in §5, can further improve the translation correction process.

LLMs for Translating and Explaining Translation Errors. LLMs have been increasingly employed for translation tasks. TOWER (Alves et al., 2024) and ALMA (Xu et al., 2024) are notable examples of models designed specifically for translation-related tasks. InstructScore, a recent work by Xu et al. (2023), uses LLMs to provide explanations for translation errors. However, in contrast to xTOWER, InstructScore relies on reference translations, sidesteps the information in the source sentence, and produces explanations only as a by-product to improve quality score predictions. Additionally, while InstructScore focuses on producing a single quality score to reflect overall translation quality, xTOWER not only provides plausible and helpful explanations for humans, but also generates translation corrections.

7 Conclusions

In this paper, we introduced xTOWER, a multilingual LLM designed to provide free-text explanations for translation errors and generate corrected translations. By leveraging the strengths of TOWER and integrating specialized error detection from xCOMET, xTOWER can improve the interpretability of machine translation outputs in an automatic process. Our evaluations demonstrate that xTOWER not only produces high-quality and helpful explanations, as assessed by human evaluation, but can also significantly improves translation quality, especially when combined with accurate error spans. Furthermore, we propose a hybrid approach that dynamically selects between using the original translation or querying xTOWER for a correction, resulting in overall improvements in translation quality for all language pairs.

Limitations

While xTOWER significantly advances machine translation interpretability, it has various limitations. Even though the model's dependence on external error span detector tools like xCOMET brings modularity and flexibility, it also introduces pipeline complexity. Our evaluation, focused on the few language pairs which have MQM annotations available, may not generalize across all languages and domains. Additionally, the computational resources required for distillation and finetuning are substantial, limiting reproducibility for some users. The generated explanations, though helpful, may not always faithfully represent the model's reasoning or effectively guide users. Lastly, potential biases in the training data could affect translation and explanation quality, requiring further work to ensure fairness and reliability.

Potential Risks

The use of xTOWER may carry potential risks. One concern is the possibility of the model generating fluent but misleading explanations, which could affect user trust. There are also fairness considerations; as discussed above, the model might inadvertently reinforce biases present in the training data, potentially disadvantaging historically marginalized groups. Lastly, the focus on certain languages and datasets could lead to the underrepresentation of less commonly spoken languages. Careful monitoring and ongoing evaluation, such as detecting and overcoming hallucinations (Guerreiro et al., 2023b; Dale et al., 2023), can help mitigate these risks and ensure the model's responsible use.

References

Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, et al. 2024. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. 2023. Faithfulness tests for natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 283–294, Toronto, Canada. Association for Computational Linguistics.

Pushpak Bhattacharyya, Rajen Chatterjee, Markus Freitag, Diptesh Kanojia, Matteo Negri, and Marco Turchi. 2023. Findings of the WMT 2023 shared task on automatic post-editing. In *Proceedings of the Eighth Conference on Machine Translation*, pages 672–681, Singapore. Association for Computational Linguistics.

Yanda Chen, Ruiqi Zhong, Narutatsu Ri, Chen Zhao, He He, Jacob Steinhardt, Zhou Yu, and Kathleen McKeown. 2023. Do models explain themselves? counterfactual simulatability of natural language explanations. *arXiv preprint arXiv:2307.08678*.

David Dale, Elena Voita, Loic Barrault, and Marta R. Costa-jussà. 2023. Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity Even better. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–50, Toronto, Canada. Association for Computational Linguistics.

Ana C Farinha, M. Amin Farajian, Marianna Buchicchio, Patrick Fernandes, José G. C. de Souza, Helena Moniz, and André F. T. Martins. 2022. Findings of the WMT 2022 shared task on chat translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 724–743, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Christian Federmann. 2018. Appraise evaluation framework for machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 86–88, Santa Fe, New Mexico. Association for Computational Linguistics.

Nils Feldhus, Leonhard Hennig, Maximilian Dustin Nasert, Christopher Ebert, Robert Schwarzenberg, and Sebastian Möller. 2023. Saliency map verbalization: Comparing feature importance representations from model-free and instruction-based methods. In *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE)*, pages 30–46, Toronto, Canada. Association for Computational Linguistics.

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

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Nuno M Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André FT Martins. 2023a. xcomet: Transparent machine translation evaluation through fine-grained error detection. *arXiv preprint arXiv:2310.10482*.

Nuno M. Guerreiro, Elena Voita, and André Martins. 2023b. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1059–1075, Dubrovnik, Croatia. Association for Computational Linguistics.

Brihi Joshi, Ziyi Liu, Sahana Ramnath, Aaron Chan, Zhewei Tong, Shaoliang Nie, Qifan Wang, Yejin Choi, and Xiang Ren. 2023. Are machine rationales (not) useful to humans? measuring and improving human utility of free-text rationales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7103–7128, Toronto, Canada. Association for Computational Linguistics.

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. 2022. Maieutic prompting: Logically consistent reasoning with recursive explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1266–1279, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dayeon Ki and Marine Carpuat. 2024. Guiding large language models to post-edit machine translation with error annotations. *arXiv preprint arXiv:2404.07851*.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Tom Kocmi, 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.

Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. 2024. Navigating the metrics maze: Reconciling score magnitudes and accuracies. *arXiv preprint arXiv:2401.06760*.

Jenny Kunz, Martin Jirenius, Oskar Holmström, and Marco Kuhlmann. 2022. Human ratings do not reflect downstream utility: A study of free-text explanations for model predictions. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 164–177, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning. *Preprint*, arXiv:2301.13688.

Ana Marasovic, Iz Beltagy, Doug Downey, and Matthew Peters. 2022. Few-shot self-rationalization with natural language prompts. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 410–424, Seattle, United States. Association for Computational Linguistics.

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.

Ellie Pavlick and Joel Tetreault. 2016. An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics*, 4:61–74.

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.

Sahana Ramnath, Brihi Joshi, Skyler Hallinan, Ximing Lu, Liunian Harold Li, Aaron Chan, Jack Hessel, Yejin Choi, and Xiang Ren. 2023. Tailoring self-rationalizers with multi-reward distillation. *arXiv preprint arXiv:2311.02805*.

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 Linguis-*

tics: EMNLP 2023, pages 12009–12024, Singapore. Association for Computational Linguistics.

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

Ricardo Rei, Nuno M. Guerreiro, Marcos Treviso, Luisa Coheur, Alon Lavie, and André Martins. 2023. The inside story: Towards better understanding of machine translation neural evaluation metrics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1089–1105, Toronto, Canada. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. 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. 2022b. 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.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Michel Simard, Cyril Goutte, and Pierre Isabelle. 2007. Statistical phrase-based post-editing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 508–515.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment. *Preprint*, arXiv:2310.16944.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

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.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. Reframing human-AI collaboration for generating free-text explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 632–658, Seattle, United States. Association for Computational Linguistics.

Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. 2021. Measuring association between labels and free-text rationales. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10266–10284, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Wang, and Lei Li. 2023. INSTRUCTSCORE: Towards explainable text generation evaluation with automatic feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5967–5994, Singapore. Association for Computational Linguistics.

	EN-DE	HE-EN	ZH-EN	EN-RU
WMT 2022				
# Samples	8,815	-	13,631	10,996
# Error Spans	14,174	-	26,506	22,508
Avg. Input Length	42.2	-	72.3	40.5
Avg. Span Length	1.8	-	2.4	1.6
WMT 2023				
# Samples	4,111	5,325	15,690	-
# Error Spans	17,439	8,476	43,649	-
Avg. Input Length	190.0	18.1	52.9	-
Avg. Span Length	3.0	1.0	2.5	-
xCOMET *spans (without references):*				
# Error Spans	33,774	16,816	57,917	-
Avg. Span Length	2.4	1.1	2.0	-
xCOMET *spans (with references):*				
# Error Spans	30,856	16,434	53,602	-
Avg. Span Length	2.5	1.1	2.0	-

Table 6: Datasets statistics.

A Datasets Statistics

We show statistics for all datasets used in this work in Table 6.

B Prompting

Prompting explanations and translation correction. For 1-shot, we pass a unique example as input: for EN-DE we pass a single EN-DE example, whereas for HE-EN and ZH-EN we pass a ZH-EN example. For 5-shot, we pass a list of 5 examples containing 3 EN-DE, 1 EN-RU, and 1 ZH-EN samples. For all models, we sample new tokens using a temperature set to zero. We provide an example of our prompt template used for 1-shot EN-DE experiments in Table 1.

Prompting translation-only LLMs. For the translation LLMs baselines, we use the prompt shown in Table 7 to obtain translations.

Translate the following English source text to German:
English source: This is a great product and suitable for all bikes, cars and commercial applications.
German translation: Dieses großartige Produkt eignet sich für alle Motorräder, Autos und gewerbliche Anwendungen.

Table 7: 0-shot prompt for generating translations.

C Human Evaluation

Detailed Task Instructions. We present the detailed task instructions provided to the annotators in Figures 5 (*relatedness*) and 6 (*helpfulness*). The interface was created using Appraise (Federmann, 2018).

SPAN	LEVEL	EN-DE r	EN-DE ρ	ZH-EN r	ZH-EN ρ
Human	Explanation	0.56	0.51	0.34	0.20
Human	Document	0.50	0.38	0.21	0.17
xCOMET	Explanation	0.47	0.46	0.54	0.46
xCOMET	Document	0.39	0.40	0.50	0.46
Average	Explanation	0.52	0.48	0.44	0.33
Average	Document	0.45	0.39	0.35	0.32

Table 8: Inter-annotator agreement at explanation and document-level, according to Pearson's r and Spearman's ρ correlation coefficients.

Inter-annotator agreement. We ask human annotators to assess the translations at both the explanation and document levels. The inter-annotator agreement was measured using two statistical metrics: Pearson correlation coefficient (r) and Spearman rank correlation coefficient (ρ). Specifically, following Pavlick and Tetreault (2016), for each instance either at explanation or document-level, we randomly choose one annotator's scores to be the scores provided by Annotator 1, and take the mean scores of the other two annotators to be the scores given by an Annotator 2. We then compute the correlation for these two simulated annotators. Table 8 presents the results. The results indicate that while human annotators exhibit higher consistency for EN-DE translations, the agreement is generally lower for ZH-EN translations. For xCOMET spans, however, annotators agree more consistently across both language pairs.

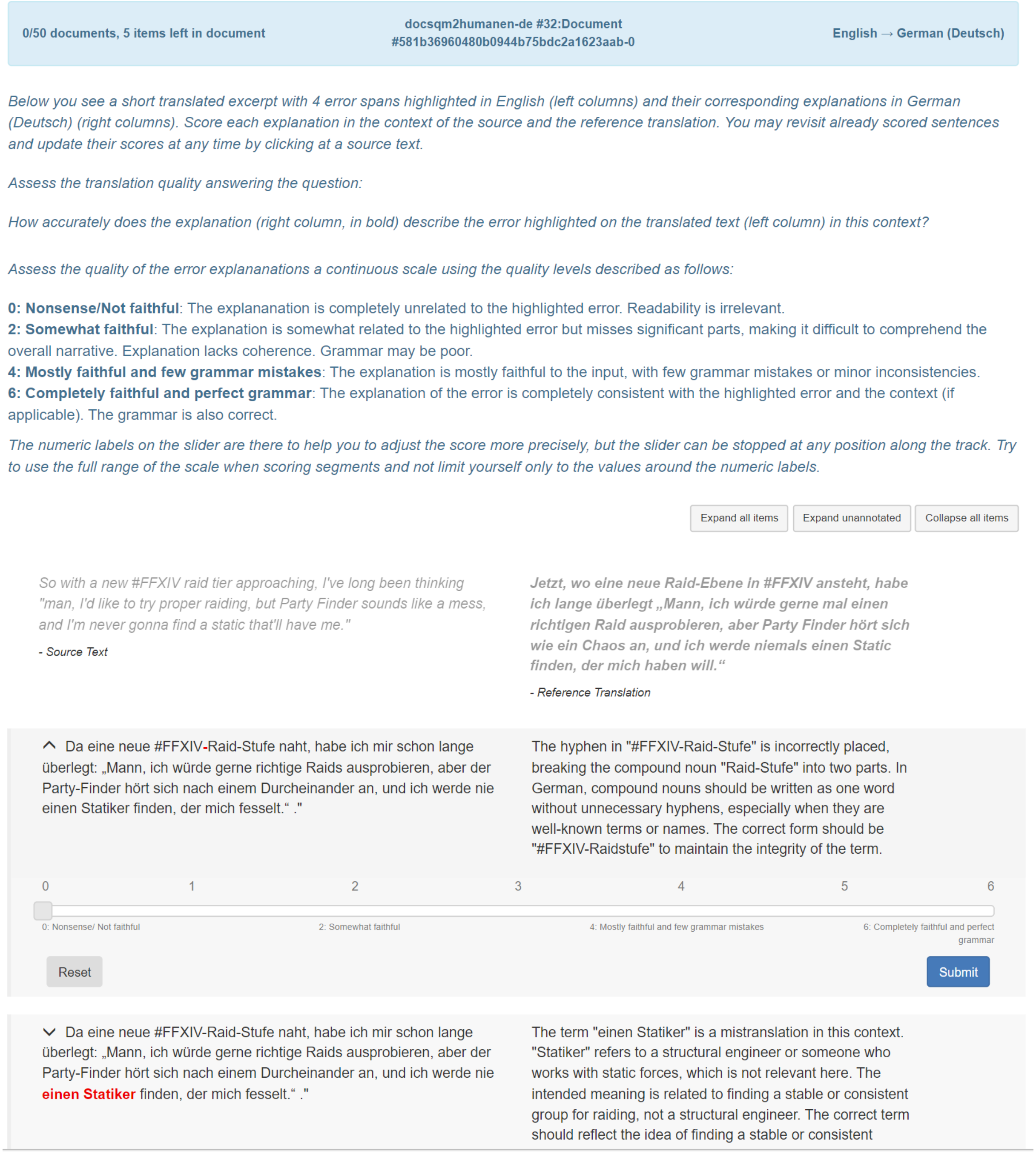
Sample size. For relatedness experiments, we evaluated a total of 282 explanations for EN-DE and 279 for ZH-EN (561 in total). For helpfulness, we evaluated 83 explanations EN-DE and 99 for ZH-EN (182 in total).

Participants Details. We hired native speakers of Chinese and German (fluent in English) for this task (four females and two males). They were compensated at $24 per hour.

D Translation Correction

D.1 Referenceless

K-shot prompt. In Table 9, we present results with k-shot for samples with human-annotated translation error spans in terms of COMET, covering both referenceless and reference-based se-



0/50 documents, 5 items left in document

docsqm2humanen-de #32:Document #581b36960480b0944b75bdc2a1623aab-0

English → German (Deutsch)

Below you see a short translated excerpt with 4 error spans highlighted in English (left columns) and their corresponding explanations in German (Deutsch) (right columns). Score each explanation in the context of the source and the reference translation. You may revisit already scored sentences and update their scores at any time by clicking at a source text.

Assess the translation quality answering the question:

How accurately does the explanation (right column, in bold) describe the error highlighted on the translated text (left column) in this context?

Assess the quality of the error explananations a continuous scale using the quality levels described as follows:

0: Nonsense/Not faithful: The explananation is completely unrelated to the highlighted error. Readability is irrelevant.
2: Somewhat faithful: The explanation is somewhat related to the highlighted error but misses significant parts, making it difficult to comprehend the overall narrative. Explanation lacks coherence. Grammar may be poor.
4: Mostly faithful and few grammar mistakes: The explanation is mostly faithful to the input, with few grammar mistakes or minor inconsistencies.
6: Completely faithful and perfect grammar: The explanation of the error is completely consistent with the highlighted error and the context (if applicable). The grammar is also correct.

The numeric labels on the slider are there to help you to adjust the score more precisely, but the slider can be stopped at any position along the track. Try to use the full range of the scale when scoring segments and not limit yourself only to the values around the numeric labels.

So with a new #FFXIV raid tier approaching, I've long been thinking "man, I'd like to try proper raiding, but Party Finder sounds like a mess, and I'm never gonna find a static that'll have me."

- Source Text

Jetzt, wo eine neue Raid-Ebene in #FFXIV ansteht, habe ich lange überlegt „Mann, ich würde gerne mal einen richtigen Raid ausprobieren, aber Party Finder hört sich wie ein Chaos an, und ich werde niemals einen Static finden, der mich haben will.“

- Reference Translation

Da eine neue #FFXIV-Raid-Stufe naht, habe ich mir schon lange überlegt: „Mann, ich würde gerne richtige Raids ausprobieren, aber der Party-Finder hört sich nach einem Durcheinander an, und ich werde nie einen Statiker finden, der mich fesselt.“ ."

The hyphen in "#FFXIV-Raid-Stufe" is incorrectly placed, breaking the compound noun "Raid-Stufe" into two parts. In German, compound nouns should be written as one word without unnecessary hyphens, especially when they are well-known terms or names. The correct form should be "#FFXIV-Raidstufe" to maintain the integrity of the term.

0
1
2
3
4
5
6
0: Nonsense/ Not faithful
2: Somewhat faithful
4: Mostly faithful and few grammar mistakes
6: Completely faithful and perfect grammar
Reset
Submit

Da eine neue #FFXIV-Raid-Stufe naht, habe ich mir schon lange überlegt: „Mann, ich würde gerne richtige Raids ausprobieren, aber der Party-Finder hört sich nach einem Durcheinander an, und ich werde nie **einen Statiker** finden, der mich fesselt.“ ."

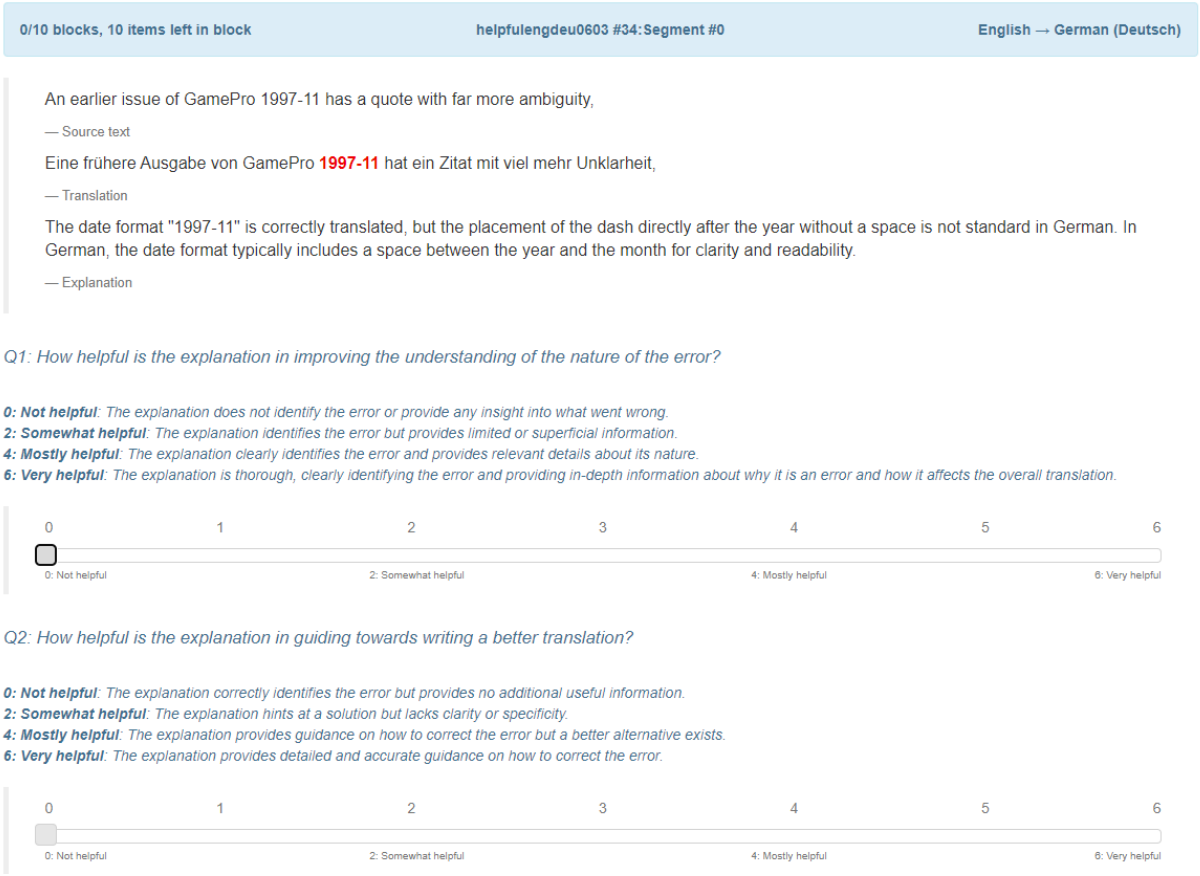
The term "einen Statiker" is a mistranslation in this context. "Statiker" refers to a structural engineer or someone who works with static forces, which is not relevant here. The intended meaning is related to finding a stable or consistent group for raiding, not a structural engineer. The correct term should reflect the idea of finding a stable or consistent

Figure 5: Screenshot of the *relatedness* task interface presented to annotators.

tups. Note that only xTower was evaluated with $k = 0$, as it was finetuned on explanations, and thus it can sidestep the in-context learning examples. Mixtral 8x7B seems to benefit more with $k = 5$ than other models for HE-EN and ZH-EN, but looses more around 4 COMET points for EN-DE. On the other hand, GPT 3.5 performs better with $k = 1$ than with $k = 5$ for referenceless experiments, with $k = 1$ results also being very close to $k = 5$ for reference-based experiments. Finally, xTower with $k = 5$ usually obtains slightly better results than with $k \in \{0, 1\}$ (delta within 0.2-0.4), but it introduces substantial runtime and memory costs as the prompt grows $\sim$5 times its original size. These findings motivated us to select $k = 1$ for Mixtral and GPT, and $k = 0$ for xTower, for all experiments in the paper.

D.2 Reference-based

For many use cases, users can provide an initial translation draft and then query XTOWER with the goal of obtaining an improved version. Here, we in-



0/10 blocks, 10 items left in block   helpfulengdeu0603 #34:Segment #0   English → German (Deutsch)

An earlier issue of GamePro 1997-11 has a quote with far more ambiguity,

— Source text

Eine frühere Ausgabe von GamePro **1997-11** hat ein Zitat mit viel mehr Unklarheit,

— Translation

The date format "1997-11" is correctly translated, but the placement of the dash directly after the year without a space is not standard in German. In German, the date format typically includes a space between the year and the month for clarity and readability.

— Explanation

Q1: How helpful is the explanation in improving the understanding of the nature of the error?

***0: Not helpful**: The explanation does not identify the error or provide any insight into what went wrong.*
***2: Somewhat helpful**: The explanation identifies the error but provides limited or superficial information.*
***4: Mostly helpful**: The explanation clearly identifies the error and provides relevant details about its nature.*
***6: Very helpful**: The explanation is thorough, clearly identifying the error and providing in-depth information about why it is an error and how it affects the overall translation.*

Q2: How helpful is the explanation in guiding towards writing a better translation?

***0: Not helpful**: The explanation correctly identifies the error but provides no additional useful information.*
***2: Somewhat helpful**: The explanation hints at a solution but lacks clarity or specificity.*
***4: Mostly helpful**: The explanation provides guidance on how to correct the error but a better alternative exists.*
***6: Very helpful**: The explanation provides detailed and accurate guidance on how to correct the error.*

Figure 6: Screenshot of the *helpfulness* task interface presented to annotators.

		Referenceless			Reference-based		
MODEL	k	en-de	he-en	zh-en	en-de	he-en	zh-en
Mixtral	1	66.8	76.0	75.7	69.9	84.5	80.0
Mixtral	5	63.1	77.1	76.6	65.4	85.3	81.4
GPT 3.5T	1	80.6	80.0	79.2	83.7	87.6	82.9
GPT 3.5T	5	79.9	79.9	79.0	81.4	87.9	83.2
xTower	0	81.3	77.7	79.4	84.1	88.2	83.6
xTower	1	81.4	77.6	79.3	84.3	88.7	83.9
xTower	5	81.2	77.9	79.2	84.4	88.6	84.0

Table 9: Results for translation refinement with k-shot prompting in terms of COMET.

vestigate the impact of providing a reference translation to the input on the quality of corrected translations.

D.2.1 Experimental Setup

Distillation data. Since references might play an important role in understanding and explaining translation errors, for example by offering context and highlighting specific areas where the translation deviates from the ideal, we include the reference in our prompts in 50% of the cases during distillation. Consequently, after finetuning, this approach allows us to balance between leveraging references for better explanations and ensuring the model engages in genuine error correction.

Prompting with XCOMET spans. Since we introduce the reference translation as an additional signal to our prompt, we rerun XCOMET with source-translation-reference triplets as input, obtaining a total of 99,892 spans.

Hybrid strategy. We use the same hybrid approach use for reference-less experiments, as defined in Equation 1. However, here we use COMET as m, a reference-based metric.

D.2.2 Results

We present our results in Table 11. Next, we discuss out main findings.

What's the gap to referenceless? Comparing the results with and without references, we find that reference-based models consistently outperform referenceless ones across all metrics and language pairs. For example, we obtain COMET boosts of roughly 3 points for EN-DE, 11 for HE-EN, and 5 for ZH-EN. Moreover, we note that human-annotated

	EN-DE			HE-EN			ZH-EN		
Spans	C	S	Δ	C	S	Δ	C	S	Δ
Without references:									
xComet	.01	.42	.44	.01	.41	.39	.01	.17	.77
Human	.01	.43	.49	.00	.40	.39	.01	.18	.77
With references:									
xComet	.10	.57	.49	.18	.72	.51	.08	.33	.79
Human	.06	.57	.55	.15	.69	.53	.07	.33	.78

Table 10: Portion of samples where the corrected translation is same as the reference ($C \downarrow$), their normalized Levenshtein similarity ($S \downarrow$), and how often the former is judged better than the latter by CometKiwi ($\Delta \uparrow$).

spans yield again similar results with xComet spans across the board. These findings indicate that xTower effectively leverages references, leading to significant improvements for the task of correcting translations.

Is xTower simply copying the reference? Since we are now providing a reference translation to xTower, it is not clear whether the quality gap that we have measured is not just an effect of copying the provided reference. To address this question, we computed two additional metrics: the percentage of translation corrections that are identical to the reference, and their closeness using normalized Levenshtein similarity. The results, presented in Table 10, indicate that xTower does not simply copy the reference. While the translation corrections become more similar to the reference, this is beneficial as it shows the model relies on the reference to generate improved translations. Furthermore, to determine the quality of these improvements, we compared the CometKiwi scores of the corrected translations and the original references relative to the source. The results show that this percentage is generally above 50%, demonstrating that xTower effectively produces translations that are on par with or better than the original references.

Is the hybrid approach effective? Our hybrid approach, which dynamically alternates between utilizing high-quality original translations and high-quality corrections, yields significant improvements, particularly in terms of Comet and BLEURT scores, just as observed in referenceless experiments in §5. Overall, these findings highlight the full potential of xTower towards improving translation quality.

D.3 Additional Results

Lexical metrics. For completeness, we include lexical metrics for referenceless and reference-based experiments for the translation refinement task in Table 12. Specifically, we include BLEU and ChrF.[17]

TowerBase vs xTower. To verify whether xTower maintains the original TowerBase translation capabilities after extending it, we also report its performance as a translation-only LLM in Table 12. That is, we prompt xTower with the 0-shot template shown in Table 7. The table shows that xTower performs on par or slightly surpass TowerInstruct for all language pairs in terms of BLEURT, Comet, and CometKiwi. This suggests that xTower not only keeps the original translation capabilities of TowerBase, but also holds potential to improve them.

Comet scores for original vs corrected translations. In Figure 3 (in §5), we show how xTower behaves depending on the quality of the original translation for EN-DE samples. Now, in Figure 7 we show plots for HE-EN and ZH-EN. Overall, we observe that the same trend remains: xTower is particularly helpful for cases where the original translation obtains weak-moderate Comet scores (from 0 to 80%).

E Computational Details

All experiments involving xTower and Mixtral 8x7B were carried on Nvidia RTX A6000 GPUS with 48GB VRAM. For GPT 4 and GPT 3.5T, we used the official API from OpenAI. We used VLLM[18] for efficient generation.

F AI Assistants

We have used Github Copilot[19] during code development, and ChatGPT[20] during paper writing for grammar correction.

[17] SacreBLEU signature: `|1|mixed|no|13a|exp|`.

[18] `https://github.com/vllm-project/vllm`

[19] `https://github.com/features/copilot`

[20] `https://chat.openai.com/`

	EN-DE			HE-EN			ZH-EN		
MODEL	BLEURT	COMET	CKIWI	BLEURT	COMET	CKIWI	BLEURT	COMET	CKIWI
Original MT	48.4	78.4	75.5	59.8	77.5	75.5	55.2	78.0	76.7
Translation-only LLMs:									
Mixtral 8x7B	46.4 ↓2.0	80.4 ↑2.0	76.6 ↑1.1	53.9 ↓5.9	71.6 ↓5.9	69.3 ↓6.2	53.5 ↓1.7	77.7 ↓0.3	77.3 ↑0.6
GPT 3.5T	51.3 ↑2.9	82.7 ↑4.3	78.6 ↑3.1	65.5 ↑5.8	80.9 ↑3.4	77.8 ↑2.2	57.1 ↑1.8	79.9 ↑2.0	79.2 ↑2.5
TOWERINST 13B	50.0 ↑1.6	82.2 ↑3.8	78.7 ↑3.2	50.7 ↓9.1	68.7 ↓8.8	66.5 ↓9.0	56.5 ↑1.3	79.1 ↑1.1	78.4 ↑1.7
With predicted error spans:									
Mixtral 8x7B	49.4 ↑1.0	70.0 ↓8.4	62.8 ↓12.7	74.1 ↑14.4	85.6 ↑8.2	77.2 ↑1.7	62.4 ↑7.1	80.6 ↑2.6	75.2 ↓1.5
GPT 3.5T	63.3 ↑15.0	85.2 ↑6.8	78.6 ↑3.1	80.2 ↑20.5	88.8 ↑11.4	79.3 ↑3.8	66.5 ↑11.2	83.3 ↑5.3	77.7 ↑0.9
xTower 13B	62.9 ↑14.6	84.6 ↑6.2	77.7 ↑2.2	80.5 ↑20.8	89.0 ↑11.5	79.3 ↑3.8	66.8 ↑11.6	83.7 ↑5.7	78.2 ↑1.5
+ Hybrid	62.4 ↑14.0	85.8 ↑7.4	79.4 ↑3.9	80.2 ↑20.4	88.4 ↑10.9	79.5 ↑4.0	66.5 ↑11.2	83.8 ↑5.8	78.0 ↑1.3
With human-annotated error spans:									
Mixtral 8x7B	46.3 ↓2.1	69.9 ↓8.5	63.5 ↓12.0	72.1 ↑12.3	84.5 ↑7.1	76.5 ↑0.9	60.7 ↑5.4	80.0 ↑2.0	75.3 ↓1.4
GPT 3.5T	58.5 ↑10.2	83.7 ↑5.3	77.9 ↑2.5	77.7 ↑18.0	87.6 ↑10.2	78.7 ↑3.2	65.3 ↑10.0	82.9 ↑4.9	77.9 ↑1.2
xTower 13B	59.1 ↑10.7	84.1 ↑5.7	77.8 ↑2.4	78.8 ↑19.1	88.2 ↑10.8	78.7 ↑3.2	66.5 ↑11.2	83.6 ↑5.6	78.5 ↑1.8
+ Hybrid	61.7 ↑13.3	86.0 ↑7.6	79.7 ↑4.2	79.6 ↑19.8	88.6 ↑11.1	80.1 ↑4.5	67.1 ↑11.9	84.3 ↑6.3	78.5 ↑1.8

Table 11: Reference-based results for correcting translations conditioned on explanations and error spans predicted via XCOMET or obtained via human annotation. We also show the absolute difference to the original translation, with red and blue denoting negative and positive deltas, respectively.

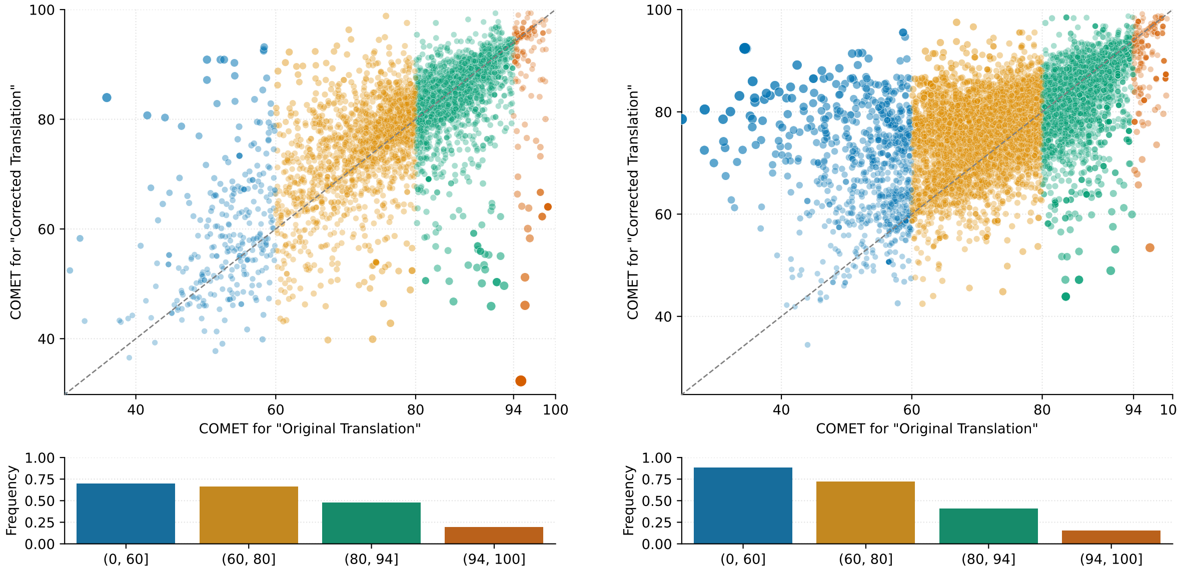


Figure 7: COMET of the original translation versus the corrected translation with human-annotated spans for HE-EN (left) and ZH-EN (right). At the bottom, we show how often the COMET for the corrected translation is higher than for the original per quality bin.

	EN-DE					HE-EN					ZH-EN				
Model	chrF	bleu	bleurt	comet	ckiwi	chrF	bleu	bleurt	comet	ckiwi	chrF	bleu	bleurt	comet	ckiwi
Original MT	64.8	39.0	48.4	78.4	75.5	56.5	33.5	59.8	77.5	75.5	49.6	23.8	55.2	78.0	76.7
Translation-only LLMs:															
Mixtral 8x7B	61.5	32.4	46.4	80.4	76.6	50.9	24.5	53.9	71.6	69.3	46.5	17.0	53.5	77.7	77.3
GPT 3.5T	68.2	41.9	51.3	82.7	78.6	64.5	43.9	65.5	80.9	77.8	50.2	22.0	57.1	79.9	79.2
TowerInst 13B	66.3	40.1	50.0	82.2	78.7	45.9	22.6	50.7	68.7	66.5	48.3	21.6	56.5	79.1	78.4
xTower 13B	66.5	40.0	50.5	82.2	78.6	45.8	22.2	50.6	69.4	67.3	48.7	21.6	56.8	79.5	78.5
Referenceless															
With predicted error spans:															
Mixtral 8x7B	29.9	10.4	42.9	64.9	58.7	53.5	31.6	58.1	76.4	73.2	41.5	18.1	51.2	74.4	73.4
GPT 3.5T	62.8	37.5	53.4	81.6	77.5	60.0	38.2	63.9	80.9	77.9	48.6	22.1	56.2	79.1	77.9
xTower 13B	59.5	34.1	52.7	81.3	77.0	57.1	34.5	60.9	78.5	75.6	48.5	20.8	56.0	79.0	78.4
+ Hybrid	64.8	38.4	52.4	82.2	80.1	59.9	37.4	62.4	80.0	78.7	51.4	24.1	55.4	79.1	78.8
With human-annotated error spans:															
Mixtral 8x7B	37.7	16.4	42.1	66.8	61.3	54.0	30.7	57.7	76.0	73.1	43.3	19.4	52.8	75.7	74.1
GPT 3.5T	63.1	37.6	50.2	80.6	76.5	58.6	36.1	62.6	80.0	77.4	48.8	22.3	56.6	79.2	77.9
xTower 13B	61.3	35.3	50.2	81.3	77.3	56.3	33.3	60.0	77.7	75.0	49.2	21.3	56.4	79.4	78.6
+ Hybrid	64.7	38.4	52.7	82.5	79.9	60.3	38.2	63.6	80.8	79.4	51.7	24.6	56.2	79.7	79.2
Reference-based															
With predicted error spans:															
Mixtral 8x7B	36.9	16.1	49.4	70.0	62.8	71.2	54.2	74.1	85.6	77.2	53.2	31.0	62.4	80.6	75.2
GPT 3.5T	74.1	55.4	63.3	85.2	78.6	79.1	65.3	80.2	88.8	79.3	58.8	36.6	66.5	83.3	77.7
xTower 13B	70.0	50.8	62.9	84.6	77.7	81.0	66.2	80.5	89.0	79.4	60.1	35.9	66.8	83.7	78.3
+ Hybrid	73.4	52.7	62.4	85.8	79.4	82.3	69.5	80.2	88.4	79.5	63.6	39.8	66.5	83.8	78.1
With human-annotated error spans:															
Mixtral 8x7B	40.3	18.7	46.3	69.9	63.5	69.0	50.4	72.1	84.5	76.4	51.6	29.1	60.7	80.0	75.3
GPT 3.5T	71.6	50.9	58.5	83.7	77.9	75.9	60.8	77.7	87.6	78.7	57.7	35.0	65.3	82.9	77.9
xTower 13B	70.6	50.7	59.1	84.1	77.9	78.9	62.5	78.8	88.2	78.7	59.8	35.6	66.5	83.6	78.5
+ Hybrid	73.5	52.5	61.7	86.0	79.7	80.6	67.1	79.6	88.6	80.1	63.3	39.7	67.1	84.3	78.5

Table 12: Full results for translation correction experiments in terms of lexical and neural metrics.