How Good Are LLMs for Literary Translation, Really? Literary Translation Evaluation with Humans and LLMs

Ran Zhang[1,4] **Wei Zhao**[2] **Steffen Eger**[3,4]

[1]University of Mannheim, School of Business Informatics and Mathematics
[2]University of Aberdeen, Department of Computing Science
[3]University of Technology Nuremberg (UTN), Dep. Computer Science & AI
[4] Natural Language Learning and Generation (NLLG) Lab

[1] ran.zhang@uni-mannheim.de [2] wei.zhao@abdn.ac.uk [3] steffen.eger@utn.de

Abstract

Recent research has focused on literary machine translation (MT) as a new challenge in MT. However, the evaluation of literary MT remains an open problem. We contribute to this ongoing discussion by introducing LITEVAL-CORPUS, a paragraph-level parallel corpus containing verified human translations and outputs from 9 MT systems, which totals over 2k translations and 13k evaluated sentences across four language pairs, costing 4.5k. This corpus enables us to (i) examine the *consistency* and *adequacy* of human evaluation schemes with various degrees of complexity, (ii) compare evaluations by students and professionals, assess the effectiveness of (iii) LLM-based metrics and (iv) LLMs themselves. Our findings indicate that the adequacy of human evaluation is controlled by two factors: the complexity of the evaluation scheme (more complex is less adequate) and the expertise of evaluators (higher expertise yields more adequate evaluations). For instance, MQM (Multidimensional Quality Metrics), a complex scheme and the de facto standard for non-literary human MT evaluation, is largely inadequate for literary translation evaluation: with student evaluators, nearly 60% of human translations are misjudged as indistinguishable or inferior to machine translations. In contrast, BWS (BEST-WORST SCALING), a much simpler scheme, identifies human translations at a rate of 80-100%. Automatic metrics fare dramatically worse, with rates of at most 20%. Our overall evaluation indicates that published human translations consistently outperform LLM translations, where even the most recent LLMs tend to produce considerably more literal and less diverse translations compared to humans.

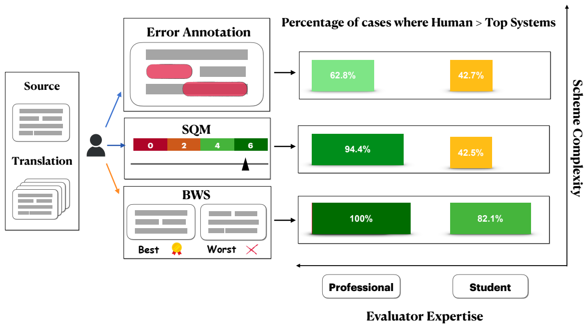


Figure 1: Adequacy comparison per scheme complexity and evaluator expertise. SQM represents Scalar Quality Metrics. Professional and student evaluators perform free and MQM error annotation, respectively. See Figure 4 and Section F.5 for more details.

1 Introduction

With the advent of Large Language Models (LLMs), literary translation, once the exclusive domain of human translators (Matusov, 2019; Hansen and Esperança-Rodier, 2022; Yan et al., 2024), now stands at the threshold of a technological transformation (Zhao et al., 2023b; Wu et al., 2024). LLMs have the potential to increase the volume of translated literary works, particularly for low-resource languages (Rutherford et al., 2024). However, they pose a dual threat: potentially decreasing the salaries of professional translators or even displacing them entirely and diminishing the aesthetic value of translations if LLMs' output quality is persistently overestimated. As a consequence, the field has received much increased attention recently, both in the academic domain (Vanroy et al., 2024; Daems et al., 2024; Resende and Hadley, 2024; Pang et al., 2024; Wang et al., 2023) and in the translation community (Nina George, 2023; Nizio, 2024; Martin, 2023; Kollektive-Intelligenz, 2024).

Evaluating literary translation poses unique challenges that extend beyond mere linguistic accuracy. It demands a delicate balance of cultural insight, authorial voice, creative interpretation, and aesthetics (Voigt and Jurafsky, 2012; Matusov, 2019; Macken, 2024). These nuances surpass the scope of traditional sentence-level MT

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 10961–10988
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

evaluation methods typically used in non-literary domains, presenting substantial obstacles for existing assessment techniques (Van Egdom et al., 2023).

We notice the following gaps in existing research regarding literary translation: (1) For **automatic evaluation**, previous metrics like d-BLEU, BLEURT and BLONDE (Liu et al., 2020; Sellam et al., 2020; Jiang et al., 2022) fall short in discrimination between professional human translations and machine outputs in literary MT (Thai et al., 2022). More powerful recent popular metrics such as XCOMET (Guerreiro et al., 2024), GEMBA-MQM (Kocmi and Federmann, 2023), and Prometheus (Kim et al., 2024) remain untested for literary texts. (2) The adequacy of **human evaluation** frameworks such as Multidimensional Quality Metrics (MQM) (Freitag et al., 2021; Lommel et al., 2014), Scalar Quality Metric (SQM) (Graham et al., 2013), and Best-Worst Scaling (BWS) (Kiritchenko and Mohammad, 2017) remains largely unexplored: to our knowledge, there is no comprehensive comparison of these schemes for evaluating literary translations. This critical gap is concerning, as current studies apply these approaches arbitrarily, resulting in findings that are not directly comparable and potentially leading to unreliable conclusions. (3) Existing **datasets** are small, cover few LLM models, and lack verification for human translations, including translator qualifications and experience (Yan et al., 2024). These issues can lead to substantial misjudgments when comparing LLM outputs to human translations, potentially overestimating the abilities of LLMs to the detriment of human professional translators.

In this paper, we address these gaps with the following key contributions:

(i) We introduce LITEVAL-CORPUS, a paragraph-level parallel corpus with verified high-quality human translations across four language pairs. This corpus also facilitates the comparison of 9 MT systems beyond the GPT series, including commercial models, popular LLMs, and previous state-of-the-art (SOTA) MT systems.[1]

(ii) We investigate multiple human evaluation schemes through both students and professional translators, illustrated in Figure 1, where we find that error annotation including MQM, the de facto measure in standard human MT evaluation, is (at least partially) inadequate for literary translation; SQM's effectiveness for comparing high-quality outputs highly depends on the expertise of evaluators; and BWS proves best for comparing high-quality outputs.

(iii) We examine the effectiveness of popular LLM-based metrics. GEMBA-MQM consistently outperforms other metrics. However, it primarily focuses on accuracy errors and struggles to differentiate between human translations and literal LLM outputs: it prefers human professional translations over LLM output at a level of 9.6%, compared to 94.4% for professional SQM, 81.1% for student BWS and 42% for student MQM and SQM.

(iv) We compare system performance, where we find that (1) human translators produce the best translations, as judged by both professionals and students; (2) GPT-4o ranks second, followed by Google Translate and DeepL. LLM outputs are more literal and less diverse than human translations with recent models decisively better and more closely approaching human literary translation.

2 Related work

Human evaluation for literary texts Literary MT evaluation primarily relies on human judgments of various types. MQM, the widely used guided error annotation scheme in the renowned MT venue WMT (Wang et al., 2023; Karpinska and Iyyer, 2023), has been shown effective for non-literary texts (Freitag et al., 2021) but its suitability for literary translation is uncertain. However, Kocmi et al. (2024) indicate that the complex MQM error categorization system requires costly trained experts and may lead to low annotation agreement. BWS is also employed for literary texts, including poetry (Thai et al., 2022; Belouadi and Eger, 2023), but has limitations in comparing multiple systems simultaneously (Kiritchenko and Mohammad, 2017). Recently, Wu et al. (2024) claim that LLM-based multi-agent literary translations achieve parity with human translations, based on preference comparison by crowdsourced and LLM evaluator—though the expertise of human translators remains unclear. SQM, a Lik-

[1] Our dataset and code are publicly accessible on GitHub: https://github.com/zhangr2021/LitMT_eval.

ert scale rating, is used in non-literary domains. To our knowledge, a comprehensive comparison of these schemes for literary MT is lacking, with different studies employing various methods ad libitum. Our work provides the first systematic comparison of these evaluation methods for literary texts.

Human evaluation of language models has been the subject of recent debate (Thomson et al., 2024; Belz et al., 2023). Human evaluation is typically assessed via *consistency*, i.e., whether human annotations are similar. Besides consistency, we focus on *adequacy*: whether the implications of evaluation are plausible (e.g., whether human reference texts are ranked the highest) (Toral et al., 2018).

Automatic metrics for MT have been extensively studied at sentence and document levels in non-literary domains. These include NLI-based MENLI (Chen and Eger, 2023), trained metrics like the COMET series (Guerreiro et al., 2024), BERT-based scores such as BLEURT (Sellam et al., 2020) and BERTScore (Zhang et al., 2020), prompting-based metrics like GEMBA-MQM (Kocmi and Federmann, 2023) and MQM-APE (Lu et al., 2025), in addition to document-level BLONDE (Jiang et al., 2022) and DiscoScore (Zhao et al., 2023a). While some researchers have applied these metrics to literary texts (Wang et al., 2023; Hansen and Esperança-Rodier, 2022), their effectiveness in literary domain remains understudied. Previous work has examined BLEU, BLEURT, and BLONDE for identifying top-performing systems (Van Egdom et al., 2023; Thai et al., 2022), with limited correlation analysis (e.g., brief discussions in appendices) of COMET, COMET-QE, BLEURT, and BERTScore (Karpinska and Iyyer, 2023). Our study addresses this gap by evaluating the effectiveness and adequacy of recent LLM-based metrics for literary translation.

Corpora for literary MT evaluation Existing parallel corpora are insufficient for literary translation evaluation. Table 4 (appendix) summarizes the statistics. *BWB* (Jiang et al., 2022) and *GuoFeng* (Xu et al., 2022) focus on recent *Zh-En* web novels with unclear translation origins, potentially post-edited MT outputs (Kolb, 2023). *PAR3* (Thai et al., 2022) lacks crucial meta-information, such as the origin of human translations, and offers limited evaluation with small-scale A/B tests comparing outputs from only three generation systems. Karpinska and Iyyer (2023) compare two systems (Google Translate and GPT-3.5) with only contemporary works—usually easier to translate than classics and potentially involving post-edited MT. Our LITEVAL-CORPUS addresses these limitations, containing over 2k paragraph-level segments with 13k sentences, comparing nine systems against verified published human translations with human evaluation under multiple schemes.

3 LITEVAL-CORPUS

We detail our LITEVAL-CORPUS in this section, covering data construction and statistics.

3.1 Dataset construction

Our data construction process consists of 3 steps: (1) preselect and construct paragraph-level parallel corpus; (2) select MT systems and collect translation hypotheses; (3) assess translation quality.

Constructing parallel corpus We consider three languages: English, Chinese, and German, covering translations between four different language pairs, i.e., English-Chinese (*En-Zh*), German-English (*De-En*)/English-German (*En-De*), and German-Chinese (*De-Zh*).[2] We begin by collecting paragraph-level parallel corpus for each language pair, including both classic and contemporary works. Classics present more challenges due to their complex syntax, language change, and cultural references that may no longer be familiar to modern audiences. Notably, older translations are especially valuable, as they were created before the widespread use of MT and are thus free from concerns of being post-edited MT outputs. Additionally, including contemporary works helps mitigate potential data contamination in LLMs, as these models are more likely to have been pre-trained on widely available texts, including classic literature.

To ensure the quality of our dataset, we derive both source and target paragraphs from verified publications, including books from Project Gutenberg (Stroube, 2003), reading samples from published works, and purchased materials. For classic works, the target paragraphs include at least two human translations, while contemporary works have at least one. For classic *De-En* works,

[2]The authors are native/proficient in these three languages and can thus verify the quality of the corpus and the evaluation.

we utilize PAR3 samples (Thai et al., 2022).[3] For contemporary *De-En* works and the other three language pairs (*De-Zh*, *En-Zh*, and *En-De*), we build the corpus from scratch, where the first author manually aligns source and translation paragraph pairs.

Selecting MT systems and collecting translation hypotheses We explore various-sized models, including cost-efficient alternatives to larger models:

- **Commercial systems** like Google Translate and DeepL. DeepL is particularly noteworthy for translators due to its widespread use within the translation community.
- Previous SOTA **transformer models** optimized for sentence-level translation, including NLLB-3.3b (Costa-jussà et al., 2022) and M2M_100-1.3b (Fan et al., 2021).
- Closed-source GPT-4o and four series of open-source **LLMs** of smaller sizes:[4] (1) Llama 3 (AI@Meta, 2024), one of the strongest open-source models; (2) Qwen 2 (Yang et al., 2024), specialized in Chinese; (3) Gemma 1.1 (Team et al., 2024), developed with the same technology as the closed-source Gemini model (Team et al., 2023); and (4) TowerInstruct (Alves et al., 2024), trained for translation-related tasks. We use identical prompts across all LLMs when requesting literary translations: *"Please translate the following literary texts from [source language] to [target language]. The texts are as follows: [texts]"*.[5]

3.2 Assessing translation quality

Our study examines three human evaluation schemes, i.e., guided/free error annotation, SQM, and BWS, using both student evaluators and professional translators. Figure 4 (appendix) summarizes the evaluation schemes. For **guided error annotation (MQM)**, evaluators read the source/candidate translation pair, identify error spans, and categorize them based on the annotation guideline with six major categories: (1) Terminology; (2) Accuracy; (3) Fluency; (4) Style; (5) Locale-convention; and (6) Non-translation.[6] Each error is then labeled as major or minor. The MQM score is thus computed with the following formula:

$$-\frac{(C_{\text{Non-translation}} \times 25 + C_{\text{major}} \times 5 + C_{\text{minor}} \times 1)}{\text{number of sentences per paragraph}}$$

C represents the count of errors. In contrast, **free error annotation** does not follow any guidelines. The evaluators highlight spans indicating erroneous/good translations and explain their highlights. We can thus compute span-level agreement with student error annotation.[7] For **SQM**, evaluators provide a scale rating between 0-6 for the overall quality of the translation, especially the translation's stylistic and artistic qualities based on the source. Evaluators may scroll up and down to check all other translations for scoring comparison. **BWS** is a direct comparison method that identifies the best and worst translations.

Evaluator statistics We involve four student evaluators, one for each language pair, and four professional translators (one for *En-Zh*, two for *En-De*, and one for *De-En*). All evaluators must be native speakers of the target language. Three students are master students with bachelor degrees in linguistics (*De-Zh*), computational linguistics (*De-En* and *En-De*) and one senior bachelor in translation studies (*En-Zh*). The student evaluators for *De-Zh*, *De-En*, and *En-Zh* are female and the *En-De* evaluator is male. For professional evaluators, we recruit two literary translators with verified publication records for ten hours each. The female translator for *En-Zh* has published a translation of a classic work with a public rating of 9.3/10. The female *En-De* translator is highly experienced, having published over 30 translated works. The other two translators volunteer to evaluate for six hours each. Both are female, recruited by universities, and highly experienced in academic translation.

We divide our human evaluation into three groups due to budget constraints:

[3]PAR3 includes 9k *De-En* paragraphs aligned with human translations but has alignment issues for some paragraphs with unidentified translation origins. To address these, the first author proofread sampled source-translation pairs and verified the origin of each human translation.

[4]For detailed parameter size, see Table 7 (appendix).

[5]We explored multiple prompts but observed no clear superiority of other prompts over the basic one. For more details, see Section C (appendix).

[6]See Section E.2 (appendix) for annotation details.

[7]Our feedback from translators revealed that MQM error annotation is unfamiliar to the translation community. Free error annotation better reflects the editing process and enables us to assess MQM protocol by comparing error span matches.

- Students conduct MQM and SQM evaluation for *all segments* using Label Studio platform (Tkachenko et al., 2020-2022).
- The same students perform BWS on *samples*, comparing translations from human translators, GPT-4o, DeepL, and Google Translate (Qwen for pairs involving Chinese) blindly.[8]
- Professional translators perform free error annotation and SQM on *samples* from human translators, GPT-4o, Google Translate, and DeepL (Qwen for pairs involving Chinese).[9]

3.3 LITEVAL-CORPUS statistics

Table 1 summarizes the statistics of LITEVAL-CORPUS. The dataset contains 188 source paragraphs, encompassing 2,188 translation segments at paragraph level including over 13k sentences. On average, source paragraphs contain 5.6 sentences, while target paragraphs have 6.1 sentences. To our knowledge, LITEVAL-CORPUS is the most extensive dataset for literary translation evaluation, offering comprehensive information on both source and target translations. It uniquely features paragraph-level translations across 10 generation systems.[10]

pair	#paragraph	#segments	#sentences	#sent/para	
				source	target
De-En	46	562	4310	7.1	7.7
En-De	46	554	2790	4.4	5.0
De-Zh	47	520	3039	5.3	5.8
En-Zh	49	552	3162	5.5	5.7
Total	188	2188	13301	5.6	6.1

Table 1: Characteristics of LITEVAL-CORPUS. #sentences and #sent/para represent the total number of annotated sentences in translations and the number of sentences per paragraph, respectively. #paragraph and #segments represent the number of source paragraphs and translation segments from all systems, respectively.

4 Human evaluation

We use LITEVAL-CORPUS to examine the consistency and adequacy of different schemes.

4.1 Consistency

Pair	MQM	SQM	Error span match		#segments	BWS
			span	label		
De-En	0.464	0.455	0.308	0.243	45	0.578
En-De	0.434	0.350	0.348	0.283	53	0.564
De-Zh	-	-	-	-	-	-
En-Zh	0.582	0.656	0.452	0.220	33	0.581
Mean	0.493	0.487	0.369	0.249	-	0.574

(a)

Pair	SQM		Error span match (span)		#segments
	S-S	S-P	S-S	S-P	
De-En	0.528	0.363	0.282	0.252	20
En-De	0.151	0.196	0.442	0.318	26
De-Zh	-	-	-	-	-
En-Zh	0.359	0.517	0.374	0.297	22
Mean	0.346	0.359	0.366	0.289	-

(b)

Table 2: Annotation agreement (a) between pairs of two student evaluators (**S-S**); (b) between student and professional translators (**S-P**). The agreement for MQM score and SQM score is measured by Kendalls Tau; the agreement for BWS and the span-level comparison is measured by Cohens kappa, where span match reflects the agreement on error spans regardless of labels, and label match on both spans and labels.

Student vs. student Table 2 (a) shows the annotation agreement between pairs of student evaluators across three language pairs: *De-En*, *En-De*, and *En-Zh*.[11] We assess MQM and SQM using Kendall's Tau (Puka, 2011). For BWS and annotated error spans, we employ Cohen's kappa (McHugh, 2012). A span match indicates agreement on error spans regardless of labels, and a label match considers agreement on both spans and labels.[12]

The results show *moderate to high agreement for both MQM and SQM*. MQM agreement ranges from 0.434 for *En-De* to 0.582 for *En-Zh*, while SQM scores the highest agreement for *En-Zh* at 0.656 and the lowest for *En-De* at 0.350. On average, MQM yields an agreement of 0.493, similar to SQM of 0.487.

In terms of error span agreement, *En-Zh* achieves the highest span match at 0.452, while *En-De* shows the highest label match at 0.283. Al-

[8]Evaluating all systems in one entry with 10-12 segments may decrease the reliability of BWS (Kiritchenko and Mohammad, 2017). We include 4-5 translation segments from the top-performing systems to achieve the best outcome.

[9]We focus on segments from top systems per feedback from professionals. Other translations often contain major issues, making them less valuable for professional assessment (e.g., highlighting several problematic sentences).

[10]See Table 5 (appendix) for information on source and translation details.

[11]*De-En* and *En-De* evaluators annotate shared segments for both language pairs. A co-author, a native Chinese speaker with a double major in linguistics and fluent in English, evaluates *En-Zh* samples. For *De-Zh*, we lack additional qualified evaluators.

[12]BWS compares 4-5 systems with samples of 15, 15, 23, and 22 paragraphs for *De-En*, *En-De*, *En-Zh*, and *De-Zh*, respectively. These correspond to 67, 65, 118, and 111 translation segments in total.

though error span agreement is generally *lower than MQM and SQM, the results remain competitive* with previous sentence-level multi-domain findings (Blain et al., 2023; Freitag et al., 2021; Leiter et al., 2023). However, a direct comparison of agreement on paragraph-level annotations in literary domain has yet to be explored, underscoring the value of this study as a baseline for future research in literary translation evaluation.

Additionally, *BWS results indicate strong performance*, with *En-Zh* achieving the highest agreement at 0.581, followed by *De-En* at 0.578 and *En-De* at 0.564 measured by Cohen's kappa. The overall mean BWS agreement is 0.574, showing a consistent agreement in selecting the best and worst translations.

Professional translators vs. student evaluators Table 2 (b) compares the annotation agreement between student and professional evaluators (S-P). For comparability, we also report the agreement between student evaluators (S-S) computed with the same segments. For SQM, S-P agreement is on average slightly higher than S-S agreement. For *En-Zh*, the agreement is 0.359 (S-S) versus 0.517 (S-P). One exception is *De-En* pair, where S-S agreement is 0.528, while S-P agreement drops to 0.363. The agreement for *En-De* is low for both groups. For error span match, the agreements between S-S and S-P both decline. Overall, S-S shows better agreement than S-P. This is understandable as student evaluators are trained with an annotation guideline, while professional translators perform free annotation without guidelines. The results indicate that both student and professional evaluators focus on joint error spans to some degree.[13] While there is *consistency between the two groups*, particularly in certain language pairs, *discrepancies exist*. We discuss these in detail in Section 4.2.

Consistency in system ranking The evaluation results in Table 7 (appendix) show strong consistency in system rankings between student and professional evaluators for both MQM and SQM. Human translation consistently ranks first in both scores when averaged across all segments in the four languages studied. This top ranking holds for both student and professional evaluators, indicating human translation's superiority. GPT-4o ranks second, followed by Google Translate and DeepL or Qwen. Smaller models such as M2M and NLLB consistently rank low. Additionally, according to student evaluators, there is a clear performance gap between the top 4 systems and the rest, i.e., 5.5 points drop from the fourth to fifth model in MQM and 1.1 points by SQM. Notably, professional translators rate human translations substantially higher than the second-place GPT-4o with a difference of 1.8 points in SQM, while this gap is comparatively smaller with 0.4 in student SQM.

4.2 Adequacy

We assess the adequacy of translation evaluation schemes by evaluating their ability to differentiate high-quality human translations from machine outputs. We believe that experienced human translators outperform MT systems, based on recent studies for other domain texts (Yan et al., 2024) and the inherent advantages of professional translators: access to full-text context, professional education, and the ability to research cultural backgrounds. These factors enable translators to develop a deeper understanding of the source material, intuitively resulting in more accurate and creative translations.

Table 3 shows the percentage of segments where human translations are preferred over other systems. The table presents two scenarios: (1) human translations compared to top-performing models like GPT-4o, DeepL, Google Translate, and Qwen, and (2) human translations compared to other MT systems, excluding top-performing models.

While student **MQM** demonstrates good performance in scenario (2), the bad performance in scenario (1) indicates that *MQM alone may not be adequate for evaluating translation quality effectively, especially for top systems*. The first reason is that MQM requires complex categorization and expects high expertise. This is evidenced by the inferior consistency score for label match in Table 2 (a) from student evaluators. The second reason is that MQM may be inherently unsuitable for literary translation evaluation. MQM struggles with evaluating top models. For example, Figure 5 (appendix) illustrates that the error distribution of top models closely resembles that of human translation. 16.6% of GPT-4o outputs are "error-free" but still fall short in wording or overall style compared to human translations. Professional translators also express concerns about

[13] See Section E.3 for detailed examples.

pair	Human > Top Systems (GPT-4o, DeepL, Google Translate, and Qwen)						Human > Other Systems (excluding Top systems)			
	Student			Professional	GEMBA-MQM		Student		GEMBA-MQM	
	MQM	SQM	BWS	SQM	Original	Literary	MQM	SQM	Original	Literary
De-En	60.0%	60.0%	66.7%	100.0%	6.7%	0.0%	86.7%	80.0%	26.7%	33.3%
En-De	35.0%	60.0%	86.7%	83.3%	0.0%	10.0%	95.0%	100.0%	70.0%	60.0%
De-Zh	30.0%	25.0%	95.0%	-	15.0%	20.0%	95.0%	90.0%	45.0%	50.0%
En-Zh	45.8%	25.0%	80.0%	100.0%	0.0%	8.3%	87.5%	79.2%	41.7%	50.0%
Mean	42.7%	42.5%	82.1%	94.4%	5.4%	9.6%	91.0%	87.3%	45.8%	48.3%

Table 3: Percentage of segments where human translations are preferred over machine translations per human evaluation schemes and GEMBA-MQM versions.

the "error-counting" schemes, particularly when assessing high-quality literary translations, as they sometimes make intentional choices, such as omitting certain elements or slightly altering meanings, to enhance the expression in the target language.[14]

On average, **SQM** with students performs similarly to MQM but professionals show a substantially higher preference for human translation (100% in *De-En* and *En-Zh*). Professional SQM outperforms student SQM by a large margin of nearly 40 points, because students—lacking sufficient competence or experience—may struggle to capture nuances through SQM rating as effectively as professionals do. This indicates that SQM's effectiveness heavily depends on the evaluator's experience and expertise. However, combining student MQM and SQM can empirically improve the adequacy by a maximum margin of 20 points for *De-En* and nearly 5 points for *En-De*, *En-Zh* and *De-Zh* compared to SQM or MQM alone (see Figure 6 in Section D.4 for more details).

BWS, on the other hand, consistently shows high ratios among top systems, with students preferring human translations at 86.7% for *En-De* and 95.0% for *De-Zh*. On average, BWS ranks second with 82.1% compared to ∼94% for professional SQM and ∼42% for student SQM and MQM. This suggests that BWS is more suitable for evaluating top systems, although it cannot provide the detailed error insights that MQM offers.

Summary: Student and professional evaluators consistently evaluate literary translations using guided/free error annotation and SQM at a moderate level with BWS showing slightly better agreement. Student MQM is not adequate to discriminate human translations from top LLM translations. Using it may yield false conclusions. The effectiveness of SQM heavily depends on the expertise of the evaluator. BWS excels at distinguishing human translations from top machine translations even among less-experienced evaluators, though this scheme offers less detailed feedback for further improvements such as post-editing.

5 Automatic metrics

We use LITEVAL-CORPUS to assess recent automatic metrics. While human MQM and SQM may not fully capture the nuances of literary translation quality, as discussed in the previous section, it is still valuable to examine how consistently current metrics align with human MQM/SQM evaluations. This is particularly relevant given that the recent top-performing metrics are built on similar principles. We focus on reference-free metrics due to the scarcity of reference translations for the majority of literary works (Rutherford et al., 2024) and measure the performance by segment-level correlation between automatic metrics and human evaluation.

We examine four SOTA automatic metrics that rely on LLMs: Prometheus 2 (7 billion parameters) (Kim et al., 2024), an open-source model trained to evaluate other language models, which is queried using a tailored prompt (shown in Table 13 in the appendix) for SQM evaluation of literary translations; XCOMET-XL/XCOMET-XXL (3.5 and 10 billion parameters) (Guerreiro et al., 2024), one of the strongest open-source MT metrics for standard MT, which is fine-tuned to assess quality, generating scores and error spans with severity labels;[15] GEMBA-MQM (Kocmi and Federmann, 2023), a top-performing prompting-based metric, which is designed to de-

[14] See examples in Table 9 (appendix).

[15] We apply XCOMET-XL and XCOMET-XXL directly to paragraph-level evaluation, based on Deutsch et al. (2023)'s finding that applying sentence-level metrics to paragraphs produces results similar to averaging sentence-level scores.

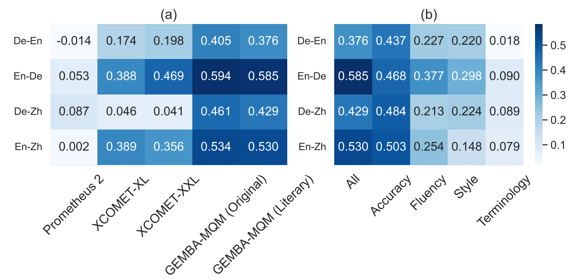


Figure 2: Segment level correlation measured by Kendalls τ between (a) human MQM and automatic metrics and (b) human MQM and GEMBA-MQM (Literary) per error categories. See Figure 9 (appendix) for SQM results.

tect translation quality errors based on the MQM framework using LLMs.[16] We modify GEMBA-MQM with the same error categories we use for student evaluators. We implement GEMBA-MQM (Original) using the original prompt template and GEMBA-MQM (Literary) adapted with domain knowledge and literary translation examples shown in Table 11.

The best metric (GEMBA-MQM) correlates moderately with human MQM but it cannot distinguish human translation from LLM outputs. Figure 2 (a) illustrates the segment-level correlation between human MQM and automatic metrics. GEMBA-MQM, particularly its Original version, consistently outperforms other metrics. GEMBA-MQM (Literary) performs similarly to GEMBA-MQM (Original), though slightly worse across the board. XCOMET-XL and XCOMET-XXL models rank second overall, demonstrating poor to moderate correlation with human evaluation. Their performance notably declines for more distant language pairs, such as *De-Zh*, possibly due to differences or limitations in their training data for this language direction. Moreover, we notice that XCOMET-XL and XCOMET-XXL are negatively correlated with the length of the target translation at -0.420 and -0.393, respectively. In contrast, the correlations between the length of the target and other metrics are all below 0.05 in absolute value. Additionally, Prometheus performs the worst, assigning a score of 4 (equivalent to "good" translation) to 54.9% of segments. Its correlation to human SQM ranks lowest among all automatic metrics, as shown in Figure 9 (appendix).

Even the best automatic metric is not adequate for literary MT evaluation. GEMBA-MQM, despite moderate correlation with human SQM and MQM, faces adequacy issues in literary translation evaluation. Table 3 shows the percentage of segments where GEMBA-MQM prefer human translations over MT. The results show GEMBA-MQM severely struggles to distinguish human translations from top LLM systems. For scenario (1), human translator vs. top systems, GEMBA-MQM (Literary)'s highest percentage (9.6%) is still 32.9 points lower than the worst human performance. This may partly stem from LLMs' tendency to act as narcissistic evaluators, favoring their own outputs—especially in reference-free scenarios (Liu et al., 2024). In scenario (2), human translator vs. systems excluding top ones, despite higher percentages, GEMBA-MQM still lags behind human evaluators by at least 39 points. GEMBA-MQM (Literary) slightly outperforms the original version by an average of 4.2 points and 2.5 points for scenarios (1) and (2), indicating that literary-specific knowledge helps but remains insufficient for discriminating human and LLM outputs. As shown in Table 14 (appendix), other metrics also exhibit substantial gaps compared to human performance. XCOMET-XL scores 26.9 points lower and XCOMET-XXL 21.2 points lower than the worst human performance in scenario (1). In scenario (2), the gaps are even larger with 39.6 points for XCOMET-XL and 30.2 for XCOMET-XXL. While XCOMET-XL and XCOMET-XXL yield better results than GEMBA-MQM on adequacy, with average performances of 15.6% and 21.3% in scenario (1) respectively, they correlate worse with human MQM and SQM than GEMBA-MQM. This is particularly evident for *De-En* and *De-Zh*, as shown in Figure 2 (a) and Figure 9 (appendix).

Additionally, our analysis reveals that GEMBA-MQM's performance is primarily driven by Accuracy aspects, with less reliability in evaluating Fluency, Style, and Terminology (see Figure 2 (b)).[17] Accuracy shows the strongest correlations across all language pairs. Accuracy for *De-En* and *De-Zh* even surpasses the full correlation. Meanwhile, Fluency and Style

[16] We utilize GPT-4o mini for GEMBA-MQM in this study due to cost reasons. See Section F.1 and F.2 (appendix) for more details.

[17] We aggregate human MQM and GEMBA-MQM per error category and then compute the correlation separately per each category.

demonstrate poor to fair correlations. Notably, Terminology consistently shows near-zero correlations, which could be problematic for distant language pairs with more culture-specific terms, e.g., *En-Zh* with Terminology as the second most frequent error category. These findings suggest that future metric development should focus on improving the assessment of Fluency, Style, and Terminology.

Human MQM vs. GEMBA-MQM: LLMs demonstrate a tendency towards more literal and less diverse translations. We further compare human MQM with GEMBA-MQM in system rankings considering characteristics of the translation systems. To assess the literalness of translation, we analyze the syntactic similarity between source-translation pairs. We utilize the recent *document-level* tree kernel-based syntactic similarity metric FastKASSIM (Chen et al., 2023). While high syntactic similarity preserves the source's form, it frequently sacrifices naturalness in the target language and hinders creativity in translations (Guerberof-Arenas and Toral, 2022; Al-Awawdeh, 2021). Diversity is another concern in literary domain (Chen et al., 2024; Zhang and Eger, 2024). To measure lexical diversity of one system compared to all other systems, we calculate pairwise lexical overlap between pairs of translations from different systems and average them per system.[18] A lower value suggests a more diverse vocabulary selection. Figure 3 shows results based on **all publications**, illustrating human MQM scores for 10 translation systems averaged over four language pairs (z-axis) against syntactic similarity (x-axis) and lexical overlap (y-axis) (see Figure 10 in the appendix for results on **contemporary works**). The rank of systems is marked as [rank per human MQM, rank per GEMBA-MQM]. We show detailed examples in Table 15 and Table 16 (appendix).

Three major clusters emerge: (1) Human Translator stands alone, characterized by a high MQM score with the lowest syntactic similarity (0.21) and lexical overlap (18.9). While human translation ranks first per human MQM by a 0.4 margin over GPT-4o, GEMBA-MQM (Literary) prefers GPT-4o with a larger margin of 1.4 over human translators. Surprisingly, GEMBA-MQM (Literary) even rates Google Translate higher than human translations. (2) GPT-4o, DeepL,

[18]See Section F.6 (appendix) for more details.

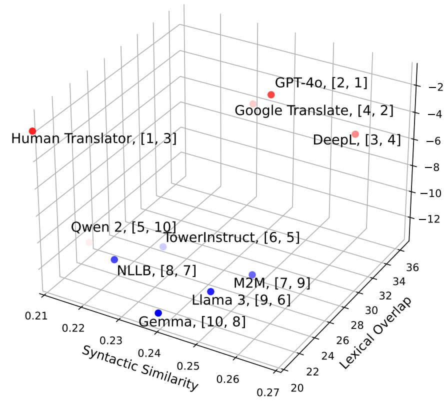


Figure 3: Distribution of human MQM (z-axis) against syntactic similarity between source-translation pairs and lexical overlap of the current system to all other systems. The system rankings are shown as [rank per human MQM, rank per GEMBA-MQM (Literary)]. See Figure 11 for a detailed heatmap.

and Google Translate all exhibit high MQM scores and high lexical overlap. The three models share high lexical overlap, contributing to their remarkably high average scores. DeepL has the highest syntactic similarity of 0.27. (3) The third cluster demonstrates a performance gap to other clusters, with Qwen 2 and TowerInstruct performing slightly better. Qwen 2 paraphrases more for *De-Zh* and *En-Zh* pairs and sometimes includes Chinese characters in *De-En* and *En-De* outputs, which results in low syntactic similarity and lexical overlap but still higher than human.

Summary: GEMBA-MQM outperforms other current automatic metrics but struggles with non-Accuracy aspects. All metrics lack adequacy in distinguishing human and LLM translations. LLMs tend to produce more literal and less diverse translations showing higher syntactic similarity to the source and greater lexical overlap with other systems than human translations. Further, LLM as evaluator prefers more literal translations and is biased towards their outputs.

6 Conclusion

This paper introduces LITEVAL-CORPUS and a comprehensive study on literary translation evaluation with human and automatic metrics.

Our study reveals the limitations of error annotation, particularly MQM when comparing top-performing systems. This may be attributed to three factors: (1) the complexity of MQM compared to other schemes; (2) the current MQM framework failing to account for "errors" in literary translation intended for certain equivalence in the target language; (3) current leading systems producing fewer errors than previous MT systems, rendering MQM insufficient. While using BWS or experienced professionals yields more plausible assessments, BWS has limitations for further improvements like post-editing when compared to MQM. These results have substantial implications, particularly for determining human parity—a crucial issue in today's LLM landscape. They are especially relevant for professional translators, as naive translation companies might eliminate human translators due to misjudgments based on MQM or SQM evaluations by inexperienced practitioners. While we highlight MQM's limitations in literary MT evaluation, our current study does not propose specific improvements to the existing MQM guidelines. Future work can explore these aspects in depth.

Furthermore, our results indicate that automatic metrics development should focus on aspects such as Style and Terminology. LLMs still need to address the issue of overly literal outputs, and a substantial gap remains between LLM and human quality in literary translation, despite the clear advancements of recent models.

Limitations

Several limitations of our current work warrant consideration:

Our study primarily focuses on prompting-based techniques with identical basic prompts across LLMs. While our initial exploration with GPT-4o showed no substantial improvement with complex prompts, we acknowledge that LLM-specific prompt optimization could potentially yield better results in literary translation. This aligns with the current understanding that prompting is a delicate process (Leiter and Eger, 2024; Mizrahi et al., 2024).

Other techniques such as fine-tuning with domain knowledge may further enhance LLM performance in this context. Conversely, data contamination may inadvertently improve translation quality for classic works already in the model's training data. However, classics unseen by the model or yet to be translated might exhibit lower translation quality.

Our focus on paragraph-level translation still does not fully capture the nuances of literary texts. The absence of evaluation on (consecutive) chapters limits our ability to comprehensively evaluate translation quality within a broader context. Although we conducted limited experiments with multi-paragraph samples, resource constraints and the complexity of the task on student evaluators prevented a more extensive exploration in this direction.

Our study covers four high-resource language pairs, each evaluated by a single evaluator except when computing agreements. We acknowledge that involving multiple evaluators with greater expertise could yield more robust results. To mitigate this limitation within our resource constraints, we implemented a rigorous selection process and provided comprehensive training for our evaluators. However, our dataset doesn't explore LLM performance in low-resource languages—an area that could potentially benefit the most from literary MT.

While we highlight the limitations of MQM in literary MT evaluation, we do not propose specific improvements to the current MQM guidelines. Professional translators agree that MQM is insufficient for literary translation but recognize that enhancing it requires extensive research and exploration. Future work should focus on improving upon MQM to better reflect aspects such as aesthetics, emotion, stylistic features, and the tone of both authors and characters.

Ethical Considerations

LLMs in literary translation have sparked widespread public interest and concern. This topic has become a focal point for researchers, translators, and publishers, generating intense debate at various forums including workshops, interviews, and book fairs.[19] The increased engagement underscores the need for responsible handling of MT outputs in general.

Low-quality translations can misrepresent or negatively affect an author's original work without

[19]We list some events below. Workshops: https://ctt2024.ccl.kuleuven.be/ and https://bohtranslations.com/blog/dual-t-end-of-project-event-recap. Interviews & discussions: https://www.goethe.de/ins/gb/en/kul/lue/ail.html, https://www.buchmesse.de/presse/pressemitteilungen/2023-10-06-kuenstliche-intelligenz, and https://www.stiftung-genshagen.de/veranstaltungen/veranstaltungen-2024/kuku/default-a39e201a92/.

their awareness. LLMs may introduce or exacerbate issues of bias and toxicity in translated literature, impacting the audience, particularly children. The growing capabilities of LLMs in translation also threaten professional translators, raising concerns about job displacement and the devaluation of human expertise due to misconceptions about LLM quality. However, the careful use of LLMs or MT models as an auxiliary tool could democratize access to global literary works by improving the accessibility of literature across languages and cultures, especially beneficial for low-resource languages. LLMs could also empower translators, allowing them to focus on the challenging aspects of translation while leaving routine tasks to MT. Additionally, LLMs could serve as potential brainstorming tools, e.g., inspiring more creative wordplay alternatives.

Regarding dataset usage: We adhere to the guidelines specified in Germany's Copyright in Academic Work regulations.[20] The dataset usage in this study has been approved by the university ethics committee.

Regarding human evaluation: We obtained consent from all participating evaluators to anonymously disclose their contributions.

Acknowledgements

We thank the anonymous reviewers for their thoughtful feedback that greatly improved the text. We also thank all evaluators for their dedication and appreciate the professional input from all translators involved. The NLLG Lab gratefully acknowledges support from the Federal Ministry of Education and Research (BMBF) via the research grant "Metrics4NLG" and the German Research Foundation (DFG) via the Heisenberg Grant EG 375/5-1.

References

AI@Meta. 2024. Llama 3 model card.

Nabil Al-Awawdeh. 2021. Translation between creativity and reproducing an equivalent original text. *Psychology and Education Journal*, 58(1):2559–2564.

Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and Andre Martins. 2024. Tower: An open multilingual large language model for translation-related tasks.

Jonas Belouadi and Steffen Eger. 2023. ByGPT5: End-to-end style-conditioned poetry generation with token-free language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7364–7381, Toronto, Canada. Association for Computational Linguistics.

Anya Belz, Craig Thomson, Ehud Reiter, Gavin Abercrombie, Jose M. Alonso-Moral, Mohammad Arvan, Anouck Braggaar, Mark Cieliebak, Elizabeth Clark, Kees van Deemter, Tanvi Dinkar, Ondřej Dušek, Steffen Eger, Qixiang Fang, Mingqi Gao, Albert Gatt, Dimitra Gkatzia, Javier González-Corbelle, Dirk Hovy, Manuela Hürlimann, Takumi Ito, John D. Kelleher, Filip Klubicka, Emiel Krahmer, Huiyuan Lai, Chris van der Lee, Yiru Li, Saad Mahamood, Margot Mieskes, Emiel van Miltenburg, Pablo Mosteiro, Malvina Nissim, Natalie Parde, Ondřej Plátek, Verena Rieser, Jie Ruan, Joel Tetreault, Antonio Toral, Xiaojun Wan, Leo Wanner, Lewis Watson, and Diyi Yang. 2023. Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in NLP. In *Proceedings of the Fourth Workshop on Insights from Negative Results in NLP*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Frederic Blain, Chrysoula Zerva, Ricardo Rei, Nuno M. Guerreiro, Diptesh Kanojia, José G. C. de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh Azadi, Constantin Orasan, and André Martins. 2023. Findings of the WMT 2023 shared task on quality estimation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 629–653, Singapore. Association for Computational Linguistics.

Maximillian Chen, Caitlyn Chen, Xiao Yu, and Zhou Yu. 2023. FastKASSIM: A fast tree kernel-based syntactic similarity metric. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 211–231, Dubrovnik, Croatia. Association for Computational Linguistics.

Yanran Chen and Steffen Eger. 2023. Menli: Robust evaluation metrics from natural language inference. *Transactions of the Association for Computational Linguistics*, 11:804–825.

Yanran Chen, Hannes Gröner, Sina Zarrieß, and Steffen Eger. 2024. Evaluating diversity in automatic poetry generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19671–19692, Miami, Florida, USA. Association for Computational Linguistics.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling

[20] `https://www.bmbf.de/SharedDocs/Publikationen/DE/FS/31580_Urheberrecht_in_der_Wissenschaft_en.pdf`

human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Joke Daems, Paola Ruffo, and Lieve Macken. 2024. Impact of translation workflows with and without MT on textual characteristics in literary translation. In *Proceedings of the 1st Workshop on Creative-text Translation and Technology*, pages 57–64, Sheffield, United Kingdom. European Association for Machine Translation.

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

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(1).

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

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Ana Guerberof-Arenas and Antonio Toral. 2022. Creativity in translation: Machine translation as a constraint for literary texts. *Translation Spaces*, 11(2):184–212.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent Machine Translation Evaluation through Fine-grained Error Detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Damien Hansen and Emmanuelle Esperança-Rodier. 2022. Human-adapted mt for literary texts: Reality or fantasy? In *NeTTT 2022*, pages 178–190.

Yuchen Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. 2022. BlonDe: An automatic evaluation metric for document-level machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1550–1565, Seattle, United States. Association for Computational Linguistics.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In *Proceedings of the Eighth Conference on Machine Translation*, pages 419–451.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, Miami, Florida, USA. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453, Miami, Florida, USA. Association for Computational Linguistics.

Waltraud Kolb. 2023. i am a bit surprised: Literary translation and post-editing processes compared. In *Computer-Assisted Literary Translation*, pages 53–68. Routledge.

Kollektive-Intelligenz. 2024. Ki aber wie? Übersetzertag 2024.

Christoph Leiter and Steffen Eger. 2024. Prexme! large scale prompt exploration of open source llms for machine translation and summarization evaluation. *arXiv preprint arXiv:2406.18528*.

Christoph Leiter, Juri Opitz, Daniel Deutsch, Yang Gao, Rotem Dror, and Steffen Eger. 2023. The Eval4NLP 2023 shared task on prompting large language models as explainable metrics. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, pages 117–138, Bali, Indonesia. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yiqi Liu, Nafise Moosavi, and Chenghua Lin. 2024. LLMs as narcissistic evaluators: When ego inflates evaluation scores. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12688–12701, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Arle Lommel, Aljoscha Burchardt, Maja Popović, Kim Harris, Eleftherios Avramidis, and Hans Uszkoreit. 2014. Using a new analytic measure for the annotation and analysis of mt errors on real data. In *Proceedings of the 17th Annual conference of the European Association for Machine Translation*, pages 165–172.

Qingyu Lu, Liang Ding, Kanjian Zhang, Jinxia Zhang, and Dacheng Tao. 2025. MQM-APE: Toward high-quality error annotation predictors with automatic post-editing in LLM translation evaluators. pages 5570–5587.

Lieve Macken. 2024. Machine translation meets large language models: Evaluating ChatGPT's ability to automatically post-edit literary texts. In *Proceedings of the 1st Workshop on Creative-text Translation and Technology*, pages 65–81, Sheffield, United Kingdom. European Association for Machine Translation.

Ruth Martin. 2023. Reflections from an ai translation slam: (wo)man versus machine.

Evgeny Matusov. 2019. The challenges of using neural machine translation for literature. In *Proceedings of the Qualities of Literary Machine Translation*, pages 10–19, Dublin, Ireland. European Association for Machine Translation.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? a call for multi-prompt llm evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949.

Monika Pfundmeier. Redaktion: Dorrit Bartel Nina George, André Hansen. 2023. Anwendungen von fortgeschrittener informatik und generativer systeme im buchsektor.

Magdalena Nizio. 2024. Künstliche intelligenz gefahr oder hilfe für literarische Übersetzung?

Jianhui Pang, Fanghua Ye, Longyue Wang, Dian Yu, Derek F Wong, Shuming Shi, and Zhaopeng Tu. 2024. Salute the classic: Revisiting challenges of machine translation in the age of large language models. *arXiv preprint arXiv:2401.08350*.

S Patro. 2015. Normalization: A preprocessing stage. *arXiv preprint arXiv:1503.06462*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Llukan Puka. 2011. *Kendall's Tau*, pages 713–715. Springer Berlin Heidelberg, Berlin, Heidelberg.

Natália Resende and James Hadley. 2024. The translator's canvas: Using LLMs to enhance poetry translation. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 178–189, Chicago, USA. Association for Machine Translation in the Americas.

Markella B Rutherford, Peggy Levitt, and Erika Zhang. 2024. Whence the 3 percent?: How far have we come toward decentering americas literary preference? *Global Perspectives*, 5(1):93034.

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.

Bryan Stroube. 2003. Literary freedom: Project gutenberg. *XRDS: Crossroads, The ACM Magazine for Students*, 10(1):3–3.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. 2022. Exploring document-level literary machine translation with parallel paragraphs from world literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9882–9902.

Craig Thomson, Ehud Reiter, and Anya Belz. 2024. Common flaws in running human evaluation experiments in NLP. *Computational Linguistics*, 50(2):795–805.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

Gys-Walt Van Egdom, Onno Kosters, and Christophe Declercq. 2023. The riddle of (literary) machine translation quality. *Tradumàtica tecnologies de la traducció*, (21):129–159.

Bram Vanroy, Marie-Aude Lefer, Lieve Macken, and Paola Ruffo, editors. 2024. *Proceedings of the 1st Workshop on Creative-text Translation and Technology*. European Association for Machine Translation, Sheffield, United Kingdom.

Rob Voigt and Dan Jurafsky. 2012. Towards a literary machine translation: The role of referential cohesion. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 18–25, Montréal, Canada. Association for Computational Linguistics.

Longyue Wang, Zhaopeng Tu, Yan Gu, Siyou Liu, Dian Yu, Qingsong Ma, Chenyang Lyu, Liting Zhou, Chao-Hong Liu, Yufeng Ma, Weiyu Chen, Yvette Graham, Bonnie Webber, Philipp Koehn, Andy Way, Yulin Yuan, and Shuming Shi. 2023. Findings of the WMT 2023 shared task on discourse-level literary translation: A fresh orb in the cosmos of LLMs. In *Proceedings of the Eighth Conference on Machine Translation*, pages 55–67, Singapore. Association for Computational Linguistics.

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

Mingzhou Xu, Longyue Wang, Derek F. Wong, Hongye Liu, Linfeng Song, Lidia S. Chao, Shuming Shi, and Zhaopeng Tu. 2022. GuoFeng: A benchmark for zero pronoun recovery and translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11266–11278, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jianhao Yan, Pingchuan Yan, Yulong Chen, Judy Li, Xianchao Zhu, and Yue Zhang. 2024. Gpt-4 vs. human translators: A comprehensive evaluation of translation quality across languages, domains, and expertise levels. *arXiv preprint arXiv:2407.03658*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Ran Zhang and Steffen Eger. 2024. Llm-based multi-agent poetry generation in non-cooperative environments. *arXiv preprint arXiv:2409.03659*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wei Zhao, Michael Strube, and Steffen Eger. 2023a. Discoscore: Evaluating text generation with bert and discourse coherence. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3865–3883.

Zoie Zhao, Sophie Song, Bridget Duah, Jamie Macbeth, Scott Carter, Monica P Van, Nayeli Suseth Bravo, Matthew Klenk, Kate Sick, and Alexandre LS Filipowicz. 2023b. More human than human: Llm-generated narratives outperform human-llm interleaved narratives. In *Proceedings of the 15th Conference on Creativity and Cognition*, pages 368–370.

A Corpora comparison

Table 4 summarizes the corpora for literary MT (evaluation). We report the annotated dataset and the entire MT parallel corpus separately.

Dataset	Evaluated Corpus								Entire Parallel Corpus			
	Scheme	#Evaluated segments			#systems	Source publication		Evaluation avialability				
		#doc	#sentences	#pairs		Classics	Contemporary		Source	Target	#pair	#doc
BWB (Jiang et al., 2022)	Span highlight (discourse phenomena)	80	2,632	1	1		✓ (Src: ?) (Tgt: ?)	✓	Chinese	English	1	196,304
GUOFENG (Xu et al., 2022)	Zero Pronoun	24	1, 658	1	1		✓ (Src: ?) (Tgt: 2005?-?)	✓	Chinese	English	1	1.5M
PAR3 (Thai et al., 2022)	Translation quality: BWS (2)	300?	unclear	3	3	✓ (Src: 1400-1966?) (Tgt: ?)		?	(French, Russian, German) + 13 more	English	16	121,385
Karpinska and Iyyer (2023)	Translation quality; lite-MQM, BWS (2)	1080	5301	**18**	3		✓ (Src: 1992-2021, + 1884 and 1956) (Tgt: 2017-2022)	✓	English, Polish, Russian, Czech, French, German, Japanese, and Chinese	English, Japanese, Polish	-	-
Ours	**Translation quality: MQM, SQM, BWS (4)**	**2197**	**13,346**	4	**10**	✓ (Src: 1774-1927) (Tgt: 1831-2023)	✓ (Src: 2010-2024) (Tgt: 2022-2024)	✓	**English, Chinese, German**	**English, Chinese, German**	**-**	**-**

Table 4: Summary of corpora for literary MT (evaluation). We separate the statistics of evaluated samples and the entire MT parallel corpus. We categorize the source works as either classics or contemporary in the source publication columns. We also indicate the time spans for both the source (src) and human translation (tgt) in brackets. "**?**" denotes an unclear time span. The other 13 languages for PAR3 are Spanish, Czech, Norwegian, Swedish, Portuguese, Italian, Japanese, Bengali, Tamil, Danish, Chinese, Dutch, Hungarian, Polish, Sesotho, and Persian. The **?** symbol in column evaluation availability indicates that the human evaluation referenced in the paper is not available in the specified GitHub repository.

B Corpus details

Table 5 summarizes the details of the publications and the corresponding translation versions.

C Prompt experiments for literary translation

We use GPT-4o to experiment with four types of prompts. The results are shown in Table 6.

D Human evaluation

D.1 Summary of human evaluation schemes

Figure 4 shows human evaluation schemes with an example.

D.2 System ranking

Table 7 shows the system ranking per evaluator and evaluation schemes.

D.3 Error distribution

Figure 5 shows the distribution of mean error count per segment (i.e., the total error count divided by the number of segments) aggregated over error categories.

D.4 Adequacy plot for scheme combination

Figure 6 shows the percentage change in distinguishing human translation from top system outputs by combining student SQM and MQM results. We use min-max scaling (Patro, 2015) to adjust MQM to the same range as SQM (0-6). The combined score is calculated as $(1-\alpha) \times S_{\text{MQM}} + \alpha \times S_{\text{SQM}}$, where S represents the score and α represents the ratio between SQM and MQM.

E Error annotation details

E.1 Total cost

The four student evaluators each work 60 hours at a rate of 14 per hour. They are given two months to complete the task. Prior to annotation, the authors and students collaboratively refine the guidelines through several iterations to address any unclear points. Additionally, two professional translators work a total of 20 hours at a rate of approximately 50 per hour. In total, this comprehensive evaluation project amounts to nearly 4,500.

E.2 MQM Annotation guidelines

We present our annotation guideline in this section and definition of translation errors in Table 8.

E.2.1 Annotation guideline

Introduction

For each annotation, you will be presented with two paragraphs: the source paragraph and the translation of the source paragraph. You need to read both paragraphs carefully and finish the following tasks.

Task description

Task 1: MQM error annotation

Step 1: You will be given the definition of translation errors. You need to highlight the spans from the translation containing errors and choose the corresponding error categories.

Please consider following the order of questions during error annotation:

- Evaluate whether the sentence is comprehensible. If it is completely incomprehensible, mark it as "non-translation."
- Are there any untranslated texts?
- Is there any cultural-specific terminology mistranslation or inconsistency?
- Any accuracy errors with addition/omission/misnomer or mistranslation especially due to overly literal translation or temporal component?
- Are there any style errors?
- Are there any fluency errors that cause inconsistency in non-terminology usage and coherence errors?

Step 2: Rate the severity of errors

Major Errors are errors that seriously distort the meaning of the source, in such a way that it becomes completely incomprehensible or that the essence of the message is lost. Another example is errors that may confuse or mislead the reader due to substantial changes in meaning.

Minor Errors dont lead to loss of meaning and wouldnt confuse or mislead the reader but would be noticed, would decrease stylistic quality, fluency, or clarity, or make the content less appealing.

You only need to annotate once if the error is repeated several times.

Task 2: SQM assessment

You need to rate the overall translation quality. How do you like the translation, especially in reflecting the stylistic, artistic, and emotional qualities of the original work? Give a score from 0-6.

Pairs	Book	Publication year				#Paragraph
		Source	Translation1	Translation2	Translation3	
En-De	Frankenstein	1818	1831	2018	2017	9
	Jane Eyre	1847	1864	2020	2008	13
	Wuthering Heights	1847	1941	1949		8
	Birnan Wood	2023	2023			16
De-En	Death in Venice	1912	1925	2012	2021	11
	The Metamorphosis	1915	2003	2003	2014	13
	Steppenwolf	1927	2013	2012		4
	Kairos	2021	2023			18
De-Zh	The sorrows of young werther	1774	2018	1980		4
	The Metamorphosis	1915	2014	2023		2
	Demian: Die Geschichte von Emil Sinclairs Jugend	1921	2019	2020	2014	16
	Steppenwolf	1927	2023	2011	2022	13
	Heimsuchung	2010	11.2022			12
En-Zh	Frankenstein	1818	2010	2016		10
	Jane Eyre	1847	2012	1989		13
	Wuthering Heights	1847	1956	1998	2023	7
	Chlorine	2024	2024			19

Table 5: Meta information for source and translation. #Paragraph denotes the number of paragraphs sampled from the book. Detailed information such as author, translator, and links to the source and translation are contained in the CSV file in the GitHub repository. Classic works, i.e., before year 1980, are underlined.

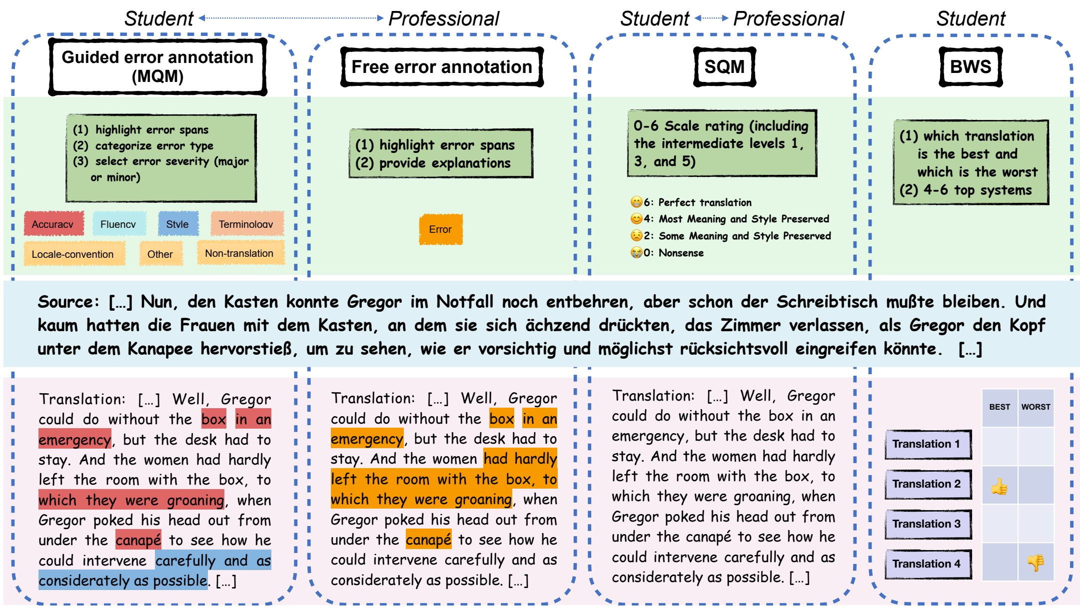
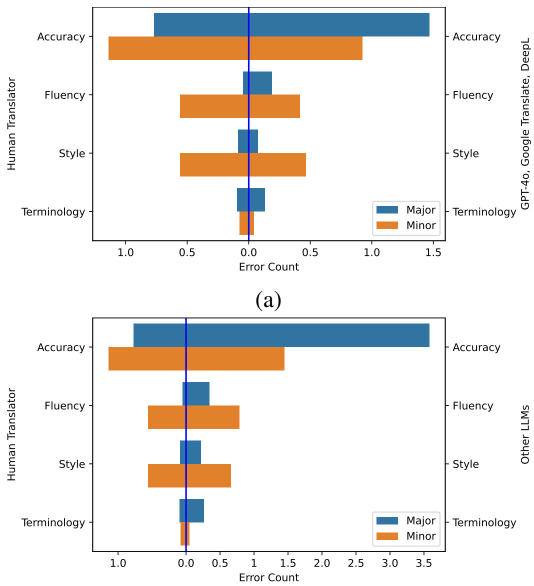


Figure 4: Human evaluation schemes conducted by student evaluators and professional translators with *De-En* example. MQM, SQM, and BWS represent Multidimensional Quality Metrics, Scalar Quality Metrics, and BEST-WORST SCALING. MQM is guided by an error categorization guideline and Free annotation permits span highlighting with comments, without specific guidelines.

mode	prompt	% winning
baseline	Please translate the following literary texts from src_lang to tgt_lang. The texts are as follows: text	27%
baseline + role	As a professional literary translator, your task is to translate the provided text with a focus on preserving the originals literary style, emotional depth, and cultural nuances. Your translation should not be a word-for-word rendering, but rather an adaptation that captures the aesthetic qualities, tone, and creative expression of the source. Ensure the translation reads naturally in the target language, maintaining fluidity and resonance with its literary conventions. Pay close attention to conveying the original mood, themes, and subtleties of meaning, while respecting both the integrity of the original work and the target audiences cultural context. Please translate the following literary texts from src_lang to tgt_lang. The texts are as follows: text	-10%
baseline + multistep error critics	Please translate the following literary texts from src_lang to tgt_lang. The texts are as follows: text 1. Please translate the source texts in [target language]. 2. Correct the text with attention to the following error categories: accuracy (including additions, omissions, misnomers, mistranslations such as overly literal translations and temporal inconsistencies, as well as untranslated text), fluency (ensuring consistency, coherence, proper grammar, punctuation, and spelling), style (avoiding awkward phrasing, incorrect register, inconsistencies, and unidiomatic expressions), terminology (using context-appropriate and consistent terms, particularly for culturally specific items and extra-linguistic terms), non-translation (ensuring all text is translated), and adherence to locale conventions. Avoid making any of these errors during the correction process. 3. Return the corrected translation.	3%
baseline + role + multistep error critics	As a professional literary translator, your task is to translate the provided text with a focus on preserving the originals literary style, emotional depth, and cultural nuances. Your translation should not be a word-for-word rendering, but rather an adaptation that captures the aesthetic qualities, tone, and creative expression of the source. Ensure the translation reads naturally in the target language, maintaining fluidity and resonance with its literary conventions. Pay close attention to conveying the original mood, themes, and subtleties of meaning, while respecting both the integrity of the original work and the target audiences cultural context. Please translate the following literary texts from src_lang to tgt_lang. The texts are as follows: text 1. First translate the source texts in tgt_lang. 2. Correct the text with attention to the following error categories: accuracy (including additions, omissions, misnomers, mistranslations such as overly literal translations and temporal inconsistencies, as well as untranslated text), fluency (ensuring consistency, coherence, proper grammar, punctuation, and spelling), style (avoiding awkward phrasing, incorrect register, inconsistencies, and unidiomatic expressions), terminology (using context-appropriate and consistent terms, particularly for culturally specific items and extra-linguistic terms), non-translation (ensuring all text is translated), and adherence to locale conventions. Avoid making any of these errors during the correction process. 3. Return the corrected translation.	-20%

Table 6: Prompt experiments with GPT-4o. The comparison is conducted using the BEST-WORST SCALING scheme. We report the average percentage of language pairs computed with Best-Count minus Worst-count divided by total evaluation count.



(a)

(b)

Figure 5: Distribution of mean error count per segment (i.e., the total error count divided by the number of segments) aggregated over error categories: (a) Human Translator vs. GPT-4o, Google Translate, DeepL. (b) Human Translator vs. other systems.

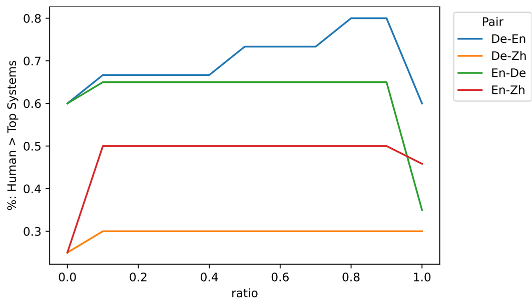


Figure 6: Percentage of segments where human translations are preferred over top systems by different combination ratios α, where $\alpha = 0$ corresponds to SQM alone and $\alpha = 1$ corresponds to MQM alone.

System	Model size (billion)	Student				Professional	
		MQM	Rank	SQM	Rank	SQM	Rank
Human	-	-1.3	**1**	5.0	**1**	5.0	**1**
Google Translate	?	-3.1	*3*	4.0	*3*	2.8	*3*
DeepL	?	-3.2	4	3.8	4	2.2	5
GPT-4o	>200	-1.7	<u>2</u>	4.6	<u>2</u>	3.2	<u>2</u>
Qwen	7.0	-8.7	5	2.7	5	(2.5)	(4)
Towerinstruct	7.0	-9.0	6	1.8	7		
Llama 3	8.0	-12.3	9	1.8	6		
Gemma	7.0	-12.6	10	1.3	8		
M2M	3.3	-11.2	7	1.1	9		
NLLB	1.3	-11.9	8	1.1	10		

Table 7: System rankings by evaluator and human evaluation scheme. We highlight the top three ranks in **bold**, <u>underline</u>, and *italics*, respectively. Professional translators evaluate Qwen outputs only for *De-Zh* and *En-Zh* pairs, which may result in higher ratings compared to Qwen's performance across all languages.

E.2.2 Error definition

The definition of translation errors is shown in Table 8. We derive our category definitions primarily from the MQM official website[21] and the study by Wang et al. (2023). We add two additional categories: **Non-translation** defined by Freitag et al. (2021) and **Temporal effect** defined by ourselves.

E.2.3 Screenshot of error annotation platform

Figure 7 demonstrates the screenshot of the error annotation platform.

E.3 Examples of human evaluation results

E.3.1 *De-En*

Figure 8 shows examples of SQM and error annotations by student and professional evaluators.

E.4 Example of human translation labelled as "error" according to student MQM

Table 9 shows examples of MQM errors labeled by student evaluators due to cultural adaptation in the translation.

F Automatic metrics

F.1 Detail of automatic metrics

Table 10 summarizes the characteristics of automatic metrics. For cost reasons, we use GPT-4o-mini instead of GPT-4 used in the original paper. This may induce a slight performance reduction. But compared to the cost for GPT-4 (on average 0.12 per segment, approximately 250 for complete evaluation per prompt), GPT-4o-mini is over 200 times cheaper (1.2 for full evaluation per prompt).

[21] `https://themqm.org/error-types-2/typology/`

	aspect	finetuned	base model	#parameter
Prometheus 2			prometheus2	7 billion
XCOMET-XL		X	XLMR-XL	3.5 billion
XCOMET-XXL		X	XLMR-XXL	10 billion
GEMBA-MQM	X		GPT-4o mini	?

Table 10: Summary of characteristics for automatic metrics. *Aspect* indicates whether the metric support evaluation per aspect such as error categories and #parameter indicates the size of the base model.

F.2 GEMBA-MQM detail

F.2.1 GEMBA-MQM prompt templates

Table 11 showcases the prompt templates.

F.2.2 Consistency analysis of GEMBA-MQM (Literary) under different temperature

Table 12 shows the consistency of GEMBA-MQM (Literary) under different temperature. We evaluate the consistency of GEMBA-MQM (Literary) on the full *De-En* dataset under varying temperatures, each temperature querying three times. From Table 12, we observe that even at temperature 0 (the deterministic setting), the resulting score shows slight variation while maintaining high overall consistency. The correlation coefficient between different queries decreases from 0.910 at temperature 0.0 to 0.856 at temperature 1.0. Similarly, the proportion of cases with small score differences ($\delta \leq 1$) drops from 78.8% to 73.3%, while larger differences ($\delta > 5$) increase from 4.3% to 7.5% when temperature increases. These results demonstrate that the scores remain robust and consistent at temperature 0 and across low temperatures but variability increases with higher temperatures.

Error Category			Definition
Accuracy	Addition		Error occuring in the target content that includes content not present in the source.
	Omission		Error where content present in the source is missing in the target.
	Misnomer		The target text is more/less specific than the raw
	Mistranslation	General	Translation does not accurately represent the source. Wrong translation in the target language. E.g., the wrong word is used in the target language; mistranslation of a false friend (word or expression that has a similar form in the source and target languages, but a different meaning);
		Overly literal	Direct translation of idioms, sentences, and structures;
		Temporal effect	Historical linguistic elements such as spelling, semantics, and morphology that affect the accuracy of translation.
	Untranslated		Error occuring when a text segment that was intended for translation is omitted in the target content.
Fluency	Punctuation/Spelling		Punctuation marks missing or used in the wrong way/Issues related to the spelling of words (Including those of capitalization hyphenated words and use of as risk for censored swear words.)
	Grammar		Issues related to the grammar or syntax of the text. other than spelling and orthography (especially inconsistency of the tenses and conditionals)
	Inconsistency		The text shows internal inconsistency
	coherence		The text is not semantically clear, logical, and consistent, and, therefore, it cannot be understood by the reader; E.g., - Unclear reference: Relative pronouns or other referential mechanisms unclear in their reference.
Style	Awkwardness		Style involving excessive wordiness or overly embedded clauses, often due to inappropriate retention of source text style in the target text.
	Register		Characteristic of text that uses a level of formality higher or lower than required by the specifications or general language conventions.
	Inconsistent		Style is inconsistent within a text
	Unidiomatic		Style that is grammatical, but unnatural.
Terminology	Mistranslation		A genre-specific or cultural-specific terminology is wrongly translated. Use of term that it is not the term a domain expert would use or because it gives rise to a conceptual mismatch.
	Inconsistent		Terminology is used in a consistent manner within the text
Locale Convention (non-cultrual)	Location Format		Using the wrong format for address, name etc.
	Number Format		The translated date, time, currency, and telephone use formats inappropriate for its locale.
Others			Other issues that havent been included in this list.
Non-translation			Non-translation error that can be used to tag an entire sentence which is too badly garbled to permit reliable identification of individual errors.

Table 8: MQM error definition

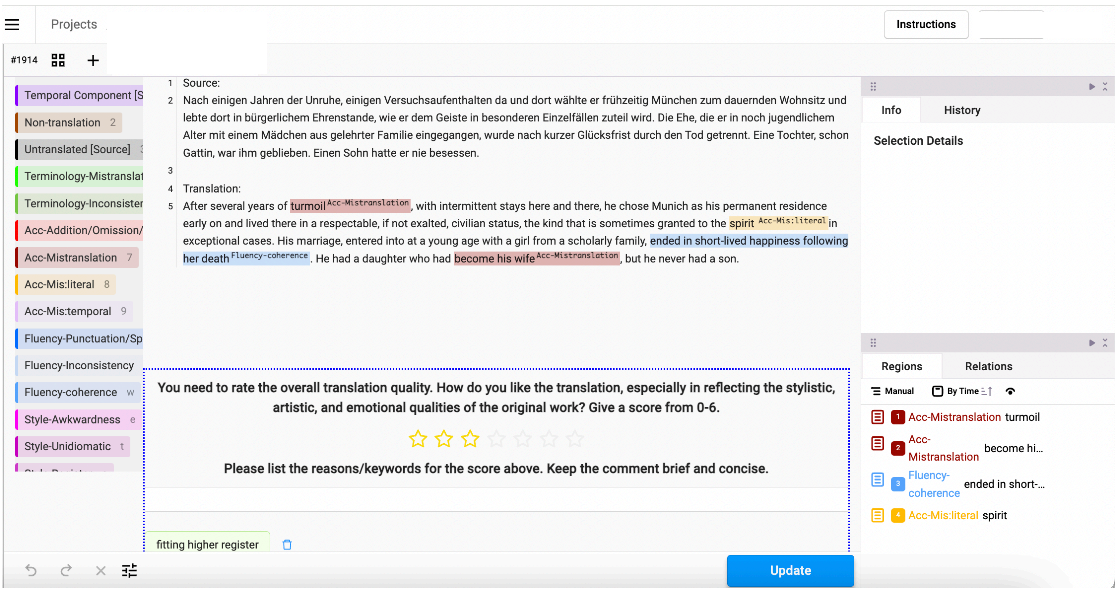


Figure 7: Screenshot of error annotation interface

Source	Und so ließ sie sich von ihrem Entschlusse durch die Mutter nicht abbringen, die auch in diesem Zimmer vor lauter Unruhe unsicher schien, bald verstummte und der Schwester nach Kräften beim Hinausschaffen des Kastens half. Nun, den Kasten konnte Gregor im Notfall noch entbehren, aber schon der Schreibtisch mußte bleiben. Und kaum hatten die Frauen mit dem Kasten, an dem sie sich ächzend drückten, das Zimmer verlassen, als Gregor den Kopf unter dem Kanapee hervorstieß, um zu sehen, wie er vorsichtig und möglichst rücksichtsvoll eingreifen könnte. Aber zum Unglück war es gerade die Mutter, welche zuerst zurückkehrte, während Grete im Nebenzimmer den Kasten umfangen hielt und ihn allein hin und her schwang, ohne ihn natürlich von der Stelle zu bringen. Die Mutter aber war Gregors Anblick nicht gewöhnt, er hätte sie krank machen können, und so eilte Gregor erschrocken im Rückwärtslauf bis an das andere Ende des Kanapees, konnte es aber nicht mehr verhindern, daß das Leintuch vorne ein wenig sich bewegte. Das genügte, um die Mutter aufmerksam zu machen. Sie stockte, stand einen Augenblick still und ging dann zu Grete zurück.			
Target (DeepL)	And so she did not allow herself to be dissuaded from her decision by her mother, who also seemed unsure in this room because of her restlessness, soon fell silent and helped her sister as much as she could to remove the box. Well, Gregor could do without the box in an emergency, but the desk had to stay. And the women had hardly left the room with the box, to which they were groaning, when Gregor poked his head out from under the canape to see how he could intervene carefully and as considerately as possible. But unfortunately, it was the mother who returned first, while Grete held the box in the next room and swung it back and forth without, of course, moving it. The mother, however, was not used to the sight of Gregor, it could have made her ill, and so Gregor, frightened, rushed backwards to the other end of the canape, but could no longer prevent the sheet in front from moving a little. That was enough to alert his mother. She stopped, stood still for a moment and then went back to Grete.	And so she did not allow herself to be dissuaded from her decision by her mother, who also seemed unsure in this room because of her restlessness, soon fell silent and helped her sister as much as she could to remove the box. Well, Gregor could do without the box in an emergency, but the desk had to stay. And the women had hardly left the room with the box, to which they were groaning, when Gregor poked his head out from under the canape to see how he could intervene carefully and as considerately as possible. But unfortunately, it was the mother who returned first, while Grete held the box in the next room and swung it back and forth without, of course, moving it. The mother, however, was not used to the sight of Gregor, it could have made her ill, and so Gregor, frightened, rushed backwards to the other end of the canape, but could no longer prevent the sheet in front from moving a little. That was enough to alert his mother. She stopped, stood still for a moment and then went back to Grete.	And so she did not allow herself to be dissuaded from her decision by her mother, who also seemed unsure in this room because of her restlessness, soon fell silent and helped her sister as much as she could to remove the box. Well, Gregor could do without the box in an emergency, but the desk had to stay. And the women had hardly left the room with the box, to which they were groaning, when Gregor poked his head out from under the canapé to see how he could intervene carefully and as considerately as possible. But unfortunately, it was the mother who returned first, while Grete held the box in the next room and swung it back and forth without, of course, moving it. The mother, however, was not used to the sight of Gregor, it could have made her ill, and so Gregor, frightened, rushed backwards to the other end of the canapé, but could no longer prevent the sheet in front from moving a little. That was enough to alert his mother. She stopped, stood still for a moment and then went back to Grete.	**Accuracy** Accuracy-Mistranslation Accuracy-Mistranslation (literal) Accuracy-Addition/Omission/Misnomer **Style** Style-Awkwardness Style-Register **Fluency** **Fluency-Coherence** **Fluency-Grammar** **Error (Translator)**
Annotator	Student 1	Student 2	Professional translator	
SQM	2	2	2	
Agreement (error span)	**student 1 - 2**: 0.400; **student 1 - translator**: 0.657; **student 2 - translator**: 0.284			

Figure 8: Annotation example for *De-En*

pair	Source	Human Translation	Interpretation
En-Zh	... a wilderness of crumbling (griffins) and shameless little boys ...	... () ...	The text in brackets is marked as a misnomer (a subcategory of accuracy error) by the student evaluator. While the source text mentions griffins, the translation simply uses "monster," since in the target culture, griffins as mythical creatures have no cultural significance.
De-Zh	Als (Lateinschüler) und Herrensöhnchen ...	() ...	The text in brackets is marked as a mistranslation (a subcategory of accuracy error) by the student evaluator. Though the source text mentions "Latin school," the translation uses "prestigious school" because Latin schools don't exist in the target culture and therefore lack the same connotation of academic prestige.
De-En	... ein (dreister Schlager) in unverständlichem Dialekt ...	... a ("big number") in incomprehensible dialect ...	The text in brackets is marked as a mistranslation (a subcategory of accuracy error) by the student evaluator, which is a culturally adapted translation.

Table 9: Mislabelled MQM error due to cultural adaptation

temperature	corr	score difference δ		
		$\delta \leq 1$	$1 < \delta \leq 5$	$\delta > 5$
0.0	0.910	78.8%	16.9%	4.3%
0.1	0.902	76.9%	16.9%	6.2%
0.3	0.882	74.7%	18.7%	6.6%
0.5	0.887	71.9%	22.6%	5.5%
0.7	0.874	72.2%	20.5%	7.3%
1.0	0.856	73.3%	19.2%	7.5%

Table 12: Consistency of GEMBA-MQM scores across different temperatures. We query three times at each temperature. Corr calculates the mean rank correlation between queries using Spearman correlation. The score difference shows the percentage of score changes under different thresholds. $\delta = 1$ represents one minor error in difference, while $\delta = 5$ represents one major error.

F.3 Prompt for Prometheus 2

Component	prompt
ABS_SYSTEM PROMPT	As a literary translation critic, your role is to identify errors and evaluate the translations quality. Focus on the subtleties of literary style, emotional impact, and creative expression. An excellent translation captures the original works aesthetic qualities, tone, and cultural nuances, rather than adhering to a word-for-word approach. Translations that are excessively literal and fail to adapt to the target languages literary conventions and natural flow should be critiqued accordingly.
rubric_data	**"criteria"**: "What is the overall quality of the given literary translation based on the source texts?" **"score0_description"**: "Nonsense: Nearly all information is lost between the translation and source. Grammar and style are irrelevant." **"score2_description"**: "Some Meaning and Style Preserved: The translation preserves some of the meaning and style of the source but misses significant parts. The narrative is hard to follow due to fundamental errors. Grammar may be poor. Style may be inconsistent." **"score4_description"**: "Most Meaning and Style Preserved and Few Grammar Mistakes: The translation retains most of the meaning and style of the source. This may contain some grammar mistakes or minor style and contextual inconsistencies." **"score6_description"**: "Perfect Meaning and Style Preserved: The meaning and style of the translation are completely consistent with the source and the surrounding context."

Table 13: Prompts for Prometheus 2

F.4 Correlation between SQM and automatic metrics

Figure 9 illustrates segment-level correlation between human SQM and automatic metrics.

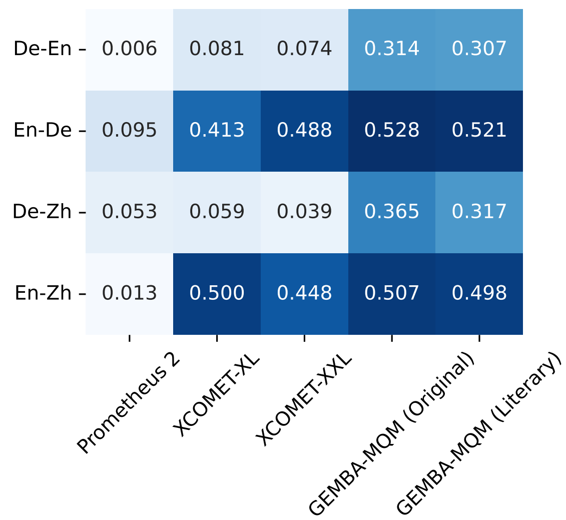


Figure 9: Segment level correlation measured by Kendalls tau between human SQM and automatic metrics.

F.5 Human vs. LLMs for metrics.

Table 14 shows the percentage of segments where human translations are preferred over machine translations per evaluation schemes and automatic metrics.

	GEMBA-MQM (Original)	GEMBA-MQM (Literary)
system	You are an annotator for the quality of machine translation. Your task is to identify errors and assess the quality of the translation.	As a literary translation critic, your role is to identify errors and evaluate the translation's quality. Focus on the subtleties of literary style, emotional impact, and creative expression. An excellent translation captures the original work's aesthetic qualities, tone, and cultural nuances, rather than adhering to a word-for-word approach. Translations that are excessively literal and fail to adapt to the target language's literary conventions and natural flow should be critiqued accordingly.
prompt	Based on the source segment and machine translation surrounded with triple backticks, identify error types in the translation and classify them. The categories of errors are: accuracy (addition, omission, misnomer, mistranslation [including too-literal translation and temporal effect], untranslated text), fluency (inconsistency, coherence, grammar, punctuation, spelling), style (awkward, register, inconsistent, unidiomatic), terminology (inappropriate for context, inconsistent use, please pay attention to cultural specific items and extra-linguistic terms), non-translation, other, locale convention, or no error. Each error is classified as one of three categories: critical, major, and minor. Critical errors inhibit comprehension of the text. Major errors disrupt the flow, but what the text is trying to say is still understandable. Minor errors are technically errors, but do not disrupt the flow or hinder comprehension.	
few shot example 1	*En-De*: { "source_lang": "English", "source_seg": "I do apologise about this, we must gain permission from the account holder to discuss an order with another person, I apologise if this was done previously, however, I would not be able to discuss this with yourself without the account holders permission.", "target_lang": "German", "target_seg": "Ich entschuldige mich dafür, wir müssen die Erlaubnis einholen, um eine Bestellung mit einer anderen Person zu besprechen. Ich entschuldige mich, falls dies zuvor geschehen wäre, aber ohne die Erlaubnis des Kontoinhabers wäre ich nicht in der Lage, dies mit dir involvement.", "answer": Critical: no-error Major: accuracy/mistranslation - "involvement" accuracy/omission - "the account holder" Minor: fluency/grammar - "wäre" style/register - "dir" }	*En-De*: { "source_lang": "English", "source_seg": "At intervals, while turning over the leaves of my book, I studied the aspect of that winter afternoon.", "target_lang": "German", "target_seg": "Von Zeit zu Zeit, während ich in meinem Buch blätterte, studierte ich das Aussehen dieses Winternachmittags.", "answer": Critical: accuracy/mistranslation (Too-literal) - "studierte" Major: accuracy/omission - "das Aussehen" Minor: no-error }
few shot example 2	*Zh-En*: {"source_lang": "Chinese", "source_seg": "", "target_lang": "English", "target_seg": "Urumqi Home Furnishing Store Channel provides you with the latest business information such as the address, telephone number, business hours, etc., of high-speed rail, and find a decoration company, and go to the reviews.", "answer": Critical: accuracy/addition - "of high-speed rail" Major: accuracy/mistranslation - "go to the reviews" Minor: style/awkward - "etc.," }	*Zh-En*: { "source_lang": "Chinese", "source_seg": "", "target_lang": "English", "target_seg": "The sun he has feet, ah, gently and quietly moved.", "answer": Critical: style/awkward - "ah" Major: fluency/grammar - "gently and quietly moved" Minor: accuracy/mistranslation (Too-literal) - "he has feet" }

Table 11: GEMBA-MQM prompt templates.

pair	Human > Top Systems (GPT-4o, DeepL, Google Translate, Qwen 2)							Human > Other Systems (excluding GPT-4o, DeepL, Google Translate)						
	Student		Prometheus 2	XCOMET		GEMBA-MQM		Student		Prometheus 2	XCOMET		GEMBA-MQM	
	MQM	SQM		XL	XXL	Original	Literary	MQM	SQM		XL	XXL	Original	Literary
De-En	60.0%	60.0%	6.7%	6.7%	13.3%	6.7%	0.0%	86.7%	80.0%	6.7%	20.0%	33.3%	26.7%	33.3%
En-De	35.0%	60.0%	10.0%	5.0%	20.0%	0.0%	10.0%	95.0%	100.0%	10.0%	70.0%	75.0%	70.0%	60.0%
De-Zh	30.0%	25.0%	15.0%	30.0%	10.0%	15.0%	20.0%	95.0%	90.0%	25.0%	30.0%	45.0%	45.0%	50.0%
En-Zh	45.8%	25.0%	8.3%	20.8%	41.7%	0.0%	8.3%	87.5%	79.2%	4.2%	70.8%	75.0%	41.7%	50.0%
Mean	42.7%	42.5%	10.0%	15.6%	21.3%	5.4%	9.6%	91.0%	87.3%	11.5%	47.7%	57.1%	45.8%	48.3%

Table 14: Percentage of segments where human translations are preferred over machine translations per human evaluation schemes and automatic metrics.

Figure 1 is based on Table 3 and Table 14 (appendix). For the percentage computation of scheme professional error annotation, we first convert error span annotation counts to numerical scores using the formula Score $=$ $|\text{Good solution}| - |\text{Error}|$, then rank these scores to compute the percentage. The percentage of human $>$ top systems averaged over three language pairs is 62.8%, with *De-En* at 75.0%, *En-De* at 33.3%, and *En-Zh* 80.0%, respectively.

The professional BWS statistics shown in Figure 1 are based on the results from *En-Zh*, with 14 samples evaluated by the *En-Zh* translator.

F.6 Computation details for diversity measure

To assess a system's lexical diversity compared to others, we calculate the pairwise lexical overlap between translations from different systems. For each paragraph, we compute the lexical overlap between the current system's translation and those of all other systems. This overlap quantifies the vocabulary shared between the current system and others. We then average these overlaps for all paragraphs, resulting in an overall lexical overlap score for the system. A lower average overlap indicates that the system uses a more diverse and distinct vocabulary, while a higher score suggests greater similarity in word choice with other systems.

The mean lexical overlap for system j across all paragraphs is given by:

$$\text{Avg.overlap}_j = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{M-1}\sum_{\substack{k=1\\k\neq j}}^{M} S(T_{ij}, T_{ik})\right)$$

where N is the total number of paragraphs. M is the total number of systems. T_{ij} is the translation of paragraph i by system j. $S(T_{ij}, T_{ik})$ is the function, i.e., BLEU by SACREBLEU (Post, 2018), to compute lexical overlap between the translation from system j and system k for paragraph i.

F.7 Comparison of classic and contemporary works

Figure 10 shows the scatter plots of human MQM (z-axis) per average syntactic similarity between source and system translation and average lexical overlap of the current system to all other systems for all contemporary works after 2000. Human translation still holds the lowest lexical overlap (19.6) than other systems except for Qwen 2 (19.5). The syntactic similarity of human translation (0.267) ranks third slightly higher than Qwen2 (0.252) and NLLB (0.253).

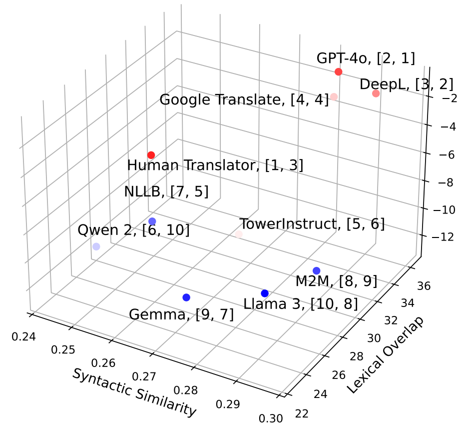


Figure 10: (Contemporary works after 2000) Scatter plots of human MQM (z-axis) per average syntactic similarity between source and system translation and average lexical overlap of the current system to all other systems. The rank of systems is marked as [rank per human MQM, rank per GEMBA-MQM (Literary)].

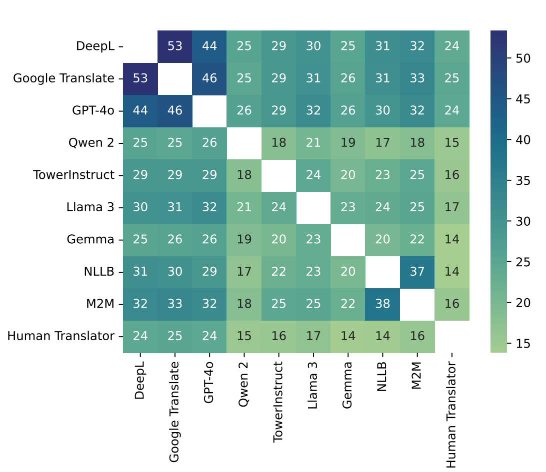


Figure 11: Heatmap for pairwise lexical overlap between translation systems (calculation based on the entire LITEVAL-CORPUS).

Figure 11 shows the heatmap of pairwise lexical overlap between systems based on the entire LITEVAL-CORPUS. The results reveal exceptionally high lexical overlap among Google Translate, GPT-4o, and DeepL, while NLLB and M2M (all from META AI) also show notably high overlap. This pattern may stem from shared training corpora or training data distillation from other models.

F.8 Comparison of different human translation versions

Figure 12 compares different human translation versions over time of the same source per lexical overlap (upper) and syntactic similarity (lower). Versions 1-3 represent translations from the oldest to newest versions in our study. Our calculations are based on paragraphs from our LITEVAL-CORPUS. For lexical overlap with other LLM systems (upper), 8 out of 12 books show that the newest translations show substantially higher overlap with other system translations (e.g., Wuthering Heights, Frankenstein, and Jane Eyre) compared to previous versions. However, some older translations (e.g., Wuthering Heights *En-Zh* and Death in Venice *De-En*) also show high overlap, likely because they are part of the training data of the translation systems. Regarding syntactic similarity between source and translation (lower), we observe no consistent pattern across versions, with some translations like the Metamorphosis *De-Zh* showing higher syntactic similarity while others remain lower or comparable.

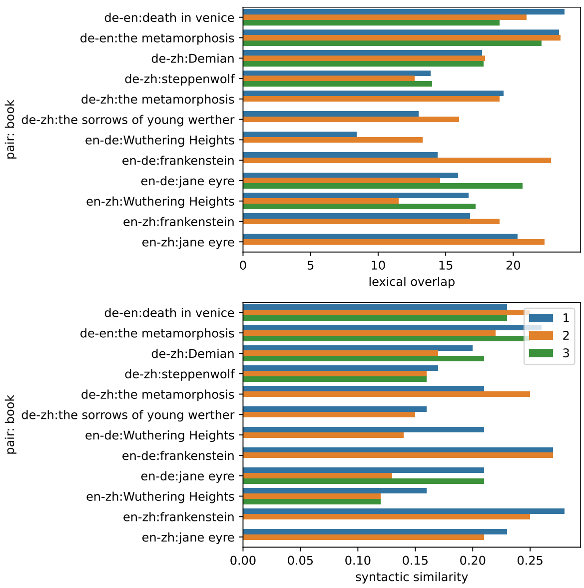


Figure 12: Comparison of lexical overlap and syntactic similarity for different human translation versions.

F.9 Example of system outputs with human and metric scores, syntactic similarity, and lexical overlap

Table 15 and 16 demonstrate a full example including all system outputs with evaluation from human and automatic metrics, syntactic similarity to the source, and lexical overlap to all other systems.

Source						
Aus diesem Traum erwachte der Heimgesuchte entnervt, zerrüttet und kraftlos dem Dämon verfallen. Er scheute nicht mehr die beobachtenden Blicke der Menschen; ob er sich ihrem Verdacht aussetze, kümmerte ihn nicht. Auch flohen sie ja, reisten ab; zahlreiche Strandhütten standen leer, die Besetzung des Speisesaals wies gröSSere Lücken auf, und in der Stadt sah man selten noch einen Fremden. Die Wahrheit schien durchgesickert, die Panik, trotz zähen Zusammenhaltens der Interessenten, nicht länger hintanzuhalten. Aber die Frau im Perlenschmuck blieb mit den Ihren, sei es, weil die Gerüchte nicht zu ihr drangen, oder weil sie zu stolz und furchtlos war, um ihnen zu weichen: Tadzio blieb; und jenem, in seiner Umfangenheit, war es zuweilen, als könne Flucht und Tod alles störende Leben in der Runde entfernen und er allein mit dem Schönen auf dieser Insel zurückbleiben,ja, wenn vormittags am Meere sein Blick schwer, unverantwortlich, unverwandt auf dem Begehrten ruhte, wenn er bei sinkendem Tage durch Gassen, in denen verheimlichterweise das ekle Sterben umging, ihm unwürdig nachfolgte, so schien das Ungeheuerliche ihm aussichtsreich und hinfällig das Sittengesetz.						
Model	**Translation**	**Syntactic similarity (Src-Tgt)**	**Lexical overlap**	**Human MQM**	**SQM**	**GEMBA-MQM (Literary)**
Human Translator 1	When he awakened from the affliction of this dream he was unnerved, shattered, and hopelessly under the power of the demon. He no longer avoided the inquisitive glances of other people; he did not care if he was exciting their suspicions. And as a matter of fact they were fleeing, travelling elsewhere. Numerous bathing houses stood empty, the occupants of the dining-hall became more and more scattered, and in the city now one rarely saw a foreigner. The truth seemed to have leaked out; the panic, despite the reticence of those whose interests were involved, seemed no longer avoidable. But the woman with the pearls remained with her family, either because the rumours had not yet reached her, or because she was too proud and fearless to heed them. Tadzio remained. And to Aschenbach, in his infatuation, it seemed at times as though flight and death might remove all the disturbing elements of life around them, and he stay here alone with the boy. Yes, by the sea in the forenoon when his eyes rested heavily, irresponsibly, unwaveringly on the thing he coveted, or when, as the day was ending, he followed shamelessly after him through streets where the hideous death lurked in secretat such times the atrocious seemed to him rich in possibilities, and laws of morality had dropped away.	0.147	17.1	-0.5	5	-2.6
Human Translator 2	The stricken man awoke from his dream unnerved and shaken, powerless in the demons thrall. He no longer shunned the curious glances of the people around him: whether he aroused their suspicion was of no concern to him. They were fleeing, were they not, leaving Venice: many of the cabanas were vacant, a number of tables in the dining room unoccupied, and in the city there was scarcely a foreigner to be seen. The truth seemed to have leaked out despite the tenacious solidarity among interested parties in their attempt to ward off panic. But the woman in pearls stayed on with her family either because the rumors had not reached her or because she was too proud and fearless to heed them. Thus Tadzio stayed on, and Aschenbach, in his beleaguered state, occasionally had the feeling that flight and death might eliminate all life standing in his way and leave him alone on this island with the beautiful boy; indeed, when he sat in the morning by the sea, his gazeheavy, injudicious, and fixedresting on the object of his desire, or when, as evening fell, he resumed his undignified pursuit through the narrow streets clandestinely haunted by loathsome dying, things monstrous seemed auspicious and the moral code null and void.	0.157	19.1	-0.4	6	-2.6
Qwen 2	From that dream, the wanderer awoke, unsettled, broken, and without strength, surrendering to the demon. He no longer feared the watchful gazes of people; it didn't bother him if they attributed their suspicions. Indeed, they too had fled or left; many beach huts stood empty, the dining hall indicated greater vacuity, and in the city, one seldom saw strangers anymore. The truth seemed to spread, but despite the stubborn unity of the interested parties, panic could not be held back. But the woman adorned with pearls remained, whether shielded from the rumors or too proud and fearless to evade them - Tadzio stayed. In his expanse, it sometimes felt as though flight and death could remove all distractions from the circle, leaving him alone with the Beauty on this island. Indeed, when, in the early morning, his gaze, deliberate, unaccountable, and unwavering, rested on the desired object across the sea, or as he followed the illicit death in hidden streets during descending days, feeling unworthy, he found the abyssal allure irresistible and the moral law.	0.194	18.6	-3.4	3	-7.6
Llama 3	From this dream, the Heimgesuchte woke up, worn out, shattered, and completely under the demon's control. He no longer feared the observing gazes of people; whether he risked being suspected by them or not, he didn't care. After all, they were fleeing, traveling away; many beach huts stood empty, the occupancy of the dining hall showed larger gaps, and in the city, it was rare to see a stranger. The truth had seeped in, and despite the persistent unity of the interested parties, panic was no longer possible to conceal. But the woman in the pearl jewelry remained with hers, whether the rumors didn't reach her or because she was too proud and fearless to leave them; Tadzio stayed; and to him, in his limited scope, it was sometimes as if flight and death could remove all disturbing life in the circle and leave him alone with the beautiful on this island. - yes, when, in the morning, his gaze, heavy, unaccountable, unattached, rested on the coveted object, when, with the sinking day, he followed unworthily through alleys where concealed, secretly, death was spreading, he seemed to find it promising and futile the law of morals.	0.208	24.2	-5	4	-7.6
TowerInstruct	From this dream the tormented one awoke weary, destroyed, and spiritually weakened to the demon's hold. He no longer shied away from people's gazes; whether he exposed himself to their suspicion did not concern him. Even they, in fact, fled, traveled away; numerous beach huts stood empty, the occupancy of the dining hall showed larger gaps, and one seldom still saw a stranger in the city. The truth seemed to seep through, and the panic, despite the interested party's staunch collective effort, was no longer to be kept at bay. But the woman in pearls stayed with her own, whether the rumors did not reach her, or whether she was too proud and fearless to budge: Tadzio remained; and to that extent, it was as though fleeing and death could all disturbing life around him and allow him alone with the beautiful person on this island to remainyes, when in the morning by the sea his gaze was heavy, irresponsible, uncharming on the desired one, when he, discreetly, followed him through alleys in which the uncomfortable death circled around him, then it seemed like the monstrous was promising and the moral code fleeting.	0.252	23.2	-2.2	1	-7.4

Table 15: (Part 1) Example of system outputs showing syntactic similarity to the source and lexical overlap with all other systems. Systems are sorted from low to high syntactic similarity.

Model	Translation	Syntactic similarity (Src-Tgt)	Lexical overlap	Human MQM	SQM	GEMBA-MQM (Literary)
Gemma	The troubled awakened from his dream, unnerved, discomposed and adrift before the demon. He no longer feared the watchful eyes of the people; whether he should evade their suspicions mattered not. They were fleeing too, traveling away; many beach shacks stood empty, the dining hall was sparsely populated, and the city held scarcely a stranger. Truth seemed to have drained away, panic struggling to contain itself despite the desperate clinging of the interested parties. But the woman in the pearl jewelry stayed with him, whether the rumors had not reached her, or because she was too proud and fearless to yield to them: Tadzio remained; and in his isolation, it sometimes seemed that flight and death could erase all that was disturbing in lifes round and leave him alone with the beautiful on this islandyes, even if this morning, at the seas edge, his gaze hung heavily, irresponsible and untethered, on the beloved one, while behind him, in the whispering gutters, death lingered in silent mockery, making him unworthy of its pursuit.	0.255	20.2	-2.6	4	-7.6
M2M	From this dream the home-seeking man woke up nervous, crushed, and powerlessly fallen to the devil. He no longer feared the observing eyes of the people; whether he exposed himself to their suspicion, he did not care. They also fled, they went away; numerous beach chalets were empty, the occupation of the dining room showed greater gaps, and in the city one rarely saw a stranger. The truth seemed to have leaked, the panic, despite the tight cohesion of the interested parties, no longer holds back. But the woman in the pearl jewelry remained with her own, either because the rumors did not come to her, or because she was too proud and fearless to evade them: Tadzio remained; and to him, in his possession, it was sometimes as if escape and death could remove all the disturbing life in the circle and he remained alone with the Beautiful on this islandyes, when in the morning at the sea his look rested heavy, irresponsible, unrelated to the desired, when he in the sinking day by the streets, in which mysteriously the ugly died, followed him unworthy, so the monstrous seemed to him in sight and fall the law.	0.256	27.9	-5	2	-7.4
GPT-4o	Of this dream, the afflicted one awoke unnerved, shattered, and powerless, succumbed to the demon. He no longer shunned the observing gazes of people; whether he exposed himself to their suspicion did not concern him. They too fled, departed; numerous beach huts stood empty, the occupancy of the dining room showed greater gaps, and in the town, one rarely saw a stranger anymore. The truth seemed to have leaked out, and the panic, despite the stubborn cohesion of interests, could no longer be held back. But the woman with the pearl jewelry stayed with her companions, whether because the rumors did not reach her or because she was too proud and fearless to yield to them: Tadzio stayed; and to the one enveloped by his feelings, it sometimes seemed as if flight and death could remove all the disturbing life around them, and he alone would remain with the beautiful one on this island,yes, when in the mornings by the sea his gaze rested heavily, irresponsibly, unwaveringly on the desired one, when he at the end of the day followed him unworthily through streets where hidden death crept, the monstrous seemed promising, and the moral law obsolete.	0.256	29.7	-0.6	5	-2.6
Google Translate	The afflicted man awoke from this dream, unnerved, shattered and powerless, succumbing to the demon. He no longer feared the watching eyes of people; he did not care whether he exposed himself to their suspicions. And they fled, departed; many beach huts stood empty, the dining room was empty and strangers were rarely seen in the town. The truth seemed to have leaked out, and the panic, despite the stubborn solidarity of those interested, could no longer be contained. But the woman in the pearl jewelry stayed with her family, either because the rumors did not reach her or because she was too proud and fearless to give in to them: Tadzio stayed; and to him, in his embarrassment, it sometimes seemed as if flight and death could remove all disturbing life in the area and he alone with the beautiful on this island - yes, when in the morning by the sea his gaze rested heavily, irresponsibly, unflinchingly on the desired one, when as the day was setting he unworthily followed him through streets in which the disgusting death was secretly going around, then the monstrous seemed to him promising and the moral law invalid.	0.257	30.6	-0.8	3	-7.4
NLLB	From this dream the house-boy awoke, nervous, broken, and powerless to fall to the demon. He no longer feared the watching eyes of men; whether he was subject to their suspicion, he did not care. They also fled, they traveled away; numerous beach huts were empty, the occupation of the dining-room showed greater gaps, and in the city one seldom saw a stranger again. The truth seemed to leak through, the panic, despite the tough unity of the interested parties, no longer held back. But the woman in pearl jewelry looked with your eyes, whether it was because the rumors did not press her, or because she was too proud and fearless to reproach them: that one, who remained in captivity, and who, with his mind, could escape and rest for days, when death seemed to him more difficult than the pursuit of the immortal guest, and who, by the unfailing grace of the unknown, returned to the unknown, and so untroubled by the unknown, was able to escape from the shame of the unknown, and to rest with the unforgettable life, as he had done in the unforgettable manner of the unknown.	0.260	20.5	-4.4	1	-12.4
DeepL	The haunted man awoke from this dream exasperated, shattered and powerless, having succumbed to the demon. He no longer shied away from the observing eyes of the people; whether he exposed himself to their suspicion was of no concern to him. They also fled, left; numerous beach huts stood empty, there were large gaps in the dining room staff, and it was rare to see a stranger in the town. The truth seemed to have leaked out and the panic could no longer be contained, despite the tenacious efforts of the interested parties. But the woman in the pearl jewelry stayed with her own, either because the rumors did not reach her or because she was too proud and fearless to give way to them: Tadzio remained; and to him, in his entrapment, it sometimes seemed as if flight and death could remove all disturbing life in the circle, and he could remain alone with the beautiful one on this island,-yes, when in the morning by the sea his gaze rested heavily, irresponsibly, unwaveringly on the desired one, when, as the day sank, he followed him unworthily through alleys where secretly the nasty dying was going on, the monstrous seemed to him promising and the moral law obsolete.	0.261	31.3	-2.6	4	-2.6

Table 16: (Part 2) Example of system outputs showing syntactic similarity to the source and lexical overlap with all other systems. Systems are sorted from low to high syntactic similarity.