LITRANSPROQA: An LLM-based LITerary TRANSlation Evaluation Metric with PROfessional Question Answering

Ran Zhang[1,5], Wei Zhao[2], Lieve Macken[3], Steffen Eger[4,5],
[1]University of Mannheim, School of Business Informatics and Mathematics
[2]University of Aberdeen, Department of Computing Science
[3]University of Gent, Department of Translation, Interpreting and Communication
[4]University of Technology Nuremberg (UTN), Department Engineering
[5]Natural Language Learning and Generation (NLLG) Lab
[1] ran.zhang@uni-mannheim.de [2] wei.zhao@abdn.ac.uk [3] lieve.macken@ugent.be [4] steffen.eger@utn.de

Abstract

The impact of Large Language Models (LLMs) has extended into literary domains. However, existing evaluation metrics for literature prioritize mechanical accuracy over artistic expression and tend to overrate machine translation as being superior to human translation from experienced professionals. In the long run, this bias could result in an irreversible decline in translation quality and cultural authenticity. In response to the urgent need for a specialized literary evaluation metric, we introduce LITRANSPROQA, a novel, reference-free, LLM-based question-answering framework designed for literary translation evaluation. LITRANSPROQA integrates humans in the loop to incorporate insights from professional literary translators and researchers, focusing on critical elements in literary quality assessment such as literary devices, cultural understanding, and authorial voice. Our extensive evaluation shows that while literary-finetuned XCOMET-XL yields marginal gains, LITRANSPROQA substantially outperforms current metrics, achieving up to 0.07 gain in correlation and surpassing the best state-of-the-art metrics by over 15 points in adequacy assessments. Incorporating professional translator insights as weights further improves performance, highlighting the value of translator inputs. Notably, LITRANSPROQA reaches an adequacy performance comparable to trained linguistic student evaluators, though it still falls behind experienced professional translators. LITRANSPROQA shows broad applicability to open-source models like LLaMA3.3-70b and Qwen2.5-32b, indicating its potential as an accessible and training-free tool for evaluating literary translations that require local processing due to copyright or ethical considerations.

1 Introduction

Large Language Models (LLMs) have shown remarkable capabilities in linguistic tasks, emerging as potentially transformative tools across many domains (Grattafiori et al., 2024; Yang et al., 2024; Achiam et al., 2023; Eger et al., 2025). However, their suitability for more nuanced creative areas—such as literary translation or poetry generation—remains uncertain (Macken, 2024; Chakrabarty et al., 2024; Al-Awawdeh, 2021; Chen et al., 2024; Belouadi and Eger, 2023; Zhang and Eger, 2024). Literary translation requires not just lexical and syntactic precision, but also a deep understanding of cultural context, aesthetic style, and interpretive nuance (Wang et al., 2023; Karpinska and Iyyer, 2023; Matusov, 2019; Pang et al., 2024). To critically assess the suitability of LLMs for such complex creative work, it is essential to establish robust and systematic *evaluation methods* that can truly capture the essence of literary translation.

While human evaluation appears ideal for assessing literary translation qualities, it becomes economically impractical at scale, given the vast corpus of world literature and LLMs' unprecedented generation capabilities. Moreover, proper evaluation requires input from trained literary professionals, making human assessment prohibitively expensive (Zhang et al., 2025; Yan et al., 2024).

The inherent nature of literary translation compounds this evaluation challenge. While technical texts often have clear "correct" translations, literary works demand creative reinterpretation across linguistic and cultural boundaries. This poses fundamental problems for reference-based evaluation methods, as generating reference translations for literary texts is not only resource-intensive but also conceptually problematic, since multiple valid interpretations can exist simultaneously.

Existing automatic evaluation approaches fall short fundamentally: previous metrics like BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005), and later embedding-based approaches such as BERTScore (Zhang et al., 2020) and BLEURT (Sellam et al., 2020) primarily mea-

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 29099–29121
November 4-9, 2025

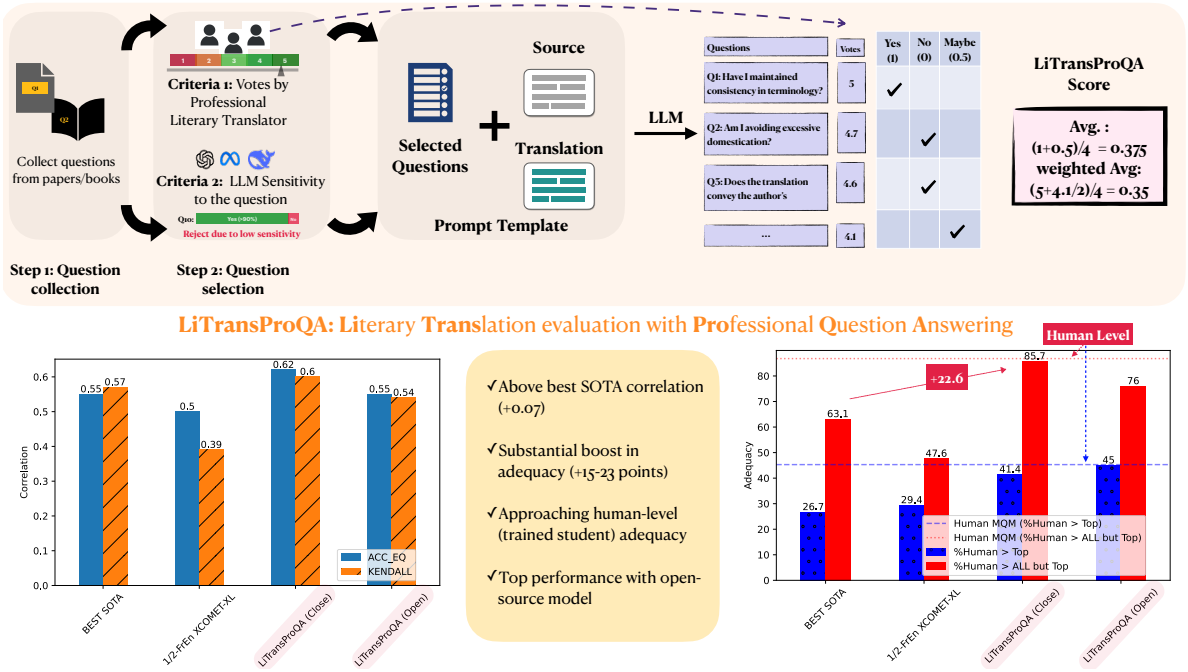


Figure 1: Overview of LITRANSPROQA and its performance compared to finetuned and LLM-based SOTA in correlation and adequacy (the ability to rate high-quality human translation better than MT). Human MQM (dashed lines) represents the adequacy level of trained students using MQM. Human level refers to trained student evaluators, who still lag behind experienced professional translators by a large margin according to Zhang et al. (2024).

sure semantic equivalence and linguistic accuracy, overlooking core literary attributes, such as tone, cultural specificity, and figurative language.

Even recent state-of-the-art (SOTA) automatic evaluation approaches, including XCOMET-XL & XCOMET-XXL (Guerreiro et al., 2024), LLM-based metrics like GEMBA-MQM (Kocmi and Federmann, 2023), and hybrid LLMs like Prometheus (Kim et al., 2024a), show limitations in literary contexts, despite their design to handle reference-free evaluation and potentially cover surface-level stylistic elements. Zhang et al. (2025) demonstrate that these metrics may consistently prefer machine-generated literary translations over translations written by experienced professional translators, which severely misaligns with judgments from human experts. The lack of professional translation expertise in current metrics, particularly regarding literary translation nuances and quality assessment standards, contributes to their limited ability to conduct meaningful evaluations.

This paper addresses these limitations by introducing LITRANSPROQA, as shown in Figure 1, a novel LLM-based LITerary TRANSlation evaluation metric focusing on PROfessional Question Answering. LITRANSPROQA integrates humans in the loop to reflect professional translators' quality control and assessment processes.[1] Unlike existing QA-based MT metrics (Krubiński et al., 2021; Fernandes et al., 2025), LITRANSPROQA focuses on core elements in literary translation proposed and verified by researchers and professional literary translators for better alignment with human experts. Our work presents a detailed question development and selection process for an optimized set of questions, as our analysis indicates that LLMs are not yet fully trustworthy for automatic question generation and QA evaluation in the literary domain.

In addition to LITRANSPROQA, we finetune XCOMET-XL, one of the dominant metrics for standard MT, with literary domain tasks for comparison. We evaluate metric performance on carefully selected human-annotated datasets where we find that (1) while finetuning XCOMET-XL yields only marginal gains, LITRANSPROQA delivers substantial performance gains even compared to the best SOTA in both correlation with human judgments and adequacy in rating published human translations as better than MT outputs, with gains of nearly 0.07 in correlation measured by ACC-EQ and Kendall's τ and over 15 points in adequacy; (2) incorporating translator

[1]The code and datasets are available under: https://github.com/NL2G/LiTransProQA.

votes improves LITRANSPROQA's performance, showing the value of professional input in evaluation; (3) LITRANSPROQA demonstrates substantial progress in adequacy, approaching the level of trained linguistic student annotators though still lagging behind experienced professionals, which highlights its potential and room for improvements; and (4) LITRANSPROQA shows robust performance using open-source models, demonstrating its value as an accessible, training-free metric for evaluating literary texts—particularly those requiring local processing for copyright or ethical reasons.

2 Background & related work

Dataset for literary translation Several large-scale **parallel corpora** exist in the literary domain. *BWB* (Jiang et al., 2022) and *GuoFeng* (Xu et al., 2022) contain recent Chinese-English web novels, but their unspecified human reference quality and potential use of post-edited MT (Kolb, 2023) make them unsuitable for our study. *PAR3* (Thai et al., 2022), another large-scale paragraph-level multilingual-to-English parallel corpus, includes published human translations. Although PAR3 lacks detailed metadata, its human translations can be manually verified (Zhang et al., 2025).

Existing **evaluation corpora with human judgments**, such as the WMT shared tasks, mainly focus on technical or news domains (Specia et al., 2020, 2021; Zerva et al., 2022; Blain et al., 2023)—content that differs substantially from literary texts (Voigt and Jurafsky, 2012; Matusov, 2019; Macken, 2024; Van Egdom et al., 2023). The recent WMT24 shared task, which consists of evaluation corpora with human judgments, includes literary samples across 7 language pairs from English. However, their human annotation results show MT systems outperform human translations in 4 out of 7 language pairs, likely due to suboptimal human references from less experienced or non-literary translators (Zerva et al., 2024). Three additional datasets, i.e., LITERARYTRAN (Karpinska and Iyyer, 2023), LITEVAL-CORPUS (Zhang et al., 2025), and PAR3-ANNOTATED, contain published human references with identifiable translators. While WMT24 and LITEVAL-CORPUS contain numerical scores as quality annotations, LITERARYTRAN and PAR3-ANNOTATED contain pairwise preference rankings via direct comparison.

Automatic metrics for translation evaluation MT evaluation has evolved substantially from lexical overlap metrics like BLEU (Papineni et al., 2002) to more sophisticated approaches using embeddings from pretrained models such as BERTScore (Zhang et al., 2020), Natural language inference-based MENLI (Chen and Eger, 2023), or discourse-based DiscoScore (Zhao et al., 2023). The COMET series (Guerreiro et al., 2024) mark a breakthrough in finetuning-based evaluation frameworks by effectively modeling the relationship between source and candidate translation (and reference) to better align with human judgments on translation quality. Moreover, the rise of LLMs has enabled powerful prompting metrics such as GEMBA-MQM and finetuned LLM Prometheus (Kim et al., 2024a), pushing the boundaries of general-domain MT evaluation.

QA-based metrics present another promising direction, showing potential in various NLP tasks, including summarization (Kim et al., 2024b; Fabbri et al., 2022) and translation evaluation. For MT, SimQA (Han et al., 2022), MTEQA (Krubiński et al., 2021), AskQE (Ki et al., 2025), and TREQA (Fernandes et al., 2025) all follow a similar process: generating questions, obtaining answers, and evaluating answers. While SimQA and MTEQA (both reference-based) focus on keyphrase-driven question generation, AskQE and TREQA emphasize direct question generation from candidates and source context, both relying on comparisons between LLM-generated ground-truth and candidate answers. QA metrics specifically designed for literary text remain unexplored. Though TREQA includes evaluation on literary translation, it primarily focuses on measuring meaning and information accuracy, missing the nuanced literary elements discussed above. In our work, we finetune XCOMET-XL with literary tasks and propose LITRANSPROQA, a QA-based metric that largely differs from previous attempts: (1) rather than relying on automatic question generation, we collaborate with professional literary translators and researchers to develop a specialized set of questions. This approach ensures better alignment with the expertise and practices of domain experts and practitioners; (2) instead of using open-ended generated answers, our question design simplifies responses to *yes, no, or maybe* to avoid overwhelming and confusing LLMs (Kamoi et al., 2023); and (3) our answer evaluation does not rely on a generated ground-truth whose reliability remains untested (Huang et al., 2025).

3 Dataset

We begin by introducing the datasets for metric evaluation (LITEVAL-CORPUS, LITERARYTRAN, and PAR3-ANNOTATED) and for metric finetuning (WMT24 and PAR3-UNANNOTATED). Additional details and statistical summaries are provided in Section A.1 and Table 6 (appendix).

3.1 Evaluation dataset

We use 3 human-annotated datasets with verified published human references as evaluation sets.[2]

LITEVAL-CORPUS LITEVAL-CORPUS is a benchmark dataset for examining metric performance on literary translation evaluation. It contains paragraph-level parallel data with verified high-quality human translations across four language pairs: German–English (De-En), English–German (En-De), German–Chinese (De-Zh), and English–Chinese (En-Zh), comprising over 2k paragraphs. The corpus includes outputs from 9 MT systems. Both human and MT translations are annotated with SOTA human evaluation scheme, Multidimensional Quality Metrics (MQM) (Freitag et al., 2021; Lommel et al., 2014), allowing us to examine how metrics correlate with human judgments.[3] This correlation is measured using Kendall's τ and its variant ACC-EQ (Deutsch et al., 2023) with the official packages from WMT. The dataset includes human translations, enabling examination of metric *adequacy*—how well metrics rank human translations above MT following Zhang et al. (2025). We compare human translations against three cases: (1) the top 4 MT systems (GPT-4o, DeepL, Google Translate, Qwen 2) as identified by Zhang et al. (2025) (top-level adequacy, the most challenging case); (2) all 9 MT systems (overall adequacy); and (3) all MT systems but top performers (low-level adequacy, the simplest case).

LITERARYTRAN LITERARYTRAN is a multilingual evaluation dataset containing source paragraphs from contemporary literature in English (En), German (De), French (Fr), Russian (Ru), Czech (Cs), and Japanese (Ja), with target translations in English, Japanese, and Polish (Pl). The dataset contains outputs from two MT systems: GPT-3.5 under three prompting methods and Google Translate. The dataset contains 540 direct pairwise preference annotations (1,080 source-target pairs). We compute the ACC-EQ and Kendall's τ of the metric against pairwise human preference judgments. The dataset also includes 180 human references for adequacy examination.

PAR3-ANNOTATED PAR3-ANNOTATED covers 3 language pairs (Fr-, Ru-, and De-En) from the PAR3 parallel corpus with direct preference annotations. The dataset includes translations from 2 MT systems (Google Translate and GPT-3.5). Notably, PAR3-ANNOTATED uses monolingual experts (writers and editors) for evaluation rather than translation professionals or linguists. From the dataset's 450 comparison cases (900 source-target pairs), we use 372 (744 pairs), excluding cases where evaluators rate Google Translate or GPT-3.5 outputs over human translations.

3.2 Finetuning dataset

We use 2 datasets to finetune XCOMET-XL, each corresponding to a different literary task.

PAR3-UNANNOTATED: pairwise ranking task We utilize PAR3-UNANNOTATED for a literary ranking task. The corpus comprises classic literary paragraphs with human translations and Google Translate outputs, covering translation pairs spanning both close (e.g., Fr-En) and distant language pairs (e.g., Zh-En). Its extensive size and diverse collection of human-translated literary texts make it ideal for model finetuning and domain adaptation. To expand our comparison between human and machine translations, we augment the dataset with outputs from smaller SOTA LLMs (GPT-4o-mini, TowerInstruct-13b/7b, Qwen2.5-7b, and LLaMA3-8b). We select these models for their cost efficiency and clear quality distinction from human translations, creating an effective ranking task for finetuning XCOMET-XL. These machine-generated translations are paired with their corresponding human translations, as shown in Table 7 (appendix).

WMT24: regression task WMT24 consists of evaluation corpora with human judgment using error span annotation (ESA) proposed recently by Kocmi et al. (2024). As discussed in Section 2, it

[2]For more details, see Section A.1 (appendix).

[3]Our analysis uses the complete LITEVAL-CORPUS annotation dataset, created by four student evaluators with linguistics or translation study backgrounds. These evaluators are all native speakers of the target language. While LITEVAL-CORPUS also contains annotations from professional literary translators, this professional dataset is limited in size and covers only 3 language pairs. Therefore, we primarily use the student annotations for our correlation analysis and as our main human performance benchmark.

1. Finetuning method		
finetuning task	**finetuned layers**	
	quarter (20.4%, layer 28-36)	**half (40.7%, layer 19-36)**
Reg.	1/4-WMT24	-
Bi-ranking	1/4-FREN	1/2-FREN
Bi-ranking + Reg.	1/4-FREN-WMT24	-
Multi-ranking	1/4-MULTI	1/2-MULTI
Multi-ranking + Reg.	1/4-MULTI-WMT24	-
2. Prompting method		
prompt design	**weighted by translator votes**	
	Yes	**No**
Vanilla	Vanilla_w	Vanilla
+ prompt instruction	PromptStep_w	PromptStep
+ question instruction	QuestionStep_w	QuestionStep

Table 1: Experimental setup for finetuning and prompting methods. Reg. stands for regression task using WMT24 dataset. Bi- and Multi-ranking stand for bi- and multilingual ranking tasks using PAR3-UNANNOTATED Fr-En and XX-En datasets, respectively.

contains literary samples across 7 language pairs from English with 8-13 MT systems. We exclude samples with fewer than 10 tokens, resulting in 4,500 source-target pairs. We use this human-annotated corpus as a regression finetuning task.

4 Experiment Design

We introduce our metric development methods in the following section: (1) finetuning XCOMET-XL with literary tasks and (2) developing an LLM-based QA metric reflecting the quality assessment process employed by professional literary translators. Table 1 summarizes the methodologies.

4.1 Finetuning XCOMET-XL

XCOMET-XL, built on the pretrained RoBERTa-XL model of 3b parameters, is finetuned on parallel translations with human judgments and error labels to predict quality scores given source-translation(-reference). However, since XCOMET-XL is mainly trained on non-literary texts like news, it lacks domain knowledge in literary translation evaluation (Zhang et al., 2025). To address this, we finetune it on literary datasets using (1) a ranking task and (2) a regression task.

Finetuning setup For the **ranking task**, we use triplet training loss to finetune the XCOMET-XL encoder, positioning human translations closer to their source texts than machine translations in the embedding space. For the **regression task**, we employ mean squared error loss to align quality assessment predictions with human-annotated scores. We selectively finetune specific model layers to adapt XCOMET-XL for literary domain while preserving its capabilities. For the ranking task, we test 2 layer-wise configurations: finetuning the top quarter layers (20.4%, 28–36) and top half (40.7%, 19–36). In parallel, we test 2 dataset configurations with 50k-paragraph translations: (1) a bilingual Fr-En PAR3-UNANNOTATED (the most common source language pair in PAR3) and (2) a multilingual PAR3-UNANNOTATED with various language pairs (XX-En). These configurations create four variants, shown in Table 1: quarter- & half-layer Fr-En (1/4-FREN & 1/2-FREN) and quarter- & half-layer multilingual (1/4-MULTI & 1/2-MULTI). For regression, we finetune only the top quarter layers of XCOMET-XL or of the finetuned 1/4-FREN & 1/4-MULTI using WMT24. This setup helps evaluate how dataset diversity, layer depth, and different tasks affect performance. The finetuning parameters are reported in Table 8 (appendix).

4.2 QA-based LITRANSPROQA

In addition to finetuning XCOMET-XL, we introduce an LLM-based QA metric that reflects the professional translator's quality control and assessment process. LITRANSPROQA consists of two key components: a prompting template and a question list paired with translator votes.

Template The Vanilla template, shown in Table 2, follows a simple structure: We first instruct LLM to be a professional literary translator. Next, we present a source-translation pair. Finally, we provide a list of evaluation questions that mirror the quality checks professional translators perform. The LLM answers each question with *Yes, No* or *Maybe*, which we map to scores of *1, 0, or 0.5* respectively. Each translation receives a list of scores corresponding to the list of questions. The overall translation score is calculated as either an unweighted or translator-vote-weighted mean (denoted with $_w$). We also include template variations by introducing stepwise instructions to Vanilla version or to individual questions to test whether more specific instructions could enhance performance.

Question list The development of question list consists of three key steps: (1) question collection—gathering diverse literary translation-related questions from textbooks, studies, and practical sources such as blogs and translator interviews, (2) question selection through professional literary translators' votes to identify the most critical and rel-

Vanilla	PromptStep
You are a professional literary translator with extensive experience. Now you are translating a work of great aesthetic value and cultural significance. You need to check if the translation covers all translation aspects by answering YES, NO or MAYBE to the following questions. Please be honest with your assessment and consider all aspects of translation quality.	You are a professional literary translator with extensive experience. Now you're translating a work of great aesthetic value and cultural significance. You need to check if the translation covers all translation aspects by answering YES, NO or MAYBE to the following questions. For each of the questions, 1. Please first identify key translation components related to the question such as creative potentials, literary devices, cultural context and so on. 2. After thoughtful reflection, clearly indicate your answer by responding YES, NO, or MAYBE. Be honest and precise in your assessment, ensuring each judgment is thoughtfully justified by your analysis.
shared part	
Source text: {source} **Translation**: {translation} Please answer YES, NO, or MAYBE to each of the following questions: {questions} Format your response as a JSON object where each question number is a key and the answer (YES, NO, or MAYBE) is the value. Do not include explanations, only YES, NO, or MAYBE answers. Example format: {{ '1': 'YES', '2': 'NO', '3': 'MAYBE' }} Answer:	

Table 2: LITRANSPROQA Vanilla and PromptStep templates. Shared parts show texts used in both templates.

evant assessment criteria, and (3) question selection through LLM sensitivity analysis to determine which aspects LLMs can effectively assess.

- **Step 1: Question collection.** We begin by collecting translation-related questions from literary translation research and practices. After refinement, we compile 45 questions covering 6 aspects, as shown in Table 12 and 13 (appendix): (1) Grammar & linguistics, (2) Literary devices, (3) Cultural understanding, context, & adaptation, (4) Tone & authorial voice, (5) Consistency & coherence, and (6) General equivalence. This ensures comprehensive coverage of linguistic, stylistic, and cultural features unique to the literary domain.
- **Step 2: Question selection via professional translators' votes.** We recruit professional translators to conduct a survey assessing all 45 questions collected previously. Seven professional literary translators (3 male and 4 female) with proof of experience in literary translation from English to other languages (work experience, publications, or educational background) are hired from Upwork.[4] The translators rate each question on a scale of 5 and give reasons for scores. Surveys take 12 hours in total, averaging 1.7 hours per survey. Translators receive $12 to $35 per hour based on experience, totaling $217.5. We rank questions based on their average ratings and eliminate questions scoring below 4.
- **Step 3: Question selection via LLM sensitivity.** We divide the evaluation dataset LITEVAL-CORPUS and LITERARYTRAN into development and test sets as shown in Table 6 (see Section A.3.1 for details). All test sets remain unseen during the development process. We perform a sensitivity analysis on all 45 questions. Using Vanilla template defined above, we query answers for all 45 questions. We then eliminate questions with low distinguishing power where answer distributions are heavily skewed toward one response, e.g., over 90% *yes* as shown in Table 12 and 13.

Following the selection process, each question undergoes both professional voting and LLM sensitivity checks independently. Our final list contains 25 questions. We incorporate these into the evaluation prompt by replacing question with the 25 questions. Table 12 and Table 13 (appendix) show the complete question list with translator votes and selection status. Our question selection step improves the cost- and compute-efficiency by reducing 20 questions of 451 tokens per query and boosts the metric performance by 0.05 in correlation compared to the unselected list on the development set.

Prompting setup We evaluate several prompting templates, as detailed in Table 1. The Vanilla template (shown in Table 2) employs minimal instructions. For more structured approaches, we develop templates with explicit stepwise instructions at two levels: at the entire prompt level ($\text{PromptStep}_\text{w}$ vs. PromptStep) and at the granular question level ($\text{QuestionStep}_\text{w}$ vs. QuestionStep). For QuestionStep, we craft step-by-step instructed questions, as demonstrated in Table 14 (using Vanilla template). This setup allows us to examine how stepwise instructions and translator-vote weighting influence the evaluation of literary translation quality.

5 Experiment results

We evaluate metric performance on 3 datasets: LITEVAL-CORPUS in Table 3, LITERARYTRAN in

[4] https://www.upwork.com/. For survey details, see Section A.3 including screenshots in Figures 2 & 5 and distribution of inter-annotation agreements in Figure 3 (appendix).

Test set 1: LITEVAL-CORPUS										
Metric	ACC-EQ		Kendall's τ		human > top systems (GPT-4o, GTR, DeepL, Qwen)		human > all systems		human > all excluding top systems	
Human Evaluation										
MQM	-		-		**45.3%**		43.6%		86.8%	
SOTA metrics										
GEMBA-MQM	0.534		0.561		6.1%		6.1%		**63.1%**	
COMET-KIWI	**0.552**		0.455		7.3%		6.2%		52.6%	
XCOMET-XL	0.528		0.387		17.0%		12.0%		54.5%	
XCOMET-XXL	0.540		0.400		**26.7%**		**23.9%**		61.2%	
M-Prometheus	0.445		**0.570**		16.5%		14.8%		56.7%	
TREQA-QE	0.469		0.314		12.0%		6.6%		22.0%	
XCOMET-XL Finetuned										
XCOMET-XL	0.528	Δ	0.387	Δ	17.0%	Δ	12.0%	Δ	54.5%	Δ
1/4-FREN	**0.542**	**0.014**	0.406	0.019	10.3%	-6.7%	8.1%	-3.9%	49.2%	-5.3%
1/2-FREN	0.500	-0.028	0.394	0.007	**29.4%**	**12.4%**	**22.3%**	**10.3%**	47.6%	-6.9%
1/4-MULTI	0.493	-0.035	0.348	-0.039	23.4%	6.4%	21.1%	9.2%	42.3%	-12.1%
1/2-MULTI	0.515	-0.013	0.388	0.001	20.2%	3.2%	18.5%	6.5%	50.2%	-4.2%
1/4-WMT24	0.479	-0.049	0.326	-0.061	19.9%	2.9%	12.7%	0.8%	45.0%	-9.5%
1/4-FREN-WMT24	0.442	-0.086	**0.416**	**0.029**	12.1%	-4.9%	10.4%	-1.5%	36.7%	-17.8%
1/4-MULTI-WMT24	**0.542**	**0.014**	0.397	0.010	12.9%	-4.1%	11.8%	-0.1%	**50.8%**	**-3.6%**
Avg.	0.502	-0.026	0.382	-0.005	18.3%	1.3%	15.0%	3.0%	46.0%	-8.5%
LITRANSPROQA										
BEST SOTA	0.552	Δ	0.570	Δ	26.7%	Δ	23.9%	Δ	63.1%	Δ
Vanilla †	0.606	0.054	**0.605**	**0.035**	38.7%	12.0%	37.0%	13.1%	**85.7%**	**22.5%**
Vanilla$_w$ †	**0.616**	**0.063**	**0.605**	**0.035**	**41.4%**	**14.7%**	**40.3%**	**16.4%**	**85.7%**	**22.5%**
PromptStep †	0.585	0.033	0.585	0.015	31.9%	5.3%	29.7%	5.8%	82.3%	19.2%
PromptStep$_w$ †	0.594	0.042	0.587	0.017	36.3%	9.6%	34.1%	10.2%	84.0%	20.9%
QuestionStep †	0.595	0.043	0.594	0.024	25.9%	-0.8%	22.5%	-1.4%	80.1%	17.0%
QuestionStep$_w$ †	0.600	0.048	0.597	0.027	27.0%	0.3%	23.7%	-0.2%	80.7%	17.6%
Avg.	0.599	0.047	0.595	0.026	33.5%	6.9%	31.2%	7.3%	83.1%	20.0%

Table 3: Results for LITEVAL-CORPUS. ACC-EQ and Kendall's τ measure the segment-level correlation between human MQM and metrics. Δ indicates changes in absolute value. The metric adequacy is reported as the percentage of cases where the best human translation is scored higher than outputs from (1) top systems, (2) all systems, and (3) all systems but top. † indicates significantly better ACC-EQ and Kendall's τ compared to the best SOTA on at least 3 out of 4 language pairs with permutation test at $p < 0.05$.

Table 4 and PAR3-ANNOTATED in Table 11 (appendix). We analyze metrics' correlation with human judgments and their adequacy, i.e., the ability to rank human translations over MT outputs. We include SOTA metrics as baselines: finetuned metrics (XCOMET-XL, XCOMET-XXL, and COMET-KIWI), LLM-based prompting metric GEMBA-MQM, recent QA metric TREQA, and recent multilingual LLM M-Prometheus (14b), trained for general evaluation purposes (Pombal et al., 2025). We compare literary-finetuned XCOMET-XL against the original XCOMET-XL and compare LITRANSPROQA against the best available metric performance. To ensure comparability, all LLM-based prompting metrics use the same base model (GPT-4o-mini). We also show LITRANSPROQA performance on other base models in Table 9 and 11 (appendix).

5.1 Marginal gains for finetuning XCOMET-XL

Finetuning XCOMET-XL on literary tasks offers modest improvements in some settings, though the benefits are inconsistent. For LITEVAL-CORPUS, finetuned metrics like 1/4-FREN and 1/2-FREN show mild performance gains over the base XCOMET-XL with 1/4-FREN's ACC-EQ from 0.528 to 0.542 and Kendall's τ from 0.387 to 0.406. However, these improvements do not consistently translate to adequacy gains. Other finetuned metrics like 1/4-WMT24 and 1/2-MULTI show mixed results. Resource-intensive configurations even degrade performance—1/4-MULTI shows a 12.1-point drop in *Human > all but top systems*. Joint finetuning with both tasks barely improves the performance. On average, finetuned methods show mixed results with correlation slightly below XCOMET-XL and minor adequacy gains with 1-3 points for 2 testing cases *Human > top & Human > all*. For LITERARYTRAN and PAR3-ANNOTATED, the impact of finetuning is more negative. 1/4-MULTI-WMT24 leads among the finetuned versions for LITERARYTRAN, with ACC-EQ reaching 0.643 and Kendall's τ 0.292. However, its adequacy drops from 18.5% to 14.4%. While 1/2-FREN and 1/4-MULTI show modest improvements in adequacy score by 2 points to 20.5%, other finetuned variants show barely any improvements.

Our analysis suggests that finetuning yields marginal gains with bilingual datasets or shallow

layers being more effective. This echoes the finding from Shi et al. (2024) where focused tuning on shallow layers achieves better alignment with literary translation goals by concentrating on essential features instead of noise from inconsistent literary styles in multilingual datasets.

Test set 2: LITERARYTRAN						
Metric	ACC-EQ		Kendall's τ		human > MT (GPT-3.5 & GTR)	
SOTA metrics						
GEMBA-MQM	0.419		**0.269**		11.8%	
COMET-KIWI	0.586		0.172		9.6%	
XCOMET-XL	**0.603**		0.207		18.5%	
XCOMET-XXL	0.586		0.171		**26.0%**	
M-Prometheus	0.223		0.124		21.8%	
TREQA-QE	0.519		0.038		14.4%	
XCOMET-XL Finetuned						
XCOMET-XL	0.603	Δ	0.207	Δ	18.5%	Δ
1/4-FREN	0.637	0.034	0.275	0.068	8.3%	-10.3%
1/2-FREN	0.583	-0.021	0.183	-0.024	**20.5%**	**2.0%**
1/4-MULTI	0.574	-0.029	0.148	-0.059	20.4%	1.9%
1/2-MULTI	0.567	-0.036	0.188	-0.019	12.4%	-6.1%
1/4-WMT24	0.576	-0.027	0.153	-0.054	15.7%	-2.8%
1/4-FREN-WMT24	0.474	-0.130	**0.312**	**0.105**	8.9%	-9.6%
1/4-MULTI-WMT24	**0.643**	**0.040**	0.292	0.086	14.4%	-4.1%
Avg.	0.579	-0.024	0.222	0.015	14.4%	-4.1%
LITRANSPROQA						
BEST SOTA	0.603	Δ	0.269	Δ	26.0%	Δ
Vanilla	0.519	-0.085	0.303	0.035	38.9%	12.9%
Vanilla$_w$	**0.570**	**-0.033**	**0.304**	**0.035**	**42.4%**	**16.4%**
PromptStep	0.466	-0.138	0.258	-0.010	36.3%	10.3%
PromptStep$_w$	0.510	-0.093	0.278	0.009	39.1%	13.1%
QuestionStep	0.528	-0.075	0.290	0.021	33.4%	7.4%
QuestionStep$_w$	0.554	-0.049	0.291	0.023	35.5%	9.5%
Avg.	0.525	-0.079	0.287	0.019	37.6%	11.6%

Table 4: Results for LITERARYTRAN. ACC-EQ and Kendall's τ measure the segment-level correlation between human judgments (pairwise preference) and metrics. Δ indicates changes in absolute value. The metric adequacy is reported as the percentage of cases where human translation is scored higher than the outputs from GPT-3.5 and Google Translate (GTR).

5.2 LITRANSPROQA: strong performance gain in correlation and adequacy

LITRANSPROQA demonstrates substantial and consistent improvements in both correlation with human judgments and adequacy (see example in Tables 15). For LITEVAL-CORPUS, LITRANSPROQA outperforms the best SOTA baselines by a large margin. All variants of LITRANSPROQA show significantly better correlation than the best SOTA as shown in Table 3. Vanilla$_w$ shows the strongest performance with a score of 0.616 for ACC-EQ and 0.605 for Kendall's τ. It also demonstrates the highest gains for all adequacy cases, with *Human > top systems* at 41.4%, *Human > all systems* at 40.3%, *Human > all systems but top* at 85.7%—a 14.7, 16.4, and 22.5 point increase over the best SOTA. LITRANSPROQA outperforms the best finetuned XCOMET-XL, showing an increase of nearly 0.2 in Kendall's τ, 18-point and 35-point in adequacy for *Human > all systems* and *Human > all but top systems*. For LITERARYTRAN in Table 4, LITRANSPROQA continues to excel, delivering the strongest adequacy results while maintaining top-level correlation. Vanilla$_w$ achieves a Kendall's τ of 0.304 and an adequacy of 42.4%, marking a 16.4-point gain over the SOTA XCOMET-XXL. For PAR3-ANNOTATED, LITRANSPROQA again outperforms best SOTA metrics by 0.06 in Kendall's τ and nearly 18 points in adequacy on average. Also worth noting is that step instructions, both PromptStep and QuestionStep, perform worse than Vanilla setting. This may suggest that detailed, literary-specific instructions, particularly QuestionStep, could dilute the effectiveness of LLMs' judgments in complex tasks compared to simpler instructions (see example in Tables 16).

Translator votes improve metric performance. Translator vote-weighted variants improve correlation and adequacy compared to unweighted scores, as shown in Table 10, Figure 4, and significance analysis in Section A.3.8 (appendix). For LITEVAL-CORPUS, the weighted versions achieve a better score by nearly 0.01 in ACC-EQ and over 4 points in adequacy compared to their unweighted versions. For LITERARYTRAN, weighting with translator votes is a key differentiator, with all evaluation cases showing improvements: 0.05 in ACC-EQ, 0.02 in Kendall's τ, and 3.5 points in adequacy. PAR3-ANNOTATED shows similar gains in ACC-EQ and adequacy. Overall, weighted scores yield average improvements of 0.02 in ACC-EQ and 2 points in adequacy. These results demonstrate the value of incorporating professional translators' perspectives into the evaluation.

LITRANSPROQA achieves high adequacy, approaching student annotator performance. For LITEVAL-CORPUS, LITRANSPROQA's adequacy results closely reach student annotator performance across all three comparison cases, with the gap narrowing to less than 4 points for all cases.[5] This marks a substantial improvement over existing SOTAs, which show gaps of 18.6 points for *human > top systems*, 19.7 points for *human > all systems*, and 23.7 points for *human > all but top systems*. While LITRANSPROQA has not yet matched

[5]Our adequacy results differ from those reported in Zhang et al. (2025) for two reasons: (1) we partition the dataset into test and development sets and report results only on the test set, while also making these splits available for reproducibility; and (2) Table 3 in Zhang et al. (2025) uses only a subset of the full data (see their footnote 12 regarding the BWS samples).

the professional human translators' performance of nearly 90% for *human > top systems* (Zhang et al., 2025), it demonstrates noticeable progress toward human-level evaluation capabilities comparable to trained linguistics students, highlighting its potential and room for future improvements for literary MT evaluation.

Can LITRANSPROQA perform well using other base models? To evaluate LITRANSPROQA's compatibility on other base models, we implement the PromptStep on open-source models of different sizes. Table 9 and 11 (appendix) indicate that open-source models LLaMA3.3-70b and Qwen2.5-32b show competitive results for all datasets. For LITEVAL-CORPUS and LITERARYTRAN, open-source models achieve even better results than GPT-4o-mini with LLaMA3.3-70b surpassing GPT-4o-mini in adequacy by 9 points on *human > top systems* and Qwen2.5-32b exceeding by over 0.1 ACC-EQ and 0.07 Kendall's τ for LITERARYTRAN. LLaMA-70b's reasoning variant DeepSeek-distilled shows unsatisfactory performance. Additionally, LLaMA's previous larger version 405b lags behind the smaller 70b model. Our results indicate that LITRANSPROQA maintains competitiveness when applied to open-source models, demonstrating its generalizability.

5.3 Ablation study for LITRANSPROQA

Test set 2: LITERARYTRAN				
Metric/Asp.	**#Qs**	ACC-EQ	Kendall's τ	human > LLM (GPT-3.5 & GTR)
Vanilla	25	0.519	0.303	38.9%
ablation setting 1: scores with one aspect alone (+)				
GL	3	0.171	0.177	11.0%
LD	4	0.231	0.199	21.3%
CCA	5	0.285	0.210	24.5%
TA	6	0.327	0.282	21.8%
CO	2	0.033	0.042	0.0%
GE	5	0.345	0.276	26.9%
ablation setting 2: scores excluding one aspect (-)				
GL	22	0.512	0.304	36.9%
LD	21	0.504	0.311	39.7%
CCA	20	0.483	0.301	36.3%
TA	19	0.428	0.260	36.3%
CO	23	0.516	0.304	38.3%
GE	20	0.488	0.295	32.8%

Table 5: Ablation results per question aspects on LITERARYTRAN. Asp. stands for aspects: (1) GL (Grammar & linguistics); (2) LD (Literary devices); (3) CCA (Cultural understanding, context, & adaptation); (4) TA (Tone & authorial voice); (5) CO (Consistency & coherence); and (6) GE (General equivalence). Vanilla represents LITRANSPROQA score using the full selected question set. #Q indicates the number of questions.

We conduct an ablation study to investigate the contribution of each literary aspect of LITRANSPROQA. Table 5 reports the results on LITERARYTRAN dataset under two ablation settings: (1) scoring with one aspect alone; (2) scoring excluding one particular aspect. We draw three key conclusions. First, the contribution of *any single aspect in isolation* (setting 1) is substantially weaker than the full question set (Vanilla). Even the strongest single aspect, GE (General equivalence), lags behind Vanilla by more than 0.17 in ACC-EQ and 12 points in adequacy. Second, *removing any single aspect* (setting 2) does not yield improvements over Vanilla across all three evaluation perspectives. While excluding LD (Literary devices) leads to slight gains in Kendall's τ and adequacy, it incurs a non-negligible drop in ACC-EQ by 0.015, while the removal of other aspects consistently produces declines. This demonstrates that LITRANSPROQA benefits from the complementary contributions of all six aspects. Third, some dimensions are more impactful than others. For example, in setting 2, excluding TA (Tone & authorial voice) causes the largest degradation in ACC-EQ by 0.09, while omitting GE results in the largest decline in adequacy, highlighting their central roles. In contrast, removing CO (Consistency & coherence) has comparatively limited effect.

6 Conclusion

In this paper, we introduce LITRANSPROQA, a novel LLM-based QA metric specifically designed for evaluating literary translations. LITRANSPROQA addresses critical shortcomings of existing approaches and achieves substantial performance gains over both finetuned XCOMET-XL and current SOTA metrics. Our results show improvements across all 3 test sets, with increases of 0.04 in Kendall's τ, 0.06 in ACC-EQ, and over 15 points in adequacy. LITRANSPROQA's strong performance with open-source models enhances its accessibility and reduces dependency on proprietary technology, while enabling broader adoption and careful consideration of ethical issues in evaluating copyrighted and culturally sensitive texts.

LITRANSPROQA is designed to better account for aspects of translation quality emphasized in professional human translation, such as creative subtleties and cultural nuances often overlooked by literal and homogenized MT outputs. In doing so, LITRANSPROQA can help mitigate the existing bias toward literal translations and support the recalibration of LLMs toward more human-like literary translation.

Limitations

While we try to cover as many languages as possible, our evaluation remains predominantly focused on high- and medium-resource language pairs due to the limited availability of suitable evaluation datasets. This underscores the necessity of developing comprehensive evaluation datasets within the literary domain, particularly targeting low-resource language pairs. Future work can also explore the effect of genre and style variation on the metric. While our current analysis mainly relies on the opinions of literary translators, incorporating feedback from broader audiences could provide additional insight.

Additionally, we currently evaluate translations at the paragraph level, which may miss subtle literary elements that span larger narrative sections. A key limitation is the absence of evaluation datasets containing extended narrative units like consecutive chapters or complete works. Future research could expand the evaluation dataset to include wider contexts.

Our experiments show that question-level instruction templates perform less effectively than the two simpler configurations. While we hypothesize this stems from LLMs' lack of specialized literary knowledge, our current study does not analyze the specific nature or extent of this knowledge gap. Further investigation into LLMs' specialized knowledge of literary and creative tasks remains a valuable research direction.

Finally, although LITRANSPROQA performs substantially well on both closed- and open-source models, it could benefit further from domain-specific finetuning of LLMs (Rafailov et al., 2023). Future research can delve into even smaller models using this method to improve efficiency.

Ethical Considerations

We utilize open-source datasets for evaluation and finetuning. For datasets containing copyrighted content, we use them following fair use principles for research and academic purposes.

For human evaluation, we obtained informed consent from all participating professional translators. Their contributions are disclosed anonymously and do not include any protected demographic or personal information.

Potential risks Potential risks of LITRANSPROQA include reinforcing biases toward high- and medium-resource languages, while effects remain unknown for underrepresented low-resource languages in test data. To ensure equitable distribution of LITRANSPROQA's benefits, these risks demand careful attention through responsible deployment and more comprehensive dataset coverage.

Licensing and intended use Our implementation builds on components from GEMBA-MQM (CC-BY-SA-4.0), M-Prometheus (Apache-2.0), and COMET (Apache-2.0). Because GEMBA-MQM is licensed under CC-BY-SA-4.0, which includes a ShareAlike clause, our released code is distributed under the same license. By contrast, Apache-2.0 components are permissive and compatible with redistribution under CC-BY-SA-4.0. Our use of these artifacts, as well as the LLMs utilized in this work, complies with their respective licenses and use policies. This release is intended solely for research and evaluation purposes in literary translation.

PII in data We rely exclusively on existing publicly available datasets. We have not performed independent, systematic checks for personally identifiable information (PII) within these datasets, as we consider this the responsibility of the original dataset creators. As our work involves literary excerpts, some texts may contain offensive language. In literary contexts, the presence of such language is not inherently undesirable, as it reflects the source material and, in some cases, is directly relevant to the research question.

Packages We use the mt-metrics-eval package (version 2, commit 6d4b0bb), with a modification to meta_info.py to include our test dataset meta information. The following important dependencies are used: scipy 1.10.1, seaborn 0.13.2, transformers 4.50.0, and openai 1.68.2 (for API calls).

Acknowledgements

We thank the anonymous reviewers for their feedback, which greatly improved the work. We appreciate the professional input from all professional 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

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

Ralph A Alexander. 1990. A note on averaging correlations. *Bulletin of the Psychonomic Society*, 28(4):335–336.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

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.

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.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–34.

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. *arXiv preprint arXiv:2406.15267*.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12914–12929, Singapore. Association for Computational Linguistics.

Steffen Eger, Yong Cao, Jennifer D'Souza, Andreas Geiger, Christian Greisinger, Stephanie Gross, Yufang Hou, Brigitte Krenn, Anne Lauscher, Yizhi Li, Chenghua Lin, Nafise Sadat Moosavi, Wei Zhao, and Tristan Miller. 2025. Transforming science with large language models: A survey on ai-assisted scientific discovery, experimentation, content generation, and evaluation. *ArXiv*, abs/2502.05151.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Patrick Fernandes, Sweta Agrawal, Emmanouil Zaranis, André FT Martins, and Graham Neubig. 2025. Do llms understand your translations? evaluating paragraph-level mt with question answering. *arXiv preprint arXiv:2504.07583*.

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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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

HyoJung Han, Marine Carpuat, and Jordan Boyd-Graber. 2022. SimQA: Detecting simultaneous MT errors through word-by-word question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5598–5616, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Tianbo Ji, Yvette Graham, Gareth Jones, Chenyang Lyu, and Qun Liu. 2022. Achieving reliable human assessment of open-domain dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6416–6437, Dublin, Ireland. Association for Computational Linguistics.

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.

Ryo Kamoi, Tanya Goyal, and Greg Durrett. 2023. Shortcomings of question answering based factuality frameworks for error localization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 132–146, Dubrovnik, Croatia. 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.

Dayeon Ki, Kevin Duh, and Marine Carpuat. 2025. Askqe: Question answering as automatic evaluation for machine translation. *arXiv preprint arXiv:2504.11582*.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024a. Prometheus 2: An open source language model specialized in evaluating other language models. *arXiv preprint arXiv:2405.01535*.

Yekyung Kim, Yapei Chang, Marzena Karpinska, Aparna Garimella, Varun Manjunatha, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024b. Fables: Evaluating faithfulness and content selection in book-length summarization. *Preprint*, arXiv:2404.01261.

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.

Mateusz Krubiński, Erfan Ghadery, Marie-Francine Moens, and Pavel Pecina. 2021. Just ask! evaluating machine translation by asking and answering questions. In *Proceedings of the Sixth Conference on Machine Translation*, pages 495–506, Online. 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.

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.

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.

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*.

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

José Pombal, Dongkeun Yoon, Patrick Fernandes, Ian Wu, Seungone Kim, Ricardo Rei, Graham Neubig, and André F. T. Martins. 2025. M-prometheus: A suite of open multilingual llm judges. *Preprint*, arXiv:2504.04953.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

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.

Jiang-Xin Shi, Tong Wei, Zhi Zhou, Jie-Jing Shao, Xin-Yan Han, and Yu-Feng Li. 2024. Long-tail learning with foundation model: heavy fine-tuning hurts. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André F. T. Martins. 2021. Findings of the WMT

2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 684–725, Online. Association for Computational Linguistics.

Lucia Specia, Zhenhao Li, Juan Pino, Vishrav Chaudhary, Francisco Guzmán, Graham Neubig, Nadir Durrani, Yonatan Belinkov, Philipp Koehn, Hassan Sajjad, Paul Michel, and Xian Li. 2020. Findings of the WMT 2020 shared task on machine translation robustness. In *Proceedings of the Fifth Conference on Machine Translation*, pages 76–91, Online. Association for Computational Linguistics.

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.

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.

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.

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, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Chrysoula Zerva, Frederic Blain, José G. C. De Souza, Diptesh Kanojia, Sourabh Deoghare, Nuno M. Guerreiro, Giuseppe Attanasio, Ricardo Rei, Constantin Orasan, Matteo Negri, Marco Turchi, Rajen Chatterjee, Pushpak Bhattacharyya, Markus Freitag, and André Martins. 2024. Findings of the quality estimation shared task at WMT 2024: Are LLMs closing the gap in QE? In *Proceedings of the Ninth Conference on Machine Translation*, pages 82–109, Miami, Florida, USA. Association for Computational Linguistics.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. Findings of the WMT 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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

Ran Zhang, Wei Zhao, and Steffen Eger. 2025. How good are LLMs for literary translation, really? literary translation evaluation with humans and LLMs. In *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, Albuquerque, New Mexico. Association for Computational Linguistics.

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. 2023. 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.

A Appendix

A.1 Datasets

Table 6 summarizes the statistics of both evaluation and finetuning datasets.

LITEVAL-CORPUS combines sources from contemporary and classic literary texts, including translations from 9 MT systems: the GPT series (GPT-4o), commercial models (Google Translate and DeepL), popular smaller LLMs (LLaMA3, Qwen 2, Gemini, TowerInstruct), and previous SOTA systems (M2M, NLLB). Link to the official WMT package: `https://github.com/google-research/mt-metrics-eval`. ACC-EQ

Dataset	Use case	Language Pair	Book	Size		#MT systems	Annotation Type
				test	dev		
LITEVAL-CORPUS	Test-Dev	De-En, En-De, De-Zh, En-Zh	Contemporary/Classics	1996	70	9	MQM score
LITERARYTRAN		Src: En, De, Fr, Ru, Cs, Ja Tgt: En, Ja, Pl	Contemporary	1095	165	2	Preference
PAR3-ANNOTATED	Test	Fr-En, Ru-En, De-En	Classics	744		2	Preference
PAR3-UNANNOTATED	Finetuning (ranking)	Src: 18 languages Tgt: En	Classics	50k		6	-
WMT24	Finetuning (regression)	En-Cs, Hi, Is, Ja, Ru, Uk, Zh	Contemporary	4600		8-13	ESA score

Table 6: Summary statistics of evaluation and finetuning datasets. Size indicates the number of source-target pairs of paragraphs. The ESA score, i.e., error span annotation (Kocmi et al., 2024) is an updated version of MQM (Multidimensional Quality Metrics). Preference refers to direct preference comparison between pairs of translation versions, without assigning numerical scores.

is a variant of Kendall's τ that is recently proposed and implemented by the WMT shared task (Deutsch et al., 2023). This metric evaluates pairwise accuracy while accounting for tie calibration. We report both scores to ensure broader comparability.

LITERARYTRAN Three prompting methods are examined for translating paragraphs using GPT-3.5: translating sentence-by-sentence without context (SENT), translating sentence-by-sentence with full paragraph context (PARA_SENT), and directly translating a whole paragraph (PARA). The dataset includes direct pairwise preference annotations comparing SENT vs. PARA, PARA_SENT vs. PARA, and Google Translate vs. PARA.

Adequacy For the adequacy measure of cases with multiple human translations, we consider the version rated highest in human evaluations. This approach is reasonable since older translations may be less appealing to modern annotators due to changes in language and style over time. Regarding the adequacy performance of annotators (human level), only LITEVAL-CORPUS provides annotations scoring both human translation and MT outputs. In contrast, LITERARYTRAN and PAR3 only contain human annotations of MT outputs, making it impossible to determine annotators' adequacy performance.

Correlation metric The choice of correlation metrics (ACC-EQ and Kendall's τ) is motivated by two reasons. First, both have been recently adopted in the WMT shared task, one of the most renowned venues for MT evaluation, ensuring comparability with prior work. Second, the LITERARYTRAN and PAR3 datasets contain only pairwise comparison data. In this setting, Kendall's τ and ACC-EQ are particularly well-suited, as the number of concordant and discordant pairs is well-defined even under binary labels. By contrast, Pearson and Spearman correlations require continuous or ordinal scores, making them less appropriate for pairwise judgments.

A.2 Finetuned XCOMET-XL

A.2.1 Example of finetuning dataset

Table 7 shows an example of a PAR3-UNANNOTATED paired dataset for the ranking task.

Pair	De-En	Model
Source	Am wiederholtesten aber fragte der treue Diener, fast so oft er Ottilien sah, nach der Rückkunft des Herrn und nach dem Termin derselben.	Human
Positive	But almost every time the faithful servant saw Ottilie what he most repeatedly asked about was the master's return and when that was going to happen.	Human
Negative	Most frequently, however, the faithful servant asked, almost every time he saw Ottilie, about the return of his master and the date of that return.	GPT-4o-mini

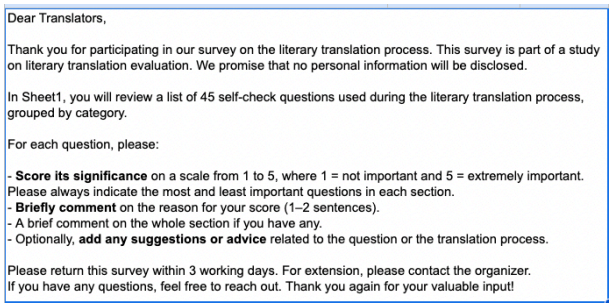
Table 7: Example of PAR3 paired dataset for the ranking task.

A.2.2 Finetuning details

parameter	value
batch_size	8
encoder_learning_rate	2.00e-5
encoder_weight_decay	0.01
max_length	512
gradient_accumulation_steps	4
early_stopping	true
epoch	3
loss	Triplet loss

Table 8: Finetuning parameters

Table 8 shows the finetuning parameters for XCOMET-XL. We use the PyTorch implementation for both losses. See https://pytorch.org/docs/stable/generated/torch.nn.TripletMarginLoss.html for the formula of triplet loss.



Dear Translators,

Thank you for participating in our survey on the literary translation process. This survey is part of a study on literary translation evaluation. We promise that no personal information will be disclosed.

In Sheet1, you will review a list of 45 self-check questions used during the literary translation process, grouped by category.

For each question, please:

- **Score its significance** on a scale from 1 to 5, where 1 = not important and 5 = extremely important. Please always indicate the most and least important questions in each section.
- **Briefly comment** on the reason for your score (1–2 sentences).
- A brief comment on the whole section if you have any.
- Optionally, **add any suggestions or advice** related to the question or the translation process.

Please return this survey within 3 working days. For extension, please contact the organizer.
If you have any questions, feel free to reach out. Thank you again for your valuable input!

Figure 2: Screenshot of the instruction page for the survey.

A.2.3 Discussion on XCOMET-XL

The limited gains from finetuned XCOMET-XL likely stem from multiple factors: (1) Domain mismatch, XCOMET is initially trained on sentence-level non-literary data, while our fine-tuning involved paragraph-level literary domain; (2) Limited training data — regression datasets are rather small and both regression/ranking datasets fail to cover all languages in the test sets; (3) Data quality — genre variation and uncertain translation quality in some datasets, as also noted in WMT 2024, may limit effective adaptation; (4) Literary texts are inherently difficult to learn, especially at paragraph level.

A.3 Details for LITRANSPROQA

A.3.1 Development set details

We sample 1-2 source paragraphs per language pair from each dataset. Our development set contains 70 source-target pairs (3.4%) on 4 language pairs from LITEVAL-CORPUS and 165 (13.1%) on 18 language pairs from LITERARYTRAN. We make sure that all test sets remain unseen during the development process. Our metric performance is reported on test data only.

A.3.2 Survey details

The instruction Figure 2 shows the screenshot of the instruction page for the survey.

The full question list Figure 5 demonstrates the screenshot of the survey page with the complete question list.

A.3.3 The selected question list with translator votes

Tables 12 and 13 present the complete question list with their status, mean translator voting, and reasons for exclusion from the final list. The status indicates one of four outcomes: (1) S for selected questions, (2) R-GI for rejected questions due to general insensitivity in GPT-4o-mini results (where one answer dominated across translations, showing poor general quality discrimination), (3) R-HI for rejected questions due to human insensitivity (where a high percentage of no/maybe responses for human translations indicated poor discrimination of human translation), and (4) R-TV for questions rejected due to low translator vote scores. We highlight the selected questions in **light green**.

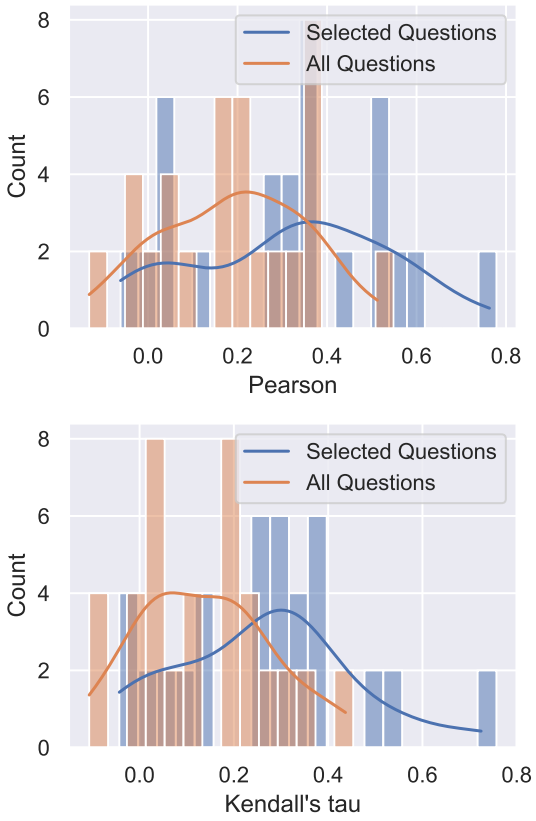


Figure 3: Pairwise inter-annotator agreement distribution for question importance ratings among the seven translators.

A.3.4 Annotation agreement among translators

As literary translation is a nuanced and inherently subjective task, some disagreement among expert annotators is expected. To obtain a more reliable consensus, we employ seven independent annotators and report pairwise agreement using Pearson and Kendall coefficients. We report the distribution of inter-annotator agreement among the seven translators (all 45 questions vs. selected 25 questions) in Figure 3. Rather than presenting averaged values, we show the full distribution because correlation measures are non-additive and their averages

LITEVAL-CORPUS					
Base	reason	ACC-EQ	Kendall's τ	human > top systems	human > all systems
GPT-4o-mini$_w$	No	0.594	**0.587**	36.3%	**84.0%**
LLaMa3.1-405b$_w$	No	0.537	0.506	36.5%	70.0%
LLaMa3.3-70b$_w$	No	0.552	0.537	**45.0%**	76.0%
DK LLaMa-70b$_w$	Yes	0.497	0.461	12.6%	40.9%
Qwen2.5-32b$_w$	No	**0.602**	0.584	31.3%	82.9%
LITERARYTRAN					
Base	reason	ACC-EQ	Kendall's τ	human > MT (GPT-3.5 & GTR)	
GPT-4o-mini$_w$	No	0.510	0.278	**39.1%**	
LLaMa3.1-405b$_w$	No	0.385	0.154	22.0%	
LLaMa3.3-70b$_w$	No	0.444	0.208	36.8%	
DK LLaMa-70b$_w$	Yes	0.405	0.069	15.0%	
Qwen2.5-32b$_w$	No	**0.616**	**0.346**	35.7%	

Table 9: LITRANSPROQA performance on open-source base models using PromptStep template weighted by translator votes. Reason denotes whether the model has reasoning capabilities. DK stands for DeepSeek distilled version.

LITEVAL-CORPUS					
template	ACC-EQ	Kendall's τ	human > top systems	human > all systems	human > all but top
Vanilla	0.009	0.000	2.7%	3.3%	0.0%
PromptStep	0.009	0.002	4.4%	4.4%	1.7%
QuestionStep	0.004	0.002	1.1%	1.1%	0.6%
LITERARYTRAN					
template	ACC-EQ	Kendall's τ	human > MT (GPT-3.5 & GTR)		
Vanilla	0.052	0.001	3.5%		
PromptStep	0.044	0.019	2.8%		
QuestionStep	0.026	0.001	2.1%		
Test set 3: PAR3 annotated					
template	ACC-EQ	Kendall's τ	human > MT (GPT-3.5 & GTR)		
Vanilla	0.001	-0.013	3.3%		
PromptStep	-0.002	-0.020	0.0%		
Avg.	0.018	-0.001	2.1%		

Table 10: Impact of translator-weighted scores for LITRANSPROQA across 3 evaluation sets. The table shows LITRANSPROQA's absolute performance gains (Δ) when using weighted versus non-weighted scoring. Avg. represents the mean across all datasets.

may be misleading and hard to interpret (Alexander, 1990; Ji et al., 2022).

The distributions in Figure 3 demonstrate moderate inter-annotator correlations, with centers around 0.3-0.4 on both measures for selected questions. This shows reasonable consistency considering the task's subjective nature. Additionally, selected questions show higher agreement compared to all questions, indicating that the selected subset contains questions with more reliable cross-translator judgments.

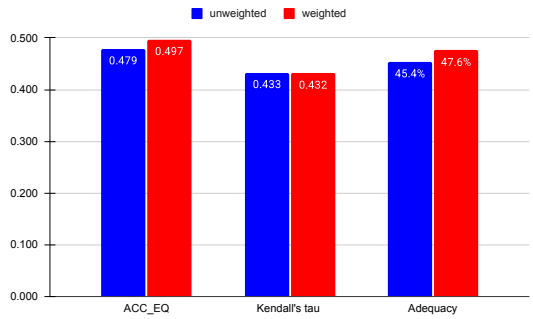


Figure 4: Comparison of weighted vs. unweighted LITRANSPROQA scores: average results across all 3 test datasets.

A.3.5 The selected question list with QuestionStep questions

Table 14 shows the selected question list for Vanilla and PromptStep in comparison to step-instructed questions for QuestionStep.

A.3.6 LITRANSPROQA results with other base models

Table 9 shows the results of PromptStep template using various base models.

A.3.7 LITRANSPROQA results for test set 3: PAR3-ANNOTATED

Table 11 presents the results for PAR3-ANNOTATED test set. Using GPT-4o-mini as the base model, LITRANSPROQA consistently outperforms SOTA metrics, showing improvements of up to 0.08 in Kendall's τ and a substantial 17.8-point increase in adequacy. When using the open-source LLaMA3.3-70b model, LITRANSPROQA performs marginally below SOTA in correlation but still exceeds the best adequacy SOTA from XCOMET-XXL by more than 10 points.

A.3.8 Impact of translator votes

Table 10 shows improvements in translator-votes weighted scores compared to non-weighted scores. For LITERARYTRAN and LITEVAL-CORPUS, improvements occur across almost all cases. PAR3-ANNOTATED also shows overall enhancements, except for the Kendall's τ metric.

We evaluate the significance of weighted vs. unweighted scores. For the 12 LITEVAL-CORPUS test cases (4 language pairs $\times$ 3 LITRANSPROQA prompts), 66.7% of Kendall's τ results and 83.3% of ACC-EQ results are significant at $p < 0.05$. For LITERARYTRAN (all language pairs combined $\times$

Test set 3: PAR3-ANNOTATED						
Metric	ACC-EQ		Kendall's τ		human > MT (GPT-3.5 & GTR)	
SOTA metrics						
GEMBA-MQM	0.132		-0.014		14.7%	
COMET-KIWI	0.231		0.155		16.7%	
XCOMET-XL	0.250		0.250		15.3%	
XCOMET-XXL	**0.269**		**0.344**		**29.3%**	
M-Prometheus	0.119		0.148		18.0%	
XCOMET-XL Finetuned						
XCOMET-XL	0.250	Δ	0.250	Δ	15.3%	Δ
1/4-FREN	0.204	-0.046	0.021	-0.229	8.0%	-7.3%
1/4-MULTI	**0.245**	**-0.005**	**0.223**	**-0.027**	18.7%	3.3%
WMT24	0.172	-0.078	-0.140	-0.110	**22.7%**	**7.3%**
1/4-FREN +WMT24	0.179	-0.071	0.101	-0.149	11.3%	-4.0%
1/4-MULTI + WMT24	0.224	-0.026	0.115	-0.135	16.0%	0.7%
Avg.	0.205	-0.045	0.064	-0.130	14.4%	-0.9%
LITRANSPROQA						
BEST SOTA	0.269	Δ	0.344	Δ	29.3%	Δ
Base model: GPT-4o-mini						
Vanilla	0.266	-0.003	0.404	0.060	46.7%	17.3%
Vanilla$_w$	**0.268**	**-0.001**	0.390	0.047	**50.0%**	**20.7%**
PromptStep	0.265	-0.005	**0.423**	**0.080**	46.0%	16.7%
PromptStep$_w$	0.263	-0.007	0.403	0.059	46.0%	16.7%
Avg.	0.265	-0.004	0.405	0.061	47.2%	17.8%
Base model: LLaMa3.3-70b						
Vanilla	0.219	-0.050	0.291	-0.053	40.0%	10.7%
Vanilla$_w$	0.214	-0.055	0.264	-0.080	40.0%	10.7%
PromptStep	0.193	-0.077	0.270	-0.074	39.3%	10.0%
PromptStep$_w$	0.195	-0.074	0.283	-0.061	38.7%	9.3%
Avg.	0.205	-0.064	0.277	-0.067	39.5%	10.2%

Table 11: Results for test set 3: PAR3-ANNOTATED. ACC-EQ and Kendall's τ measure the segment-level correlation between human judgments (pairwise preference) and metrics. The metric adequacy is reported as the percentage of cases where human translation is scored higher than the outputs from GPT-3.5 and Google Translate (GTR).

3 prompts), 66.7% of Kendall's τ results and 100% of ACC-EQ results reach significance.

A.4 Qualitative example

Table 15 and 16 demonstrate qualitative examples from LITERARYTRAN with scores from SOTA metrics and LITRANSPROQA variants. Table 15 shows that SOTA metrics tend to underestimate human translation consistently, while LITRANSPROQA does not. Table 16 demonstrates a failure case of LITRANSPROQA producing mixed results. The PromptStep template successfully ranks human translations higher than MT outputs, while the other two templates (Vanilla and QuestionStep) fail to make this distinction. Further analysis shows that both failure templates show difficulty with questions about cultural context and cultural translation, often giving these categories lower scores than PromptStep. This may occur because LITRANSPROQA misses subtle stylistic elements when using overly simple Vanilla template, yet struggles with the nuanced understanding when templates become too complex (QuestionStep). However, human evaluators also report that the differences in translation quality here are only marginal, suggesting that reliably detecting such subtle distinctions requires greater sensitivity that LITRANSPROQA should further improve upon.

Grammar and linguistics		Your score	Your comments	Cooments on section
1	**Am I on guard for "false friends" or misleading cognates between the source and target languages that could distort the meaning?**			
2	Have I checked for words that have multiple meanings in the source language and ensured the correct interpretation was chosen in translation?			
3	Have I considered the secondary meanings or feelings associated with particular words or phrases and does my translation reflect them?			
4	**Do my word choices convey the same connotations in the target language? For example, while 'Latin school' in a German context implies a prestigious institution, a direct translation into Chinese may require a term like 'better school' to convey the same idea.**			
5	Have I adjusted measurement units, currencies, numerical expressions, and titles appropriately?			
6	Does the translation accurately reflect the core meaning of the original text without unnecessary additions or omissions?			
7	**Does the translated text flow naturally in the target language, following the conventional writing style and grammar?**			
8	Do I account for grammatical differences between the source and target languages (tense, aspect, word order, etc.) so that the same information and nuances come through?			
9	Does my translation avoid literal interpretations that could distort the intended message?			
Literary devices				
1	Have I preserved literary devices (irony, symbolism, foreshadowing, etc.) so that the author's artistic intentions are carried into the translation (or provided subtle cues if direct carryover isn't possible)?			
2	Have I avoid the case where direct translation of idioms, honorifics, or set phrases creates unintended confusion?			
3	Have I adjusted idioms, humor or figures of speech to maintain the original's intent and effect within the cultural context of the target language? (e.g., "raining cats and dogs" → "llueve a cántaros" in Spanish, or "下傾盆大雨" in Chinese)			
4	When culture-specific terms (food, clothing, idioms, holidays etc.), practices, or references in the source are unfamiliar to the target readers, have I adapted these in a way that a target reader can understand and appreciate (by using the original term, adding a brief explanation, or substituting with a culturally analogous reference/an equivalent concept etc.)?			
5	Does the translation preserve the same meaning and implications of figurative language, such as metaphors or similes?			
6	Am I considering the context and unspoken implications so that the translation reflects the source text's pragmatic meaning (tone, irony, implicature), not just the dictionary definitions of individual words?			
Cultural understanding, context, and adeptation				
1	If the source text uses a reference that may not be well known in the target culture, have I replaced it with a more familiar example? (e.g., "Harvard" → "Tsinghua" when translating into Chinese to convey "prestigious university" or changing "Super Bowl" to "a major sports event" in cultures unfamiliar with American football)			
2	For references to historical or social events, have I considered whether they carry the same significance in the target culture? (e.g., mentioning "July 4th" in the US might need explanation in another culture)			
3	Am I handling the cultural differences in how certain emotions, gestures, or expressions are described in an clear and understandable way?			
4	Have I used the widely accepted translation in the target culture if certain quoted materials or references (like the title of a book, a poem, a song, or a famous saying) already have one?			
5	Am I avoiding excessive domestication that erases the source culture's identity?			
6	Am I balancing foreignizing strategies with the need for clarity so that I don't drift into exoticizing the source culture for the target reader?			
7	Have I considered the source culture's context (historical period, locale, customs) well enough to translate cultural references accurately, knowing that each language group has its own culturally specific features?			
8	Am I avoiding imposing my own culture's perspective or biases on the text?			
9	Does the translation evoke the same cultural mood as the original?			
10	Have I handled slang or dialect correctly? E.g., by finding target-language equivalents that carry a similar flavor and social meaning, or a neutral rendition to avoid confusion.			
Tone, authorial voice				
1	Does my translation maintain the same tone and level of formality for each character or narrative voice as the original?			
2	Does the use of honorifics or pronouns match the intended tone of the source text?			
3	Is the tone (formal, informal, neutral, colloquial, etc.) consistent in my translation?			
4	Does the narrative voice in my translation match the original novel's narrative voice?			
5	Am I preserving the author's unique voice or style in the source text (e.g., terse and minimalist, or elaborate and lyrical) in my translation?			
6	Have I preserved the narrative point of view and any shifts in perspective exactly as in the original?			
7	Is my translation strategy (literal, free, adaptive, etc.) aligned with the literary style of the source text and the expectations of its genre?			
8	Does the translation convey the author's descriptive imagery vividly and accurately?			
Consistensy and coherence				
1	Have I used appropriate discourse markers or transitions to maintain the logical flow of ideas?			
2	Have I maintained consistency in terminology, character names, slang, dialect and other key details throughout the text to avoid confusing the reader?			
3	Have I kept track of plot details to ensure nothing got lost or altered in translation that would cause inconsistencies?			
General equivalence				
1	Does the translation convey the same meaning and intent as the source text, even if the sentence structure is different?			
2	Do I balance well between communicative translation (focusing on the target reader's understanding and response) and semantic translation (focusing on the exact contextual meaning of the original)?			
3	Have I maintained the same level of ambiguity or precision as the source text?			
4	Does the translation elicit a similar emotional response in the target reader as the original does in its readers?			
5	Would a reader of the translation sense the same narrator persona or character as in the original?			
6	Am I conveying the author's intent behind each passage – the subtext or purpose of why they wrote it – through my translation choices?			
7	Will my target audience respond to this translation the same way the source audience responded to the original?			
8	Is the translation phrased in the most natural way for the target language, achieving the "closest natural equivalent" of the source message?			
9	Have I considered equivalence at multiple levels, at the word level, sentence level, textual level (cohesion), and pragmatic level (overall effect) – to ensure the translation works as a coherent whole?			
Your suggestions				

Figure 5: Screenshot of survey page.

ID	Asp.	Status	Questions	Score	Note
35	CO	S	Have I maintained consistency in terminology, character names, slang, dialect and other key details throughout the text to avoid confusing the reader?	5.00	
2	GL	R-GI	Have I checked for words that have multiple meanings in the source language and ensured the correct interpretation was chosen in translation?	4.86	general insensitivity: 91% yes
9	GL	R-GI	Does my translation avoid literal interpretations that could distort the intended message?	4.86	general insensitivity: 95% yes
10	LD	R-GI	Have I preserved literary devices (irony, symbolism, foreshadowing, etc.) so that the author's artistic intentions are carried into the translation (or provided subtle cues if direct carryover isn't possible)?	4.86	general insensitivity: 91% maybe
23	CCA	R-GI	Am I avoiding imposing my own culture's perspective or biases on the text?	4.86	general insensitivity: 97% yes
37	GE	R-GI	Does the translation convey the same meaning and intent as the source text, even if the sentence structure is different?	4.86	general insensitivity: 97% yes
1	GL	S	Am I on guard for "false friends" or misleading cognates between the source and target languages that could distort the meaning?	4.86	
29	TA	S	Does the narrative voice in my translation match the original novel's narrative voice?	4.86	
7	GL	S	Does the translated text flow naturally in the target language, following the conventional writing style and grammar?	4.71	
15	LD	S	Am I considering the context and unspoken implications so that the translation reflects the source text's pragmatic meaning (tone, irony, implicature), not just the dictionary definitions of individual words?	4.71	
20	CCA	S	Am I avoiding excessive domestication that erases the source culture's identity?	4.71	
21	CCA	S	Am I balancing foreignizing strategies with the need for clarity so that I don't drift into exoticizing the source culture for the target reader?	4.71	
26	TA	S	Does my translation maintain the same tone and level of formality for each character or narrative voice as the original?	4.71	
30	TA	S	Am I preserving the author's unique voice or style in the source text (e.g., terse and minimalist, or elaborate and lyrical) in my translation?	4.71	
42	GE	S	Am I conveying the author's intent behind each passage – the subtext or purpose of why they wrote it – through my translation choices?	4.71	
6	GL	R-GI	Does the translation accurately reflect the core meaning of the original text without unnecessary additions or omissions?	4.57	general insensitivity: 96% yes
40	GE	R-GI	Does the translation elicit a similar emotional response in the target reader as the original does in its readers?	4.57	general insensitivity: 91% maybe
12	LD	S	Have I adjusted idioms, humor or figures of speech to maintain the original's intent and effect within the cultural context of the target language? (e.g., "raining cats and dogs" → "llueve a cántaros" in Spanish, or "下倾盆大" in Chinese)	4.57	
13	LD	S	When culture-specific terms (food, clothing, idioms, holidays etc.), practices, or references in the source are unfamiliar to the target readers, have I adapted these in a way that a target reader can understand and appreciate (by using the original term, adding a brief explanation, or substituting with a culturally analogous reference/an equivalent concept etc.)?	4.57	
14	LD	S	Does the translation preserve the same meaning and implications of figurative language, such as metaphors or similes?	4.57	
22	CCA	S	Have I considered the source culture's context (historical period, locale, customs) well enough to translate cultural references accurately, knowing that each language group has its own culturally specific features?	4.57	
33	TA	S	Does the translation convey the author's descriptive imagery vividly and accurately?	4.57	
36	CO	S	Have I kept track of plot details to ensure nothing got lost or altered in translation that would cause inconsistencies?	4.43	
3	GL	R-GI	Have I considered the secondary meanings or feelings associated with particular words or phrases and does my translation reflect them?	4.43	general insensitivity: 91% maybe
8	GL	R-GI	Do I account for grammatical differences between the source and target languages (tense, aspect, word order, etc.) so that the same information and nuances come through?	4.43	general insensitivity: 94% yes

Table 12: Question list ranked by translator votes (top 25). ID indicate the original ID in the survey. Asp. stands for the 6 aspects: (1) GL (Grammar & linguistics); (2) LD (Literary devices); (3) CCA (Cultural understanding, context, & adaptation); (4) TA (Tone & authorial voice); (5) CO (Consistency & coherence); and (6) GE (General equivalence). Status indicates one of four outcomes: (1) S for selected questions, (2) R-GI for rejected questions due to general insensitivity in GPT-4o-mini results (where one answer dominated across translations, showing poor general quality discrimination), (3) R-HI for rejected questions due to human insensitivity (where a high percentage of no/maybe responses for human translations indicated poor discrimination of human translation), and (4) R-TV for questions rejected due to low translator vote scores.

ID	Asp.	Status	Questions	Score	Note
11	LD	R-GI	Have I avoid the case where direct translation of idioms, honorifics, or set phrases creates unintended confusion?	4.43	general insensitivity: 96% yes
4	GL	S	Do my word choices convey the same connotations in the target language? For example, while 'Latin school' in a German context implies a prestigious institution, a direct translation into Chinese may require a term like 'better school' to convey the same idea.	4.29	
18	CCA	S	Am I handling the cultural differences in how certain emotions, gestures, or expressions are described in an clear and understandable way?	4.29	
28	TA	S	Is the tone (formal, informal, neutral, colloquial, etc.) consistent in my translation?	4.29	
31	TA	S	Have I preserved the narrative point of view and any shifts in perspective exactly as in the original?	4.29	
39	GE	S	Have I maintained the same level of ambiguity or precision as the source text?	4.29	
41	GE	S	Would a reader of the translation sense the same narrator persona or character as in the original?	4.29	
38	GE	R-GI	Do I balance well between communicative translation (focusing on the target reader's understanding and response) and semantic translation (focusing on the exact contextual meaning of the original)?	4.29	general insensitivity: 91% maybe
44	GE	S	Is the translation phrased in the most natural way for the target language, achieving the "closest natural equivalent" of the source message?	4.14	
17	CCA	R-HI	For references to historical or social events, have I considered whether they carry the same significance in the target culture? (e.g., mentioning "July 4th" in the US might need explanation in another culture)	4.14	human insensitivity: 100% no/maybe
19	CCA	R-HI	Have I used the widely accepted translation in the target culture if certain quoted materials or references (like the title of a book, a poem, a song, or a famous saying) already have one?	4.14	human insensitivity: 22% no/maybe
27	TA	R-HI	Does the use of honorifics or pronouns match the intended tone of the source text?	4.14	human insensitivity: 46% no/maybe
24	CCA	S	Does the translation evoke the same cultural mood as the original?	4.00	
34	CO	R-GI	Have I used appropriate discourse markers or transitions to maintain the logical flow of ideas?	4.00	general insensitivity: 94% yes
45	GE	S	Have I considered equivalence at multiple levels, at the word level, sentence level, textual level (cohesion), and pragmatic level (overall effect) – to ensure the translation works as a coherent whole?	4.00	
32	TA	R-TV	Is my translation strategy (literal, free, adaptive, etc.) aligned with the literary style of the source text and the expectations of its genre?	3.86	translator weight < 4
25	CCA	R-TV	Have I handled slang or dialect correctly? E.g., by finding target-language equivalents that carry a similar flavor and social meaning, or a neutral rendition to avoid confusion.	3.71	translator weight < 4
5	GL	R-HI	Have I adjusted measurement units, currencies, numerical expressions, and titles appropriately?	3.43	human insensitivity: 33% no/maybe
16	CCA	R-HI	If the source text uses a reference that may not be well known in the target culture, have I replaced it with a more familiar example? (e.g., "Harvard" → "Tsinghua" when translating into Chinese to convey "prestigious university" or changing "Super Bowl" to "a major sports event" in cultures unfamiliar with American football)	3.00	human insensitivity: 100% no/maybe
43	GE	R-GI	Will my target audience respond to this translation the same way the source audience responded to the original?	3.00	general insensitivity: 97% maybe

Table 13: Question list ranked by translator votes (26-45). ID indicate the original ID in the survey. Asp. stands for the 6 aspects: (1) GL (Grammar & linguistics); (2) LD (Literary devices); (3) CCA (Cultural understanding, context, & adaptation); (4) TA (Tone & authorial voice); (5) CO (Consistency & coherence); and (6) GE (General equivalence). Status indicates one of four outcomes: (1) S for selected questions, (2) R-GI for rejected questions due to general insensitivity in GPT-4o-mini results (where one answer dominated across translations, showing poor general quality discrimination), (3) R-HI for rejected questions due to human insensitivity (where a high percentage of no/maybe responses for human translations indicated poor discrimination of human translation), and (4) R-TV for questions rejected due to low translator vote scores.

Index	ID	Asp.	Questions for Vanilla and PromptStep	Step-instructed questions for QuestionStep
1	1	GL	Am I on guard for “false friends” or misleading cognates between the source and target languages that could distort the meaning?	Identify any words that look similar in both languages. Am I on guard for “false friends” or misleading cognates between the source and target languages that could distort the meaning?
2	4	GL	Do my word choices convey the same connotations in the target language? For example, while ’Latin school’ in a German context implies a prestigious institution, a direct translation into Chinese may require a term like ’better school’ to convey the same idea.	First, identify all proper names, noun phrases, and cultural/historical references in the source. Then answer: do my word choices convey the same connotations in the target language? Example: While "Latin school" in a German context implies a prestigious institution, a direct translation into Chinese may require a term like "better school" to convey the same idea.
3	7	GL	Does the translated text flow naturally in the target language, following the conventional writing style and grammar?	Read the translation aloud. Check grammar and syntax. Does the translated text flow naturally in the target language, following conventional writing style and grammar?
4	12	LD	Have I adjusted idioms, humor or figures of speech to maintain the original’s intent and effect within the cultural context of the target language? (e.g., "raining cats and dogs" → "llueve a cántaros" in Spanish, or “下倾盆大雨” in Chinese)	Identify all idioms, jokes, metaphors, unusual expressions, and figures of speech in the source. For each, think about its intended effect or meaning. Have I adjusted them to maintain the original’s intent and effect within the cultural context of the target language? (e.g., "raining cats and dogs" → "llueve a cántaros" in Spanish, or “下倾盆大雨” in Chinese)
5	13	LD	When culture-specific terms (food, clothing, idioms, holidays etc.), practices, or references in the source are unfamiliar to the target readers, have I adapted these in a way that a target reader can understand and appreciate (by using the original term, adding a brief explanation, or substituting with a culturally analogous reference/an equivalent concept etc.)?	First, identify all culture-specific terms (food, clothing, idioms, holidays, etc.), practices, or cultural/historical references in the source. For terms that are unfamiliar to the target readers, answer: have I adapted them in a way that a target reader can understand and appreciate (by using the original term, adding a brief explanation, or substituting with a culturally analogous reference or equivalent concept)?
6	14	LD	Does the translation preserve the same meaning and implications of figurative language, such as metaphors or similes?	First, identify figurative language such as metaphors, similes, comparisons, or original images in the source. Analyze their meaning and emotional resonance. Does the translation preserve the same meaning and implications?
7	15	LD	Am I considering the context and unspoken implications so that the translation reflects the source text’s pragmatic meaning (tone, irony, implicature), not just the dictionary definitions of individual words?	Reflect on tone, irony, suggestion, or hidden implications in the source. Does the translation reflect the source text’s pragmatic meaning/function, not just the dictionary definitions of individual words?
8	18	CCA	Am I handling the cultural differences in how certain emotions, gestures, or expressions are described in an clear and understandable way?	Am I handling cultural differences in how certain emotions, gestures, or expressions are described in a clear and understandable way?
9	20	CCA	Am I avoiding excessive domestication that erases the source culture’s identity?	Revisit all localized terms. Am I avoiding excessive domestication that erases the source culture’s identity?
10	21	CCA	Am I balancing foreignizing strategies with the need for clarity so that I don’t drift into exoticizing the source culture for the target reader?	Review translations that feel “foreign” or unusual. Am I balancing foreignizing strategies with the need for clarity so that I don’t drift into exoticizing the source culture for the target reader?
11	22	CCA	Have I considered the source culture’s context (historical period, locale, customs) well enough to translate cultural references accurately, knowing that each language group has its own culturally specific features?	First, identify all cultural/historical references from the source. Think about the source culture’s context (historical period, locale, customs). Have I considered the context well enough to translate these references accurately, knowing that each language group has its own culturally specific features?
12	24	CCA	Does the translation evoke the same cultural mood as the original?	First, read the source text alone and summarize the cultural mood. Then read the translation and summarize the cultural mood. Determine: does the translation evoke the same cultural mood as the original?
13	26	TA	Does my translation maintain the same tone and level of formality for each character or narrative voice as the original?	Does my translation maintain the same tone and level of formality for each character or narrative voice as in the original?
14	28	TA	Is the tone (formal, informal, neutral, colloquial, etc.) consistent in my translation?	Is the tone (formal, informal, neutral, colloquial, etc.) consistent in my translation?
15	29	TA	Does the narrative voice in my translation match the original novel’s narrative voice?	Does the narrative voice in my translation match the original novel’s narrative voice?
16	30	TA	Am I preserving the author’s unique voice or style in the source text (e.g., terse and minimalist, or elaborate and lyrical) in my translation?	Am I preserving the author’s unique voice or style in the source text (e.g., terse and minimalist, or elaborate and lyrical) in my translation?
17	31	TA	Have I preserved the narrative point of view and any shifts in perspective exactly as in the original?	Note the narrative perspective and any changes (st person, 3rd limited, omniscient). Have I preserved the narrative point of view and any shifts in perspective exactly as in the original?
18	33	TA	Does the translation convey the author’s descriptive imagery vividly and accurately?	Highlight vivid descriptions and sensory language. Does the translation convey the author’s descriptive imagery vividly and accurately?
19	35	CO	Have I maintained consistency in terminology, character names, slang, dialect and other key details throughout the text to avoid confusing the reader?	Track key terms, character names, invented words, slang, or dialects. Have I maintained consistency throughout the text to avoid confusing the reader?
20	36	CO	Have I kept track of plot details to ensure nothing got lost or altered in translation that would cause inconsistencies?	Outline plot developments in the source. Have I kept track of plot details to ensure nothing was lost or altered in translation that would cause inconsistencies?
21	39	GE	Have I maintained the same level of ambiguity or precision as the source text?	Have I maintained the same level of ambiguity or precision as the source text?
22	41	GE	Would a reader of the translation sense the same narrator persona or character as in the original?	Understand who the narrator is and their role. Would a reader of the translation sense the same narrator persona or character as in the original?
23	42	GE	Am I conveying the author’s intent behind each passage – the subtext or purpose of why they wrote it – through my translation choices?	For each passage, ask: why did the author write it this way? Then answer: am I conveying the author’s intent behind each passage—the subtext or purpose of why they wrote it—through my translation choices?
24	44	GE	Is the translation phrased in the most natural way for the target language, achieving the “closest natural equivalent” of the source message?	Read the translation as a native would. Is it phrased in the most natural way for the target language, achieving the “closest natural equivalent” of the source message?
25	45	GE	Have I considered equivalence at multiple levels, at the word level, sentence level, textual level (cohesion), and pragmatic level (overall effect) – to ensure the translation works as a coherent whole?	Have I considered equivalence at multiple levels—word level, sentence level, textual level (cohesion), and pragmatic level (overall effect)—to ensure the translation works as a coherent whole?

Table 14: Selected questions for Vanilla and PromptStep vs. step-instructed questions for QuestionStep.

Source	Target	Model	GEM	KIWI	XCOMET		M-Pro	TRE	LITRANSPROQA					
									Vanilla		PromptStep		QuestionStep	
					XL	XXL			-	w	-	w	-	w
Bis auf Selmas Schwägerin Elsbeth waren die Leute im Dorf meistens nicht abergläubisch. Sie machten unbekümmert all das, was man bei Aberglauben nicht machen darf: Sie saßen gelassen unter Wanduhren, obwohl man bei Aberglauben daran sterben kann, sie schliefen mit dem Kopf zur Tür hin, obwohl das bei Aberglauben bedeutet, dass man durch genau die Tür bald mit den Füßen zuerst hinausgetragen wird. Sie hängten zwischen Weihnachten und Neujahr Wäsche auf, was, wie Elsbeth warnte, bei Aberglauben einem Suizid oder einer Beihilfe zum Mord gleichkommt. Sie erschraken nicht, wenn nachts das Käuzchen rief, wenn ein Pferd im Stall stark schwitzte, wenn ein Hund nachts jaulte, mit gesenktem Kopf.	Except for Selma's sister-in-law Elsbeth, people in the village were for the most part not superstitious. They blithely broke all of superstition's rules: They sat calmly under wall clocks even though the superstitious can die from it. They slept with their heads toward the door, though superstition claimed they'd soon be carried out that very door feetfirst. They hung laundry to dry between Christmas and New Year's, which, according to superstition, Elsbeth would remind them, amounts to suicide or accessory to murder. They were not frightened when owls hooted, when a horse in the stall broke into a heavy sweat, when a dog howled in the night with its head lowered.	Human	-7.0	0.71	0.67	0.54	**6**	54.1	**4.4**	**3.99**	**4.7**	**4.24**	**4.8**	**4.33**
	Apart from Selma's sister-in-law Elsbeth, most people in the village were not superstitious. They did all the things that superstition forbids: they sat calmly under wall clocks, even though superstition says that it can kill you; they slept with their heads towards the door, even though superstition says that it means you will be carried out of the door feet first; they hung out laundry between Christmas and New Year, which, as Elsbeth warned, was akin to suicide or aiding and abetting murder according to superstition. They were not scared when the screech owl called at night, when a horse sweated heavily in the stable, or when a dog howled with its head down at night.	GPT-3.5 PARA	**-6.0**	0.75	0.79	**0.66**	**6**	55.2	**4.4**	3.97	4.4	3.96	4.1	3.71
	Except for Selma's sister-in-law Elsbeth, the people in the village were mostly not superstitious. They did all the things that one should not do in superstition: They sat calmly under wall clocks, even though in superstition one can die from that, they slept with their head towards the door, even though in superstition that means that one will soon be carried out of the door feet first. They hung out laundry between Christmas and New Year's, which, as Elsbeth warned, in superstition is equivalent to suicide or aiding in murder. They did not shudder when the owl called at night, when a horse was sweating heavily in the stable, or when a dog howled with its head down.	GPT-3.5 PARA_SENT	-7.0	**0.77**	**0.82**	**0.66**	**6**	**59.4**	3.7	3.36	4.0	3.63	3.8	3.44
	Apart from Selma's sister-in-law Elsbeth, most people in the village were not superstitious. They carelessly did all the things that superstition forbids: They sat calmly beneath wall clocks, even though superstition says that it can kill you, and they slept with their heads towards the door, even though superstition says that you will soon be carried out of it feet first. They hung laundry between Christmas and New Year's, which, as Elsbeth warned, was considered superstitious to be equivalent to suicide or aiding in murder. They were not startled when the owl called out at night, when a horse sweated heavily in the stable, or when a dog howled with its head bowed at night.	GPT-3.5 SENT	**-6.0**	0.75	0.79	0.60	4	51.7	4.3	3.89	4.0	3.62	3.9	3.52

Table 15: Qualitative example from LITERARYTRAN. We use abbreviated forms of metrics: GEM (GEMBA-MQM), KIWI (COMET-KIWI), XL (XCOMET-XL), XXL (XCOMET-XXL), M-Pro (M-Prometheus), and TRE (TREQA-QE). W indicates the weighted version of LITRANSPROQA. We highlight the highest score of individual metrics in **blue** and the lowest in red.

Source	Target	Model	GEM	KIWI	XCOMET		M-Pro	TRE	LITRANSPROQA					
									Vanilla		PromptStep		QuestionStep	
					XL	XXL			-	w	-	w	-	w
Das stimmte. Der Hund war schneematschfarben, er war verwaschen grau und zottelig wie ein ausschließlicher Irischer Wolfhund ohne etwas anderes drin. Sein Körper war noch klein, aber seine Pfoten waren groß wie Bärentatzen, und wir wussten alle, was das bedeutete. Selma stand immer noch erhoben vor der Küchenbank. Sie schaute lange auf den Hund. Dann sah sie meinen Vater an, als sei er ein Geschenkideengeschäft.	It was true. The dog was the color of slush. It was watery gray and shaggy as only purebred Irish wolfhounds can be. Its body was still small, but its paws were as big as a bear's, and we all knew what that meant. Selma was still standing in front of the kitchen bench. She looked at the dog for a long time. Then she looked at my father as if he were a gift shop.	human	-7	0.69	**0.76**	**0.79**	**4**	53.9	3.2	3.26	**4.1**	**3.71**	3.3	2.98
	That was true. The dog was snow-muddy-colored, he was washed-out grey and shaggy like an exclusive Irish Wolfhound without anything else mixed in. His body was still small, but his paws were as big as bear's claws, and we all knew what that meant. Selma was still standing tall in front of the kitchen bench. She looked at the dog for a long time. Then she looked at my father as if he was a gift shop.	GPT-3.5 PARA	-37	0.76	0.75	0.71	2	65.3	3.4	3.08	3.6	3.27	**3.5**	**3.17**
	That was true. The dog was snow-mushroom colored, he was washed-out grey and shaggy like a pure Irish Wolfhound with nothing else mixed in. His body was still small, but his paws were big like bear paws, and we all knew what that meant. Selma still stood erect in front of the kitchen bench. She looked at the dog for a long time. Then she looked at my father as if he was a gift shop.	GPT-3.5 PARA_SENT	-37	**0.77**	0.75	**0.79**	**4**	**67.9**	3.4	3.08	3.6	3.27	3.2	2.88
	That's true. The dog was snow-muddy in color, he was washed-out gray and shaggy like an exclusive Irish wolfhound with nothing else mixed in. His body was still small, but his paws were as big as a bear's, and we all knew what that meant. Selma was still standing in front of the kitchen counter. She looked at the dog for a long time. Then she looked at my father as if he were a gift shop.	GPT-3.5 SENT	**-7**	0.76	0.75	0.69	2	60.5	**3.6**	**3.27**	3.6	3.27	3.4	3.07

Table 16: Qualitative failure example from LITERARYTRAN. We use abbreviated forms of metrics: GEM (GEMBA-MQM), KIWI (COMET-KIWI), XL (XCOMET-XL), XXL (XCOMET-XXL), M-Pro (M-Prometheus), and TRE (TREQA-QE). W indicates the weighted version of LITRANSPROQA. We highlight the highest score of individual metrics in **blue** and the lowest in red.