Learning from Many Voices: Literary MT Using Multi-Reference Human and Synthetic Data

Si Wu
Northeastern University
siwu@ccs.neu.edu

John Wieting
Google DeepMind
jwieting@google.com

David A. Smith
Northeastern University
dasmith@ccs.neu.edu

Abstract

Unlike many other texts, literary works are often translated multiple times. We investigate strategies for leveraging these multi-reference datasets to improve literary machine translation. We propose a filtering framework based on semantic similarity to identify source texts whose references display meaningful variation while remaining faithful. We find that fine-tuning with medium to high semantic similarity data substantially outperforms low semantic similarity data. Moreover, using medium and high semantic similarity data achieves comparable or better performance than using the full unfiltered data.

Synthetic translations generated by LLMs are economical and convenient alternatives to human expert translations; however, we find fine-tuning on human expert translations outperforms fine-tuning on synthetically augmented data in automatic metrics and human evaluations, demonstrating the indispensable value of human expert translations for fine-tuning literary machine translation models.

1 Introduction

Literary translation presents unique challenges for machine translation (MT). Literary works are deeply embedded in the source language and culture; therefore, it can be difficult to reproduce the same meaning while preserving cultural context and stylistic nuance in another language. Moreover, literary translations are meant to be read. Fully explicating all interpretations would require substantial annotation or commentary (e.g., in footnotes)—what Nabokov memorably called "translations with copious footnotes reaching up like skyscrapers" (Nabokov, 1955)—and reduce readability.

As such, literary translators must balance multiple aspects of a source text when translating: meaning, form, rhyme, style, and cultural context (Jakobson, 1987; Lefevere, 1992; Hofstadter, 1997). Some emphasize lexical and syntactic imitation (formal equivalence), while others sacrifice form for comprehensibility (dynamic/functional equivalence) (Nida, 1964). Read together, different translations of the same source text collectively provide diverse interpretations and a richer understanding of the original work. For example, in Figure 1, the three translations vary in style and word choice, yet together they reveal not only the meaning and literary style of the French text, but also some of its etymological associations: trans-

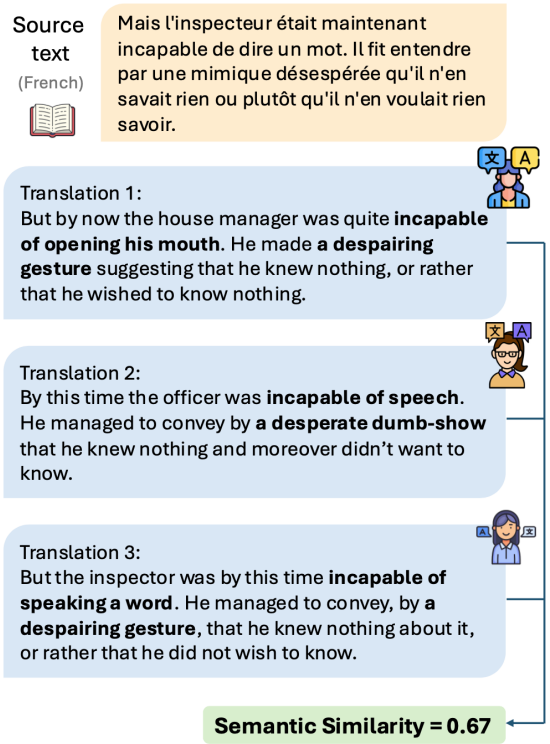

Figure 1: An example of semantic variation across human translations of the same source text. We analyze a paragraph from *Le Fantôme de l'Opéra* (*The Phantom of the Opera*) by Gaston Leroux. The original French paragraph was translated into English by three different literary translators. Using our proposed scoring system SEMSIM, the average pairwise semantic similarity across the English translations is 0.67.

Proceedings of the Eleventh Conference on Machine Translation, pages 818–841
October 28-29, 2026 ©2026 Association for Computational Linguistics

lation 2's "dumb-show" draws out the theatrical associations of the French word *mimique*, which is related to mime and gesture-based expression.

Prior work has explored methods to synthetically generate additional references for a source text via back-translation or paraphrasing (Madnani et al., 2008; Zheng et al., 2018; Khayrallah et al., 2020). However, these works are conducted within non-literary domains such as news, and the text is usually at sentence-level. News and literary texts differ fundamentally in purpose and style: news is generally written to be factual, direct, and unambiguous, whereas literary writing is sometimes intentionally open to interpretation, with authors using devices such as metaphor, imagery, symbolism, rhythm, and rhyme. Due to the very nature of literary text, scholars and readers in comparative literature and translation studies often compare multiple translations of a source text to deepen their understanding of the original and of how its meanings are rendered across languages (Rose, 1997; Stenberg, 2018; Szymanska, 2026).

Moreover, well-known literary works often have multiple translations, particularly in high-resource language pairs. It is reasonable to leverage the wealth of expert-produced translations that are already available. However, the question is how to effectively utilize them when training or fine-tuning an MT system. Additionally, it's worth investigating the performance gap between models trained on expert translations and those trained on synthetic paraphrases. Synthetic alternatives are more economical and may be particularly beneficial to low-resource languages where parallel data are scarce (Khayrallah et al., 2020; de Gibert et al., 2025).

In this paper, we investigate how to leverage multiple references for literary translation. Specifically, we propose a filtering approach based on semantic similarity scores: selecting source texts whose reference translations collectively preserve semantic fidelity while exhibiting meaningful lexical and syntactic variation, as measured by pairwise semantic similarity scores.

This paper examines three hypotheses:

H1. Different ranges of semantic similarity will lead to different performance: High and medium SEMSIM lead to substantially better translation than low SEMSIM. (Section 6)

H2. A filtered dataset is better for fine-tuning literary MT than an unfiltered dataset: Combining high and medium SEMSIM data achieves MT performance comparable or better than unfiltered dataset. (Section 7)

H3. A synthetic-augmented multi-reference dataset will perform comparably to an expert-translated dataset: Synthetic augmented dataset does not improve MT performance as reliably as human expert-translated multi-reference data and can underperform an unaugmented single-reference dataset, depending on the fine-tuning models used for augmentation. (Section 8)

For all the hypotheses, we compare performance measured by automatic metrics and evaluations from human translators. We also present zero-shot MT results from commercial and open-weight instruction-tuned models in Appendix D, and provide example outputs in Appendix I.

2 Related work

Literary machine translation is known to be a challenging task (Taivalkoski-Shilov, 2019; Brusasco, 2022; Corpas Pastor and Noriega-Santiáñez, 2024). Unlike news and academic writing, literary language often deploys devices such as imagery, symbolism, metaphor, rhythm, and rhyme. Literary text is deeply embedded in its linguistic, cultural and historical contexts. Polysemy is one example: when multiple meanings of a word in the source text are intended to be evoked, but in the target language there is no word with the same one-to-many mapping, there is no direct translation. In this situation, a translator will have to decide what to prioritize in a translation. As such, to deepen their understanding of the linguistic and cultural nuance of the source text, scholars in comparative literature and translation studies often compare multiple translations (Rose, 1997; Stenberg, 2018; Szymanska, 2026).

In machine translation, the idea of using multiple target translations of a source text to improve understanding of the source-target language pair has long been explored in various ways. Much of this work is through synthetically creating new target texts: from using pivot-based paraphrasers (Callison-Burch et al., 2006; Madnani et al., 2008), to back-translation in statistical and neural machine translation (Sennrich et al., 2016; Poncelas et al., 2018; Edunov et al., 2018). More recently, LLMs have been used for synthetic data generation in many different tasks (Long et al., 2024), and LLMs' ability to generate high-quality text is particularly

beneficial in low-resource settings (Scalvini et al., 2025; de Gibert et al., 2025). Although LLMs have improved the quality and complexity of paraphrase generation, they still make mistakes, particularly for outputs longer than a single sentence (Jayawardena and Yapa, 2024a,b; Su et al., 2022; Horvitz et al., 2024; Mishra et al., 2024). Most prior work focuses on non-literary domains and primarily at sentence-level.

Many have explored employing post-editing to improve literary MT (Besacier and Schwartz, 2015; Thai et al., 2022; Castaldo et al., 2025). While some have looked into fine-tuning methods for the literary MT task (Li et al., 2024; Mikelenić et al., 2025), we are unaware of work on how to optimally fine-tune MT model for the literary domain, particularly with multiple translations per source text. Moreover, no prior work has directly compared the effects of fine-tuning on human data vs. synthetic data for literary MT.

Finally, our filtering criterion differs from the general notion of "noise" in MT: we filter based on semantic similarity of target texts of the same source text, not the mismatches or semantic divergence between source and target. Nevertheless, prior work on data filtering shows that training data quality matters for MT performance (Carpuat et al., 2017; Khayrallah and Koehn, 2018; Briakou and Carpuat, 2021; Steingrímsson et al., 2023).

3 Experimental setup

We describe the models we use for MT and automatic metrics we use in Section 3.2, and the details of pre-processing steps in Appendix A. We discuss each hypothesis and its results using the multi-reference datasets produced from the pre-processing step in Sections 6 to 8.

3.1 Literary Translation Data

Popular works of world literature are often translated into English by multiple translators, and these translations can differ stylistically while remaining faithful to the original text. The Par3 dataset aligns multiple English translations of each non-English book at the paragraph level: each source paragraph has at least two distinct English translations produced by expert literary translators (Thai et al., 2022). Translating literature at paragraph-level instead of sentence-level incorporates more discourse features and has been proven to improve literary MT (Voigt and Jurafsky, 2012). Although Par3 covers 19 source languages, we conduct experiments only on the French-English and Russian-English language pairs, which account for 48% and 32% of the total source paragraphs, respectively. Although both language pairs exhibit similar trends in automatic metrics, our discussion in Sections 6 to 8 will mainly focus on French-English, since we only conducted human evaluations on this language pair.

3.2 Models and Automatic Metrics

We fine-tune three models for the literary machine translation task: mT5-Large (Xue et al., 2021), Gemma3-12B-PT (Team et al., 2025), and OpusMT (Tiedemann et al., 2023; Tiedemann and Thottingal, 2020), and we use BLEU as the early-stopping criterion for all models. Detailed model descriptions, training details, and the descriptions of automatic metrics are in Appendix B.

4 Measuring Semantic Similarity Among References of the Same Source Text

Intuitively, reference translations of the same source text should be semantically similar but not identical: if two references' semantic similarity is too high, the meaning of the two would be nearly identical, and the additional reference would provide little new learning signal; conversely, if the semantic similarity is too low, the discrepancy may indicate noise, such as source–target misalignment or overly divergent translator interpretations.

We measure semantic similarity between reference translations of the same source text using SEMSIM. This similarity score allows us to quantify the variations among different references in lexical choice or syntactic structure while expressing the same underlying meaning.

SEMSIM utilizes a semantic similarity scoring model PARAGRAM-SP (Wieting et al., 2022). PARAGRAM-SP score ranges from -1 to 1, where a lower score indicates low semantic similarity, and a higher score indicates high semantic similarity. A model similar to this one from Wieting et al. (2019b) was successfully used as a semantic similarity reward in minimum risk training for neural machine translation (Wieting et al., 2019a), building on prior minimum risk training methods for MT (Smith and Eisner, 2006; Shen et al., 2016). Since PARAGRAM-SP can only score the semantic similarity for two texts, with more than two translations, we take the average pairwise PARAGRAM-SP score as the source text's SEMSIM. Mathematically:

$$\text{SEMSIM}\ (s) = \frac{1}{|C|} \sum_{(a,b)\in C} \text{PARAGRAM-SP}(a, b)$$

where M is the set of human translations given a source text s, and C is all possible combinations of a set of pairs in M. PARAGRAM-SP (a, b) is the PARAGRAM-SP score given a pair of English translations a and b.

Note that PARAGRAM-SP computes similarity over embeddings and does not explicitly penalize length. There is an implicit effect only insofar as a reference that omits or heavily alters content will differ in embedding space and score lower, but length alone is not the driver.

We compute SEMSIM for each source text and use these scores to filter and create new datasets. Notice that SEMSIM is computed only with a source text's English translations without relying on the source text itself.

We assume source-target alignment is verified during preprocessing, e.g., Par3 verified source-target alignment. The alignment method used in preprocessing is itself an active research area, but we believe the proposed SEMSIM score helps further filter out misaligned examples regardless of which alignment method is used beforehand. Our filtering therefore targets stylistic and semantic divergence among references that are already aligned to the source: a translation can be faithful yet very different in style.

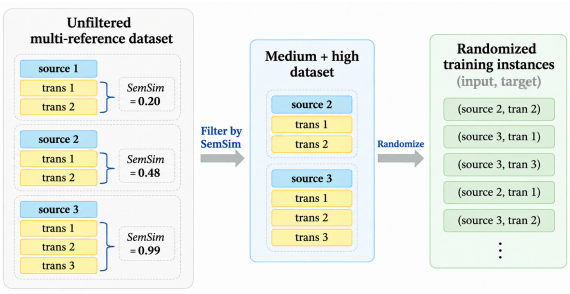

Figure 2: The pipeline of creating the medium+high semantic similarity dataset from the unfiltered dataset. SEMSIM measures the semantic similarity among a source text's target translations.

5 Human Evaluation

To better understand and characterize the outputs from the different systems, we conducted a human evaluation using pairwise A/B comparisons. We recruited 48 translators from North America, Europe, and Africa, through Upwork[1] and ProZ[2].

The translators are native speakers of either French or English and fluent in the other language, and most have prior experience translating literary texts. All evaluations were completed using Label Studio[3]. The translators were paid at the rate of 40 USD for approximately one hour of active annotation time, which is consistent with similar translation rates on Upwork and ProZ.

In each evaluation task, translators were presented with 30 questions. Each question contained a source text in French and two English translations. They were then asked to choose which translation they preferred (A or B). If they considered the two translations to be equally good or bad, they could select "tied". Then the translators were asked to provide a brief justification for their choice.

Detailed annotation instructions, along with an example evaluation question, are provided in Appendix H. We discuss human evaluation results comparing different models in the following sections, alongside automatic evaluation results.

6 Hypothesis 1: Different ranges of semantic similarity will lead to different performance.

We compute the SEMSIM score for the reference translations for each source text, and divide the source texts into three groups based on their SEMSIM scores: [-1, 0.45) is *Low*, [0.45,0.85) is *Medium*, and [0.85, 1] is *High*. We summarize the linguistic characteristics of each range in Table 11.

Each SEMSIM range is then used to fine-tune three different models described in Section 3.2 for literary machine translation. For a fair comparison, we keep the total number of training instances (#INSTN) the same, and total number of source texts (#SRC) as close as possible, and this isolates SEMSIM as the only changing variable. We run each model with three different seeds. Table 1 reports the mean and standard deviation for each automatic metric described in Appendix B.2.

Across models, metrics, and language pairs, fine-tuning on high SEMSIM data on average gives the best performance, followed by medium then low SEMSIM data. However, the performance gap be-

[1] https://www.upwork.com/

[2] ProZ is a platform for hiring freelance translators. https://www.proz.com/

[3] https://labelstud.io

French–English (FR–EN)											
SemSim Level	#src	#instn	BLEU (mean/std)			COMET (mean/std)			chrF++ (mean/std)		
			mT5	Gemma	OpusMT	mT5	Gemma	OpusMT	mT5	Gemma	OpusMT
low	5016	11 279	19.5 (3.69)	19.6 (3.47)	24 (0.5)	0.63 (0.03)	0.64 (0.02)	0.65 (0)	0.36 (0.05)	0.37 (0.02)	0.42 (0)
medium	4977	11 279	34.7 (0.35)	25.3 (1.7)	**30.6** (0.12)	0.71 (0)	0.7 (0.01)	**0.68** (0)	0.50 (0.01)	0.44 (0.01)	0.47 (0.01)
high	4779	11 279	**35.0** (1.07)	**30.6** (1.10)	30.4 (0.26)	**0.72** (0.01)	**0.73** (0)	**0.68** (0.01)	**0.52** (0.01)	**0.50** (0.01)	**0.49** (0)

Russian–English (RU–EN)											
SemSim Level	#src	#instn	BLEU (mean/std)			COMET (mean/std)			chrF++ (mean/std)		
			mT5	Gemma	OpusMT	mT5	Gemma	OpusMT	mT5	Gemma	OpusMT
low	1618	4114	5.25 (0.21)	12.0 (0.38)	12.8 (1)	0.54 (0)	0.58 (0.01)	0.57 (0.01)	0.25 (0.01)	0.30 (0.01)	0.33 (0)
medium	1529	4114	19.2 (1.39)	18.3 (0.64)	18.6 (0.46)	0.64 (0.01)	0.65 (0.01)	0.60 (0.01)	0.38 (0)	0.39 (0.01)	0.37 (0.01)
high	1874	4114	**24.3** (0.36)	**25.9** (1.56)	**19.7** (0.49)	**0.65** (0)	**0.68** (0)	**0.61** (0.01)	**0.4** (0)	**0.45** (0.01)	**0.38** (0)

Table 1: Automatic metric results for mT5, Gemma, and OpusMT after fine-tuning on datasets with different semantic similarity (SEMSIM) levels: low, medium, and high. The top table uses the French-English language pair, and the bottom table uses Russian-English. To fairly evaluate the effect of SEMSIM on literary MT performance, the datasets contain the same number of training instances (#INSTN), and similar number of source sentences (#SRC). We report the mean and standard deviation across three different seeds. Medium and high SEMSIM produce comparable performance and both substantially outperform low SEMSIM.

Setting	Preferred A	Preferred B	Tied
high vs. low	75.8	10.8	13.3
high vs. med	48.7	20.2	31.1
med vs. low	64.2	12.5	23.3

Table 2: Human annotators' vote percentages from A/B testing of outputs produced by mT5-large fine-tuned on datasets with different SEMSIM levels: high, medium, and low. Annotators strongly prefer high and medium SEMSIM outputs over low SEMSIM outputs.

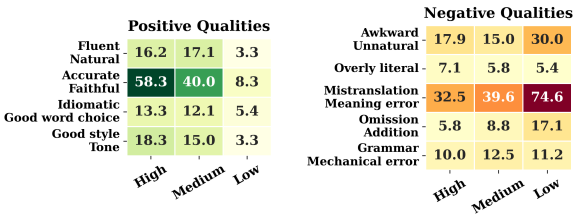

Figure 3: Frequency (%) of translation quality attributes mentioned in annotator comments. Each cell represents the percentage of comments where annotators explicitly noted a positive or negative quality attribute for a given system. High and medium SEMSIM data produce translations that are more fluent, accurate, idiomatic, and have better style and fewer errors than low SEMSIM.

tween medium and high SEMSIM is much smaller than the gap between low and medium SEMSIM.

From low to medium, mT5 shows the most gains, with BLEU increasing by 15.2 (+77.9%), COMET by 0.08 (+12.7%), and chrF++ by 0.14 (+38.9%). Gemma and OpusMT also improve from low to medium, though with smaller relative gains.

In contrast, the performance difference between medium and high is much smaller for mT5 and OpusMT: BLEU changes by less than 1 ($\pm 1\%$), COMET by at most 0.01 ($\leq 1.4\%$), and chrF++ by 0.02 ($\leq 4.3\%$). On the other hand, fine-tuning using Gemma, the performance difference between medium and high is only slightly less than low to medium: BLEU +5.3 (20.9%), COMET +0.03 (4.3%), and chrF++ increases by 0.06 (13.6%).

These results suggest that medium and high SEMSIM data are substantially better for fine-tuning literary MT than low SEMSIM data.

To assess whether these automatic metric trends align with human judgments, we hired 12 professional translators to conduct a pairwise comparison among mT5 outputs generated by fine-tuning on low, medium, and high SEMSIM data.

Source texts were randomly selected. We created four survey variants, with each variant assigned to

three translators, so that each question was rated by three translators. No source text was repeated within a survey, across survey variants, or in any other experiment in this paper.

Table 2 summarizes translators' votes as percentages in the pairwise comparison study. Translators strongly preferred high and medium SEMSIM outputs over low SEMSIM outputs, with preference rates of 75.8% vs. 10.8%, and 64.2% vs. 12.5%, respectively. By contrast, the preference difference is smaller between high and medium SEMSIM, with translators preferring high over medium by 48.7% vs. 20.2%, and many more votes on "tied" (31.1%).

To understand the linguistic characteristics of the outputs, we classify translators' comments on these pairwise comparisons using Claude Opus 4.6 API and visualize them in Figure 3. For each comment, we identify any positive and negative qualities of the translation mentioned, and we report how frequently the outputs from a system were mentioned with this quality. We provide the prompt for classifying their votes using Claude in Appendix F.

Translators described high and medium SEMSIM outputs as comparably fluent, natural, idiomatic, with high SEMSIM as slightly better than medium SEMSIM in accuracy/faithfulness and style/tone, while low SEMSIM outputs were rarely associated with positive qualities.

Conversely, translators found low SEMSIM outputs to contain substantially more mistranslation, omission and/or addition, and to be more awkward and unnatural. Many negative qualities were mentioned nearly twice as often for low SEMSIM than medium and high SEMSIM. Other negative qualities, such as overly literal and grammatical errors, occurred much less often and at similar rates as medium and high SEMSIM.

Overall, the human evaluation suggests that the automatic metric trends are consistent with human judgments. In particular, using medium and high SEMSIM data instead of low SEMSIM data improves perceived accuracy, style, word choice, and fluency, and substantially reduces mistranslation.

7 Hypothesis 2: A filtered dataset will lead to better performance than an unfiltered dataset.

Hypothesis 1 shows that both medium and high SEMSIM data produce substantially better automatic metric and human evaluation results than low SEMSIM data, and the performance between medium and high SEMSIM is comparable. We therefore investigate whether medium and high SEMSIM examples should be combined in practical fine-tuning scenarios. Prior work has also found that, when data quality is comparable, increasing the amount of training data can generally improve model performance (Madnani et al., 2007, 2008).

Similar to Hypothesis 1, we fine-tune each model with three different random seeds and report the mean and standard deviation in Table 3. We compare medium+high data against the unfiltered dataset under two settings: using all available examples, and using controlled subsets with the same number of sources and instances. Although unfiltered (all) contains 14% more training instances than medium+high (all), medium+high (all) performs better across most models and metrics. In particular, mT5 improves from 42.8 to 44.3 BLEU, 0.74 to 0.75 COMET, and 0.55 to 0.56 chrF++; OpusMT improves from 33.6 to 35.2 BLEU while maintaining comparable COMET and chrF++. Gemma is the exception, where the two settings perform similarly.

We observe a similar pattern in the controlled comparison, where medium+low, medium+high, and unfiltered all have the same #SRC and #INSTN. To create medium+low and medium+high datasets, we keep #SRC and #INSTN fixed by using the full medium SEMSIM subset and sampling the same amounts of low and high SEMSIM data. Unfiltered is randomly sampled to match #SRC and #INSTN as well. Under this setting, medium+high performs best overall. Compared with medium+low, medium+high improves BLEU by 3.3 for mT5 and 2.3 for OpusMT, and 0.7 for Gemma, while also giving small improvements in chrF++. Compared with the size-matched unfiltered dataset, medium+high also gives higher BLEU for all three models, while COMET and chrF++ remain comparable or slightly better. These results suggest that adding low SEMSIM data does not provide the same benefit as adding high SEMSIM data, even when the total amount of training data is held constant.

Translators also judged the two systems to be very close, preferring unfiltered in 37.8%, medium+high in 36.1%, with 26.1% marked as ties. Notably, the tie rate is large relative to the preference rates for either system, suggesting that translators often found the two outputs comparable.

The analysis on positive and negative qual-

French–English (FR–EN)											
	#src	#instn	BLEU (mean/std)			COMET (mean/std)			chrF++ (mean/std)		
SemSim Level			mT5	Gemma	OpusMT	mT5	Gemma	OpusMT	mT5	Gemma	OpusMT
medium+low	30 580	69 575	40.6 (1.29)	27.2 (1.15)	32.7 (0.21)	0.74 (0.01)	0.71 (0.01)	0.7 (0)	0.53 (0.01)	0.46 (0.01)	0.5 (0)
medium+high	30 595	69 575	43.9 (0.50)	**27.4** (1.0)	35.0 (0.21)	0.74 (0)	0.71 (0.01)	**0.71** (0.01)	0.55 (0)	0.46 (0.02)	**0.51** (0)
unfiltered	30 580	69 575	42.7 (0.40)	**27.4** (1.15)	34.0 (0.67)	0.74 (0)	0.71 (0)	0.70 (0.01)	0.55 (0)	0.46 (0.01)	**0.51** (0)
unfiltered (all)	39 880	91 647	42.8 (1.53)	27.2 (1.44)	33.6 (0.25)	0.74 (0.01)	0.71 (0.01)	**0.71** (0)	0.55 (0.01)	0.46 (0.01)	**0.51** (0)
medium+high (all)	34 864	80 368	**44.3** (0.60)	26.8 (0.26)	**35.2** (0.31)	**0.75** (0)	0.71 (0)	**0.71** (0.01)	**0.56** (0.01)	0.46 (0)	**0.51** (0.01)

Russian–English (RU–EN)											
	#src	#instn	BLEU (mean/std)			COMET (mean/std)			chrF++ (mean/std)		
SemSim Level			mT5	Gemma	OpusMT	mT5	Gemma	OpusMT	mT5	Gemma	OpusMT
medium+low	17 530	47 641	29.7 (1.63)	22.8 (0.74)	24.9 (0.53)	0.67 (0.01)	0.66 (0)	0.63 (0)	0.45 (0.01)	0.41 (0)	0.42 (0)
medium+high	17 969	47 641	31.6 (0.64)	23.6 (1.46)	25.9 (0.06)	0.67 (0.01)	**0.67** (0.01)	0.64 (0.01)	0.45 (0.01)	0.43 (0.01)	0.43 (0.01)
unfiltered	18 744	47 641	33.1 (0.37)	**24.9** (0.76)	26.3 (0.36)	**0.68** (0)	**0.67** (0)	0.64 (0)	**0.46** (0.01)	**0.44** (0.01)	0.43 (0.01)
unfiltered (all)	26 502	67 106	**33.6** (0.49)	24.5 (0.8)	27.8 (0.32)	**0.68** (0)	0.66 (0.01)	0.64 (0.01)	**0.46** (0.01)	0.42 (0.01)	**0.44** (0)
medium+high (all)	24 884	62 992	33.1 (0.96)	24.6 (0.96)	**28** (0.1)	**0.68** (0.01)	**0.67** (0)	**0.65** (0.01)	**0.46** (0.01)	0.43 (0.01)	**0.44** (0)

Table 3: The top section compares models fine-tuned on medium+low, medium+high, and unfiltered datasets, where all three datasets contain the same number of training instances and a similar number of source texts. The bottom section compares fine-tuning on the full unfiltered dataset with fine-tuning on the full medium+high SEMSIM subset. We report the mean and standard deviation across three different seeds. In both settings, medium+high SEMSIM gives comparable or better performance than unfiltered data.

Setting	%	BLEU	COMET	chrF++
FT w/ unfiltered	37.8	45.3	0.78	0.61
FT w/ med+high all	36.1	45.7	0.78	0.61
Tied	26.1	-	-	-

Table 4: Human annotators' vote percentages from A/B testing of outputs produced by mT5 fine-tuned on the full unfiltered and the full medium+high French-English datasets.

Figure 4: Frequency (%) of translation quality attributes mentioned in annotator comments. Annotators overall found the output from fine-tuning on the full unfiltered data and the full medium+high data to be comparable.

ities also shows a similarly mixed pattern: medium+high has a slight advantage in fluency, word choice/idiomatic, and style/tone, while unfiltered is described slightly more often as accurate. Interestingly, the positive and negative comments are almost contradictory: medium+high is more fluent/natural but also more awkward/unnatural, unfiltered is more accurate/faithful but also has more omission/addition.

Many translators also commented on the limitations of the setup. Translators noted that some translations are hard to decide which is better without even longer context such as the preceding or following sentence around a paragraph. This was especially true for abrupt paragraphs such as quotations, which are common in literary text. In such cases, some translators preferred a more literal translation to be safe, while others favored a more natural and idiomatic interpretation of the original text. A natural transition at the start of a paragraph can, in isolation, look like a mistranslation without more context, especially if a translator took creative liberties with the translation (artistic license).

French–English (FR–EN)											
	#src	#instn	BLEU (mean/std)			COMET (mean/std)			chrF++ (mean/std)		
Train Data			mT5	Gemma	OpusMT	mT5	Gemma	OpusMT	mT5	Gemma	OpusMT
human (single ref)	34 864	34 864	41.4 (0.46)	**28.8** (0.26)	33.5 (0.62)	0.74 (0)	**0.72** (0)	0.70 (0.01)	0.54 (0)	**0.47** (0.01)	0.50 (0.01)
human (multi ref)	34 864	80 368	**44.3** (0.60)	26.8 (0.26)	**35.2** (0.31)	**0.75** (0)	0.71 (0)	**0.71** (0.01)	**0.56** (0.01)	0.46 (0)	**0.51** (0.01)
synthetic (Gemma3)	34 864	80 368	42.0 (0.1)	24.3 (0.32)	35.1 (0.31)	0.74 (0)	0.7 (0)	0.7 (0)	0.55 (0)	0.43 (0.01)	**0.51** (0)
synthetic (Mistral)	34 864	80 368	40.5 (0.4)	22.9 (0.89)	33.2 (0.78)	0.74 (0)	0.7 (0)	**0.71** (0)	0.54 (0.01)	0.43 (0.01)	**0.51** (0)
synthetic (Qwen3)	34 864	80 368	40.2 (2.35)	26.0 (0.92)	34.3 (0.35)	0.72 (0.01)	0.71 (0.01)	0.7 (0)	0.50 (0.01)	0.44 (0.01)	**0.51** (0)

Russian–English (RU–EN)											
	#src	#instn	BLEU (mean/std)			COMET (mean/std)			chrF++ (mean/std)		
Train Data			mT5	Gemma	OpusMT	mT5	Gemma	OpusMT	mT5	Gemma	OpusMT
human (single ref)	24 884	24 884	31.9 (0.66)	23.6 (1.06)	25.6 (0.98)	**0.68** (0.01)	**0.67** (0.01)	0.64 (0)	0.45 (0.01)	0.34 (0.01)	0.43 (0)
human (multi-ref)	24 884	62 992	**33.1** (0.96)	**24.6** (0.96)	**28** (0.1)	**0.68** (0.01)	**0.67** (0)	**0.65** (0.01)	**0.46** (0.01)	**0.43** (0.01)	**0.44** (0)
synthetic (Gemma3)	24 884	62 992	18.9 (1.53)	11.9 (0.69)	25.2 (0.12)	0.63 (0)	0.61 (0.01)	0.63 (0.01)	0.39 (0.01)	0.32 (0.01)	0.41 (0.01)
synthetic (Mistral)	24 884	62 992	23.2 (1.08)	13 (1.24)	25.1 (0.25)	0.65 (0.01)	0.62 (0)	0.63 (0.01)	0.40 (0.01)	0.34 (0.01)	0.42 (0.01)
synthetic (Qwen3)	24 884	62 992	20.7 (1.63)	14.1 (1.13)	25.3 (0.1)	0.64 (0.01)	0.63 (0)	0.63 (0)	0.39 (0)	0.34 (0.01)	0.41 (0.01)

Table 5: Performance comparison between human-only and synthetic-augmented training data. "Human multi ref" denotes training on the full medium and high SEMSIM data with multiple human references, whereas "Human single ref" uses the same source texts with one randomly selected human reference per source. "Synthetic (Gemma3)" denotes training data created by using instruction-tuned Gemma3 to create additional target texts. We report the mean and standard deviation across three different seeds. Synthetic-augmented training data underperform human multi-reference data, but synthetic references generated by a strong instruction-tuned model such as Gemma3 can improve MT performance over a single-reference dataset when fine-tuning using mT5 and OpusMT.

Translators also noted that character names were sometimes translated by both systems. While some translators considered this a serious flaw, others overlooked it.

Overall, the automatic metrics and human evaluation suggest that medium+high data is at least competitive with the unfiltered dataset, despite the unfiltered dataset containing 14% more training examples. The controlled comparison further shows that replacing high SEMSIM examples with low SEMSIM examples hurts performance, supporting the decision to exclude low SEMSIM data.

8 Hypothesis 3: A synthetic-augmented multi-reference dataset will perform comparably to an expert-translated dataset.

We investigate whether augmenting a single-reference dataset with synthetic references can achieve performance comparable to that of expert-translated multi-reference datasets.

Specifically, for each source text in medium+high, we randomly select one human translation and generate paraphrases of this translation using open-weight instruction-tuned LLMs. For all the multi-reference datasets, we keep the number of target texts per source text the same. We provide model details and the prompts we used for zero-shot paraphrase generation in Appendix C. We compare using these synthetic-augmented data against our overall best-performing model, which is fine-tuned on the full medium and high SEMSIM data.

We find that synthetic-augmented datasets provide minimal improvements over the single-reference dataset, or in some cases, even degrade the performance depending on the model used to generate the synthetic references (Table 5). The best model for generating synthetic references, Gemma3, increases BLEU by 0.6 when fine-tuning with mT5, and 1.6 with OpusMT, and improves COMET by 0.01 for both mT5 and OpusMT. However, none of the synthetic datasets improve performance when fine-tuned using Gemma.

Setting	%	BLEU	COMET	chrF++
FT w/ human	**50.4**	45.1	0.79	0.60
FT w/ human + synthetic	32.1	41.0	0.75	0.56
Tied	17.5	-	-	-

Table 6: Human annotators' vote percentages from A/B testing of outputs produced by mT5 fine-tuned on the full medium and high SEMSIM data and the synthetically augmented data. Synthetic data is generated using instruction-tuned Gemma3. Annotators preferred outputs of the MT system fine-tuned on the human data instead of synthetically augmented data by 18%.

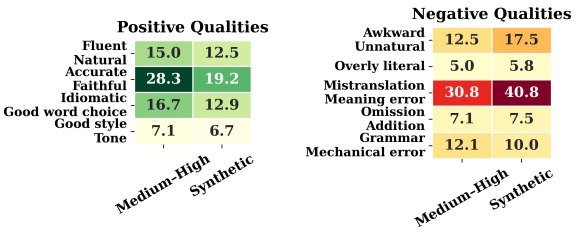

Figure 5: Frequency (%) of translation quality attributes mentioned in annotator comments. Translation outputs from model trained on human data (the full medium and high SEMSIM data) were found to be more accurate, idiomatic, and have less mistranslation than model trained on synthetically augmented data.

Moreover, none of the synthetic datasets outperform the medium+high multi-reference dataset under any model or metric. We further conduct a human evaluation comparing medium+high with the best synthetic-augmented dataset using mT5.

In Figure 8 in the appendix, we also compare the distribution of the pairwise PARAGRAM-SP scores between the ground-truth human translation and each generated synthetic translation. Compared to the PARAGRAM-SP scores between two human translations, the score between a human and a synthetic translation is on average lower across all models. The middle 50% of the distribution shifts down, as does the median. This means the SEMSIM of synthetically augmented data is lower overall.

Translators preferred translations from model fine-tuned on human data over synthetic-augmented data, with preference votes of 50.4% vs. 32.1%, and 17.5% tied.

Qualitative analysis on translators' comments further shows that expert-translated data leads to more accurate, idiomatic, and fluent translations.

Lastly, the human evaluation also includes comparisons between ground-truth expert translations and outputs from our best fine-tuned model.

Unsurprisingly, translators preferred human translations over machine translated text (55.4% vs. 31.3%). Human translations were judged to be substantially more fluent/natural (27.1% vs. 8.8%), more idiomatic (27.1% vs. 9.6%), and have much less mistranslation (12.1% vs. 34.6%).

Setting	Percentage
human GT	55.4
FT w/ med+high all	31.3
Tied	13.3

Table 7: Human annotators' vote percentages from A/B testing of outputs produced by mT5 fine-tuned on the full medium and high SEMSIM data (denoted as "FT w/ med+high all") and ground-truth human expert translations (denoted as "human GT").

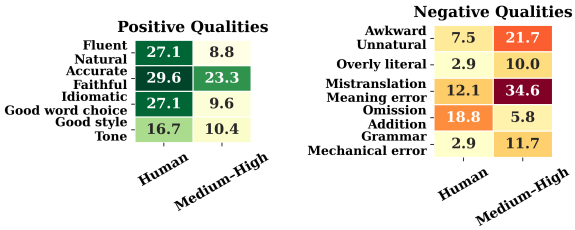

Figure 6: Ground-truth human expert translation were more fluent, idiomatic, and have better style, though more prone to omission and addition than our best MT system output.

Interestingly, translators found human translations more often involved omissions and additions (18.8% vs 5.8%), echoing the concerns raised in Hypothesis 2, where translators noted that some comparisons are difficult to judge without additional context. For example, if the name "Christine" appears in a previous paragraph, a human translator may use "she" in the following paragraph to refer to Christine, even if the original French text repeats "Christine" in that paragraph. In contrast, machine translated text is more likely to retain "Christine" in order to match the French source text.

9 Conclusion

Most MT fine-tuning work treats references as interchangeable ground truth, using a single reference per source or weighting all references equally, without considering whether the references agree with one another. In contrast, we use the semantic similarity among multiple references as a filtering signal—an idea motivated by the comparative-literature practice of reading multiple translations. To our knowledge, this is the first work to do so for literary MT.

We propose a framework for filtering a multi-reference dataset for literary MT, a setting where multiple translations of a source text often exist. Results from automatic metrics on both FR-EN and RU-EN show that high and medium semantic similarity data is better for fine-tuning MT systems than low semantic similarity data. Human evaluation confirms this: translators overwhelmingly prefer systems fine-tuned on high SEMSIM over low, and medium over low. When commenting on the outputs, translators rarely associate any positive qualities with low SEMSIM, while their assessments of high and medium are more comparable, with high SEMSIM holding a slight edge. Importantly, references with low semantic similarity among them are not merely misaligned or erroneous; they are also cases where translators diverge substantially in their interpretations.

We also find that while using more data generally improves performance, unfiltered data does not always help. Medium+high data can give comparable and at times better performance, all with less data. Human translators also found outputs from medium+high data more idiomatic and better in style, though overall performance is comparable to using unfiltered data. Translators also noted limitations of the setup that made comparison difficult: paragraphs were presented without preceding or following context, and translators differed in their own preferences, favoring more literal versus more natural, idiomatic interpretations.

We are not aware of any prior work that compares fine-tuning using human versus synthetic-augmented data for literary MT, particularly with human evaluation that examines the language of the output translations themselves. Comparing synthetic against human translations, we find that synthetic data, unlike human data, does not reliably improve MT. Performance depends heavily on the LLM used to paraphrase the ground-truth translation. For French-English, of the three LLMs we tested, only Gemma3 improved MT performance compared to using a single reference, and the improvement was small. For Russian-English, performance worsened, suggesting that paraphrase errors and deviation from the original translation act as noise. In other words, augmenting a dataset by zero-shot LLM paraphrasing of references can hurt MT performance. Interestingly, the best paraphrasing LLM also had, on average, a lower SEMSIM, while remaining mostly in the medium SEMSIM range (Figure 8).

This preference for human data is echoed in human evaluation: comparing outputs from models fine-tuned on human versus synthetic augmented data, translators preferred those trained on human data much more, finding them more accurate, better in word choice, and far less prone to mistranslation.

As a sanity check, we also compared the ground-truth translations with our best fine-tuned model (mT5 fine-tuned on all medium+high SEMSIM data), and translators preferred the ground truth at a rate similar to the earlier comparisons. While the ground-truth translations were most frequently credited with positive qualities, they were also, surprisingly, mentioned far more often for omission and addition. Considering how paragraphs flow, this makes sense: good flow often means being less literal, with translators moving content into the preceding or following paragraph.

Finally, we tested zero-shot translation (Appendix D). Compared to our best fine-tuned model (mT5 fine-tuned on all medium+high SEMSIM data), our model outperformed all open-weight models we tested. Among commercial models, two of three outperformed ours, one only marginally (44.8 vs. 44.3 in BLEU).

Meaningful human evaluation requires a sufficient amount of data, which limits us to high-resource language pairs, since low-quality output will also result in noisy human evaluation when errors are frequent and of many kinds. We hope our findings provide guidance for literary translation on both high- and low-resource language pairs: prioritize human translations, and if using synthetic LLM-based paraphrasing—even into a high-resource target language like English—choose the LLM carefully and inspect the paraphrased output, or performance may suffer.

Future work can compare LLM zero-shot paraphrasing with more controlled generation methods, such as pivot-based or neural approaches, especially for low-resource languages.

Limitations

The following limitations of our paper are also directions for future work.

1. We create additional synthetic references only using LLMs in this paper. Future work can explore different methods for generating synthetic references and compare their performance. More controlled generation could po-

tentially prevent the overall semantic similarity shift demonstrated in Figure 8.

2. While we would have liked to work on low-resource languages in this paper, the premise of the paper relies on multi-reference data, which generally exists only for high-resource language pairs. Additionally, conducting human evaluation on low-quality translations in low-resource setting would be challenging. If the outputs are not adequate, it is difficult to obtain meaningful feedback from translators. However, we believe our findings can benefit future work on low-resource settings. Specifically, it is worth investigating whether filtering by semantic similarity between synthetic references can help improve low-resource MT.

3. Due to our computational constraints, we could not experiment with LLMs with even larger numbers of parameters. For example, we could only fine-tune Gemma3 with 12 billion parameters using QLoRA, and its performance was, for the most part, worse than that of mT5-Large.

4. We would also like to extend the context window by training on longer data. As discussed in Hypothesis 2, translators mentioned that without the preceding and following paragraphs, it is sometimes difficult to judge which translation is better or whether there is a mistranslation. However, due to computational constraints and the nature of the data in Par3, we are limited to paragraph-level data, which is still better than sentence-level data. Ideally, book-chapter level translation would be a great next step.

Acknowledgments

We thank Huda Khayrallah and Brian Thompson for their early feedback on the paper. Si Wu and David Smith were supported in part by the Schmidt Sciences-funded project "Beyond Translation: Opening up the Human Record."

References

Laurent Besacier and Lane Schwartz. 2015. Automated translation of a literary work: A pilot study. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 114–122, Denver, Colorado, USA. Association for Computational Linguistics.

Eleftheria Briakou and Marine Carpuat. 2021. Beyond noise: Mitigating the impact of fine-grained semantic divergences on neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7236–7249, Online. Association for Computational Linguistics.

Paola Brusasco. 2022. Pragmatic and cognitive elements in literary machine translation: An assessment of an excerpt from j. polzin's brood translated with google, deepl, and microsoft bing. In *Using Technologies for Creative-Text Translation*, pages 139–160. Routledge.

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

Marine Carpuat, Yogarshi Vyas, and Xing Niu. 2017. Detecting cross-lingual semantic divergence for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 69–79, Vancouver. Association for Computational Linguistics.

Antonio Castaldo, Sheila Castilho, Joss Moorkens, and Johanna Monti. 2025. Extending creamt: Leveraging large language models for literary translation post-editing.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Gloria Corpas Pastor and Laura Noriega-Santiáñez. 2024. Human versus neural machine translation creativity: A study on manipulated mwes in literature. *Information*, 15(9).

Ona de Gibert, Joseph Attieh, Teemu Vahtola, Mikko Aulamo, Zihao Li, Raúl Vázquez, Tiancheng Hu, and Jörg Tiedemann. 2025. Scaling low-resource MT via synthetic data generation with LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27674–27692,

Suzhou, China. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Douglas R. Hofstadter. 1997. *Le Ton beau de Marot: In Praise of the Music of Language*. Basic Books, New York.

Zachary Horvitz, Ajay Patel, Chris Callison-Burch, Zhou Yu, and Kathleen McKeown. 2024. Paraguide: Guided diffusion paraphrasers for plug-and-play textual style transfer.

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

Roman Jakobson. 1987. Linguistics and poetics. In Krystyna Pomorska and Stephen Rudy, editors, *Language in Literature*, pages 62–94. Harvard University Press, Cambridge, MA.

Lasal Jayawardena and Prasan Yapa. 2024a. Parafusion: A large-scale llm-driven english paraphrase dataset infused with high-quality lexical and syntactic diversity. In *Artificial Intelligence and Big Data*, AIBD. Academy & Industry Research Collaboration Center.

Lasal Jayawardena and Prasan Yapa. 2024b. Parameter efficient diverse paraphrase generation using sequence-level knowledge distillation. In *2024 5th International Conference on Advancements in Computational Sciences (ICACS)*. IEEE.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia. Association for Computational Linguistics.

Huda Khayrallah, Brian Thompson, Matt Post, and Philipp Koehn. 2020. Simulated multiple reference training improves low-resource machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 82–89, Online. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

André Lefevere. 1992. *Translating literature : practice and theory in a comparative literature context*. New York :Modern Language Association of America.

Kechen Li, Yaotian Tao, Hongyi Huang, and Tianbo Ji. 2024. LinChance-NTU for unconstrained WMT2024 literary translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 987–992, Miami, Florida, USA. Association for Computational Linguistics.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On LLMs-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11065–11082, Bangkok, Thailand. Association for Computational Linguistics.

Nitin Madnani, Necip Fazil Ayan, Philip Resnik, and Bonnie Dorr. 2007. Using paraphrases for parameter tuning in statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 120–127, Prague, Czech Republic. Association for Computational Linguistics.

Nitin Madnani, Philip Resnik, Bonnie J. Dorr, and Richard Schwartz. 2008. Are multiple reference translations necessary? investigating the value of paraphrased reference translations in parameter optimization. In *Proceedings of the 8th Conference of the Association for Machine Translation in the Americas: Research Papers*, pages 143–152, Waikiki, USA. Association for Machine Translation in the Americas.

Bojana Mikelenić, Antoni Oliver, and Sergi Àlvarez Vidal. 2025. Fine-tuning and evaluation of NMT models for literary texts using RomCro v.2.0. In *Proceedings of the Second Workshop on Creative-text Translation and Technology (CTT)*, pages 44–51, Geneva, Switzerland. European Association for Machine Translation.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models.

Vladimir Nabokov. 1955. Problems of translation: "Onegin" in english. *Partisan Review*, 22(4):496–512.

E.A. Nida. 1964. *Toward a Science of Translating: With Special Reference to Principles and Procedures Involved in Bible Translating (Second edition)*. Brill.

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

Alberto Poncelas, Dimitar Shterionov, Andy Way, Gideon Maillette de Buy Wenniger, and Peyman Passban. 2018. Investigating backtranslation in neural machine translation. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pages 269–278, Alicante, Spain.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

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

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

Marilyn Gaddis Rose. 1997. *Translation and Literary Criticism: Translation as Analysis*. St. Jerome Publishing, Manchester, UK.

Barbara Scalvini, Iben Nyholm Debess, Annika Simonsen, and Hafsteinn Einarsson. 2025. Rethinking low-resource MT: the surprising effectiveness of fine-tuned multilingual models in the LLM age. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 609–621, Tallinn, Estonia. University of Tartu Library.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum risk training for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1692, Berlin, Germany. Association for Computational Linguistics.

David A. Smith and Jason Eisner. 2006. Minimum risk annealing for training log-linear models. pages 787–794.

Steinþór Steingrímsson, Hrafn Loftsson, and Andy Way. 2023. Filtering matters: Experiments in filtering training sets for machine translation. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 588–600, Tórshavn, Faroe Islands. University of Tartu Library.

Josh Stenberg. 2018. Multiplicity in lieu of authority: Translations of classical chinese poetry online. *Babel*, 64(4):579–593.

Dan Su, Xiaoguang Li, Jindi Zhang, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. 2022. Read before generate! faithful long form question answering with machine reading.

Kasia Szymanska. 2026. A manifesto on multiple translation: Bridging retranslation, world literature, and the ethics of comparative reading. *Translation Studies*.

Kristiina Taivalkoski-Shilov. 2019. Free indirect discourse: an insurmountable challenge for literary MT systems? In *Proceedings of the Qualities of Literary Machine Translation*, pages 35–39, Dublin, Ireland. European Association for Machine Translation.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric

Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report.

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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato Yves Scherrer, Raul Vazquez, and Sami Virpioja. 2023. Democratizing neural machine translation with OPUS-MT. *Language Resources and Evaluation*, (58):713–755.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

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.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019a. Beyond BLEU:training neural machine translation with semantic similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4344–4355, Florence, Italy. Association for Computational Linguistics.

John Wieting, Kevin Gimpel, Graham Neubig, and Taylor Berg-Kirkpatrick. 2019b. Simple and effective paraphrastic similarity from parallel translations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4602–4608, Florence, Italy. Association for Computational Linguistics.

John Wieting, Kevin Gimpel, Graham Neubig, and Taylor Berg-kirkpatrick. 2022. Paraphrastic representations at scale. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 379–388, Abu Dhabi, UAE. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Renjie Zheng, Mingbo Ma, and Liang Huang. 2018. Multi-reference training with pseudo-references for neural translation and text generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3188–3197, Brussels, Belgium. Association for Computational Linguistics.

A Pre-Processing

We filter out source paragraphs and translations longer than 250 words to remove extreme outliers and ensure reliable processing by sentence-based MT models like OpusMT. The 250-word cutoff is also informed by the distribution of word length in the dataset: in Figure 7, we show the distributions of both source and target paragraphs length measured by word count in the French train and validation datasets. After filtering, we only keep instances where each source paragraph has two or more human translations.

Then, for each language, we construct train, validation, and test splits at the book level, preventing any paragraph overlap across splits. We target an 80/10/10 paragraph ratio. Exact statistics for training instances for all splits and datasets are reported in Table 8 and 9.

The train and validation datasets are the starting points for more targeted experiments from Sections 6 to 8, where we will further filter by the semantic similarity among the human translations. The test dataset created is used to test all experiments for a specific language pair. The French test set contains 5,030 source texts, and Russian has 3,292. Each source text in the test set has multiple human expert translations as targets.

	Train		Validation	
Setting	**src**	**instance**	**src**	**instance**
low	5,016	11,279	470	940
med	4,977	11,279	470	940
high	4,779	11,279	470	940
med+low	30,580	69,575	4,208	8,416
med+high	30,595	69,575	4,208	8,416
unfiltered	30,580	69,575	4,208	8,416
unfiltered	39,880	91,647	4,987	9,974
medium+high all	34,864	80,368	4,517	9,034
human (single ref)	34,864	34,864	4,517	4,517
synthetic (Gemma3)	34,864	80,368	4,517	9,034
synthetic (Mistral)	34,864	80,368	4,517	9,034
synthetic (Qwen3)	34,864	80,368	4,517	9,034

Table 8: French train and validation source and instance counts.

	Train		Validation	
Setting	**src**	**instance**	**src**	**instance**
low	1,618	4,114	384	1,087
med	1,529	4,114	506	1,087
high	1,874	4,114	520	1,087
med+low	17,530	47,641	2,322	5,491
med+high	17,969	47,641	2,534	2,534
unfiltered	18,744	47,641	2,344	5,491
unfiltered	26,502	67,106	3,296	7,675
medium+high all	24,884	62,992	2,534	5,491
human (single ref)	24,884	24,884	2,534	2,534
synthetic (Gemma3)	24,884	62,992	2,534	5,491
synthetic (Mistral)	24,884	62,992	2,534	5,491
synthetic (Qwen3)	24,884	62,992	2,534	5,491

Table 9: Russian train and validation source and instance counts.

B MT Models and Automatic Metrics

B.1 Models

mT5-large A multilingual encoder–decoder Transformer model with approximately 1.2B parameters. It is pre-trained on the mC4 corpus, which covers 101 languages, including French and Russian (Xue et al., 2021).[4] Although mT5-large is not specifically pre-trained for translation, it can be fine-tuned for MT and other downstream tasks. We fine-tuned this model using the Hugging Face `Trainer` class. Each experiment takes less than 24 hours on a single NVIDIA H200 GPU. The learning rate $= 1e-3$, batch size $= 4$, and gradient accumulation $= 4$.

Gemma-3-12B-PT An open-weight, non-instruction-tuned language model with 12 billion parameters.[5] Given its scale and our computational constraints, we adapt the model using LoRA with 4-bit quantization (Hu et al., 2022; Dettmers et al., 2023). Gemma3 is trained on multilingual web text covering 140 languages, although the language distribution is not publicly specified (Team et al., 2025). We use LoRA rank $r = 32$ and scaling parameter $\alpha = 32$. Each experiment takes from 48 to 120 hours on a single NVIDIA H200 GPU, depending on the size of the dataset. The learning rate$= 1e-4$, batch size $= 4$, and gradient accumulation $= 8$.

OpusMT An open-source multilingual machine translation model from the OPUS-MT project (Tiedemann et al., 2023; Tiedemann and Thottingal, 2020).[6] The model supports multiple source languages with English as the target language and is trained on the OPUS corpus using the Marian framework (Junczys-Dowmunt et al., 2018). Each experiment takes less than 12 hours on a single NVIDIA Tesla V100 SXM2 GPU. The learning

[4] This model is available on Hugging Face at https://huggingface.co/google/mt5-large

[5] https://huggingface.co/google/gemma-3-12b-pt

[6] https://huggingface.co/Helsinki-NLP/opus-mt-mul-en

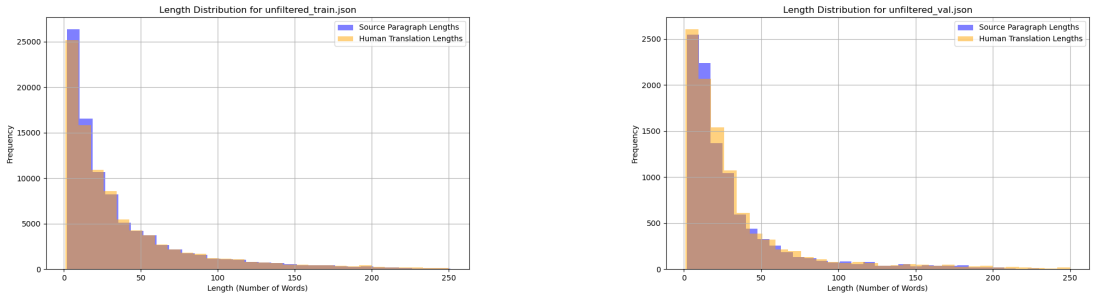

Figure 7: Distribution of paragraph word lengths in the French TRAIN (left) and VAL (right) datasets. Each instance represents a source paragraph or translation instance, showing the frequency of paragraphs by word count.

rate $= 5e-5$, batch size $= 8$, and gradient accumulation $= 4$.

B.2 Automatic Evaluation Metrics

BLEU An MT evaluation metric based on modified n-gram precision with a brevity penalty (Papineni et al., 2002). We compute BLEU using multiple human references for each source input in the test set.

$\text{COMET}_{22}^{\text{DA}}$ An MT evaluation metric that is based on sentence embeddings (Rei et al., 2020). We use wmt22-comet-da by Unbabel (Rei et al., 2022).[7] It was trained on a transformer-based multilingual masked language model, XLM-R (Conneau et al., 2020). Since $\text{COMET}_{22}^{\text{DA}}$ only accepts one reference per source text for evaluation, for each source text, we take the average of their $\text{COMET}_{22}^{\text{DA}}$ scores for each reference.

chrF++ An MT evaluation metric based on character and word-level n-gram precision and recall (Popović, 2017).

[7] https://huggingface.co/Unbabel/wmt22-comet-da

C Models and Prompts Used for Generating Synthetic Data

We first randomly select one human translation from each source text in the full medium+high dataset. Then, each one of these human translations is used to generate paraphrase with the following models. The following models were all able to zero-shot generate paraphrases for at least 99% of the data.

We use 1,000 as the limit for max new token, and retry with 1,200 if the output is cut off. For the few instances where we could not successfully parse the model output because the model did not follow the instructions, we salvaged as many generated paraphrases as possible, then replaced any missing paraphrases with human translations.

gemma-3-27b-it[8]: gemma-3-27b-it is Google's 27-billion-parameter instruction-tuned Gemma3 model (different from the 12b-pt variant we use for fine-tuning). It is designed for instruction following, and it can handle long-context window. To generate text more efficiently, we perform zero-shot text generation using vLLM (Kwon et al., 2023) with a batch size of 16. For the French TRAIN split, generation completed on a single H200 GPU in under 17 hours. We obtained outputs for all 34,864 instances. Only 2 instances produced a single paraphrase rather than the requested two. For the validation split, we successfully processed all 4,517 instances, with no single-paraphrase outputs and no discarded instances.

> System prompt: You are a helpful assistant.
>
> User prompt: Generate two paraphrases of the paragraph below. Each paraphrase should preserve the same meaning as the original, but express it in a different way. The two paraphrases

[8] Available to download on Hugging Face at https://huggingface.co/google/gemma-3-27b-it

should not be identical or trivial rewrites. Set `reference` to exactly the paragraph provided. Return ONLY valid JSON in the following format:

```
{
  "reference": "...",
  "paraphrase_1": "...",
  "paraphrase_2": "..."
}

[INPUT]
```

Qwen3-30B-A3B-Instruct-2507[9]**:** An instruction-tuned mixture-of-experts language model from the Qwen3 family, with 30.5 billion total parameters and 3.3 billion activated parameters. It has strong instruction following and general reasoning capabilities. Like Gemma3, we ran inference using vLLM with a batch size of 8. The TRAIN split finished under 5 hours. Of the 34,864 source paragraphs in the TRAIN split, only 54 outputs contained a single paraphrase rather than the requested two paraphrases. For the validation split, all outputs were successfully salvaged; only 1 instance contained a single paraphrase.

User prompt: Generate two paraphrases of the paragraph below. Each paraphrase should convey the same meaning as the original, but be expressed in a different way. The two paraphrases should not be identical or trivial rewrites. Set `reference` to be exactly the paragraph provided, without any changes. Return only valid JSON in the following format:

```
{
  "reference": "...",
  "paraphrase_1": "...",
  "paraphrase_2": "..."
}

[INPUT]
```

Mistral-Small-3.2-24B-Instruct-2506[10]**:** A 24-billion-parameter instruction-tuned model from Mistral AI. It is designed for efficient instruction following as well as long context use. We ran inference using vLLM with a batch size of 8. The TRAIN split finished in under 24 hours. For the TRAIN split, 85 outputs could not be salvaged. 1,959 instances had one single paraphrase. For the validation split, all 4,517 instances were processed: 4,290 produced complete paraphrase pairs, 220 produced a single recoverable paraphrase, and 7 were discarded and replaced with human translations.

User prompt: Generate two paraphrases of the paragraph below. Each paraphrase must preserve the original meaning but be written differently. Avoid trivial rewrites.

Return ONLY valid JSON in the following format:

```
{
  "reference": "...",
  "paraphrase_1": "...",
  "paraphrase_2": "..."
}

[INPUT]
```

[9] Available to download on Hugging Face at https://huggingface.co/Qwen/Qwen3-30B-A3B-Instruct-2507

[10] Available to download on Hugging Face at https://huggingface.co/mistralai/Mistral-Small-3.2-24B-Instruct-2506

D Zero-shot Translation with Instruction-tuned Models

While this paper primarily explores fine-tuning MT models using datasets with different levels of semantic similarity using multi-reference data, it is also informative to know the zero-shot translation performance of commercial models (via API) and open-weight instruction-tuned LLMs. In Table 10 we show their performance measured by automatic metrics on the French test set.

Results from the automatic metrics show that while commercial models like GPT 5.4 Pro outperform our best model (medium+high (all) with mT5) by 4.2 in BLEU, 0.01 in COMET, and 0.04 in chrF++, other state-of-the-art commercial models like Gemini 3.1 pro and Claude Opus 4.6 are only marginally better or worse than our best model.

Open-weight models not only all underperform our best fine-tuned model, some even struggle to process and translate given the instructions, with a valid output rate as low as 52.9% using Mistral-Small-3.2-24B-Instruct-2506 (Table 10).

model	cost	BLEU	COMET	chrF++	Valid%
GPT 5.4 pro	$15.47	**48.5**	**0.76**	**0.6**	100
Gemini 3.1 Pro	$37	**44.8**	**0.76**	0.55	99.9
Claude Opus 4.6	$7.95	44.2	**0.76**	**0.6**	100
Gemma3-27b-instruct	0	40.6	0.75	0.55	99.5
Qwen3-instruct	0	40.1	0.73	0.54	75.9
Mistral-instruct	0	35.4	0.66	0.50	52.9
our best FT	0	44.3	0.75	0.56	100

Table 10: Zero-shot machine translation performance on French-English test data. Scores in bold font are better than our best fine-tuned model—fine-tuning mT5 using the full medium and high SEMSIM data. "Valid%" counts how many test data were able to be processed and translated. Cost for Gemini is estimated.

Below, we provide details of these models.

D.1 Commercial Models

GPT 5.4 Pro A commercial model developed by OpenAI with long context window, multilingual capability, and instruction following ability. We used the model with reasoning effort set to `high`.

Gemini 3.1 pro A commercial model developed by Google Deepmind. Similar to GPT 5.4 Pro, it has multilingual generation capability, long context window, and instruction following ability. We used the model with reasoning effort set to `high`.

Claude Opus 4.6 A commercial model developed by Anthropic. It is capable of multilingual tasks, has a massive 1 million token context window, and instruction following ability.

For all commercial models, we use the following system prompt for zero-shot machine translation. All APIs were called in batch mode.

> System prompt: You are an expert literary translator from French to English. Translate the following French paragraph into natural, fluent English that preserves the original tone, style, and nuance. Output ONLY the English translation with no explanation or commentary.
>
> `[INPUT]`

D.2 Open-Weight Models

For zero-shot translation, we used the same open-weight models that we used to zero-shot generate paraphrase in Appendix C. Below, we show the prompts we used for generation.

gemma-3-27b-it

> System prompt: You are a careful literary translator.
>
> User prompt:
> Task: Translate the following French paragraph into natural, fluent English.
> Style: literary; preserve tone, register, and meaning; avoid adding new facts.
> Output format: Return ONLY a valid JSON object with exactly this key:
>
> ```
> {
>  "prediction": "..."
> }
> ```
>
> Rules: Do NOT include markdown, code fences, commentary, or extra keys.
> Keep the English as a single string, including punctuation or quotes as needed.
>
> French paragraph:
>
> `[INPUT]`

Qwen3-30B-A3B-Instruct-2507

> User prompt:
> Task: Translate the following French paragraph into natural, fluent English.
> Style: literary; preserve tone, register, and meaning; do not add new facts.
>
> Return ONLY valid JSON with exactly this key:
>
> ```
> {
>  "prediction": "..."
> }
> ```
>
> Rules:
> - No markdown, no code fences, no explanations.
> - Do NOT add extra keys.
> - Keep the English as a single string.
>
> French paragraph:
> `[INPUT]`

Mistral-Small-3.2-24B-Instruct-2506

User prompt:
Task: Translate the following French paragraph into natural, fluent English.
Style: literary; preserve tone, register, and meaning; do not add new facts.

Return ONLY valid JSON with exactly this key:

```
{
  "prediction": "..."
}
```

Rules:
- No markdown, no code fences, no extra keys, no explanations.
- Keep the English as a single string.

French paragraph:
`[INPUT]`

E PARAGRAM-SP Score Distribution of the Synthetic Data

In Figure 8, we compare the distribution of the pairwise PARAGRAM-SP scores between the ground-truth human translation and each LLM's zero-shot paraphrased translation.

Compared to the PARAGRAM-SP scores between two human translations, the score between a human and a synthetic translation is on average lower across all models.

In the violin plot, the middle 50% of the distribution shifts down for all three models, as does the median. This means the SEMSIM of synthetically augmented data is lower overall, since SEMSIM is the average pairwise score among all references of the same source. Interestingly, Gemma3 deviates the most from the human score distribution, yet its paraphrases gave the best MT performance of the three models. Note that for all three models, the majority of the data falls in the medium SEMSIM range, whereas nearly half of the human data falls in the high SEMSIM range.

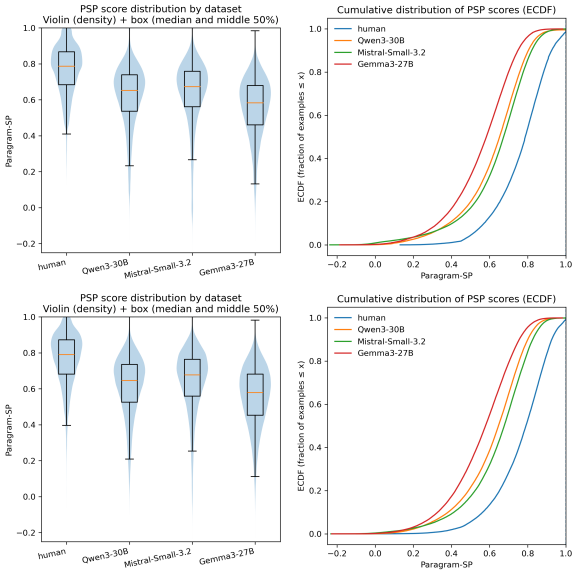

Figure 8: The distributions of the pair-wise semantic similarity score between one generated additional reference and the human translation. The top row is from the French-English data, and bottom row is from the Russian-English data. "Human" pair-wise similarity score distribution is from the multi-reference full medium+high SEMSIM dataset. Overall, the synthetic pairwise similarity score distribution shifts lower compared to the human translation distribution.

F Prompt for Classifying Translators' comments

We used Claude Opus 4.6 to label annotator comments with fine-grained translation quality categories. For each comment, the model was given the original French source text, Translation A, Translation B, the annotator's free-text comment, and the full list of quality labels. We wrote this detailed instruction with the assistance of Claude Opus 4.6 in the chat interface. We tested multiple instructions by checking 50 outputs to ensure accuracy. Below is the prompt we use:

> You are a strict annotation analyst for literary translation evaluation (French to English). You will be given:
>
> - The original French source text
> - Translation A and Translation B (the actual translations)
> - A human annotator's comment comparing them
> - A list of quality labels
>
> For EACH quality label, decide:
>
> - "A" if the comment EXPLICITLY praises or criticizes Translation A for this quality
> - "B" if the comment EXPLICITLY praises or criticizes Translation B for this quality
> - "tie" if the comment EXPLICITLY says BOTH translations share this quality
> - "" (empty) if the comment does NOT clearly address this quality
>
> STRICT RULES:
>
> - Only tag a quality if the comment clearly and directly discusses it. Do NOT infer qualities that are merely implied.
> - Most qualities should be "" (empty) for any given comment. A typical comment only addresses 1–3 qualities.
> - Annotators often quote specific words or phrases from the translations without saying "A" or "B". You MUST match quoted words back to Translation A or Translation B to determine which system they refer to.
> - If the comment says one translation "flows better" or "is more natural", that is "fluent / natural" for that system AND "awkward / unnatural" for the other ONLY if the comment indicates the other is worse.
> - Do NOT automatically assign the opposite negative quality just because one system got a positive. Only assign it if the comment actually criticizes the other.
>
> Quality definitions:
>
> - "fluent / natural": reads smoothly, sounds like natural English
> - "accurate / faithful": correctly conveys the meaning of the French source
> - "idiomatic / good word choice": uses appropriate English expressions and well-chosen words
> - "good style / tone": preserves the literary register, mood, or voice of the original
> - "awkward / unnatural": reads clumsily, sounds unnatural or stilted
> - "overly literal": follows French syntax or phrasing too closely, word-for-word translation
> - "mistranslation / meaning error": gets the meaning wrong, misinterprets the source
> - "omission / addition": leaves out content from the source or adds content not in the source
> - "grammar / mechanical error": has grammatical mistakes, typos, or mechanical issues
>
> EXAMPLES:
>
> Example 1:
>
> Original French: [French source text]
>
> Translation A: [some translation with more natural English]
>
> Translation B: [some translation with less natural English]
>
> Comment: "I've gone with A purely because the English is more natural and flows more smoothly than B. However, there are serious issues in both: neither gets d'Artagnan's name right, neither explains what a conge is (it would have been better to write 'permissions to leave' or something like that and I'd have preferred 'pistols' to 'pistoles'."
>
> Correct labels:
>
> {"fluent / natural": "A", "accurate / faithful": "", "idiomatic / good word choice": "", "good style / tone": "", "awkward / unnatural": "B", "overly literal": "", "mistranslation / meaning error": "tie", "omission / addition": "", "grammar / mechanical error": ""}
>
> Example 2:
>
> Original French: [French source text]
>
> Translation A: [some translation preserving Milady's voice]
>
> Translation B: [some translation]
>
> Comment: "A sounds more like Milady would speak."
>
> Correct labels:
>
> {"fluent / natural": "", "accurate / faithful": "", "idiomatic / good word choice": "", "good style / tone": "A", "awkward / unnatural": "", "overly literal": "", "mistranslation / meaning error": "", "omission / addition": "", "grammar / mechanical error": ""}
>
> Respond with ONLY a JSON object. No explanation, no markdown, no backticks.

For each instance, the user prompt was constructed using the following template:

Original French: {original}

Translation A: {a_pred}

Translation B: {b_pred}

Annotator comment:

"{comment}"

Quality labels:

- fluent / natural
- accurate / faithful
- idiomatic / good word choice
- good style / tone
- awkward / unnatural
- overly literal
- mistranslation / meaning error
- omission / addition
- grammar / mechanical error

Return a JSON object mapping each quality label to "A", "B", "tie", or "". Remember: most should be empty. Only tag what the comment explicitly discusses.

G Descriptions of Each SEMSIM Level (Table 11)

H Human Evaluation Interface Details

Across all tasks, no source text or translation is repeated within a survey, across survey variations for the same task, or in any other task.

Translators were informed of the duration, the pay, and the nature of the task before accepting the job.

In Figure 9 and 10, we show the instructions and the UI for annotation on Label Studio, as well as an example question.

I Example Translation Outputs from Different MT Systems (Table 12)

SEMSIM	Description	Dataset
1.0	Completely identical.	Medium+High Unfiltered
$[0.9 - 1.0)$	Almost identical, except for minor punctuation differences, e.g., single vs double quotation marks, or word replacement (by a synonym or a phrase).	Medium+High Unfiltered
$[0.85 - 0.9)$	Differences happen more often, but they are still variations that are not meaningful. Compared to the 0.9-1.0 similarity, the difference can be clausal instead of just one or two words. However, the overall amount of information is about the same.	Medium+High Unfiltered
$[0.45 - 0.85)$	Very meaningful variations. The paraphrases are more likely to be visually dissimilar. Sometimes, the amount of information can differ much more: some translators are more verbose, others are more succinct.	Medium Medium+High Medium+Low Unfiltered
$[-1 - 0.45)$	Noisy paraphrases that have an overall information mismatch. Paraphrases can be misaligned or erroneous, or translators diverge too much in their interpretations.	Medium+Low Unfiltered

Table 11: Different levels of semantic similarity score (SEMSIM) and their characteristics. See Section 4 for how we compute SEMSIM given a source text and its reference translations. The dataset column indicates the datasets that include paraphrases from the corresponding SEMSIM range.

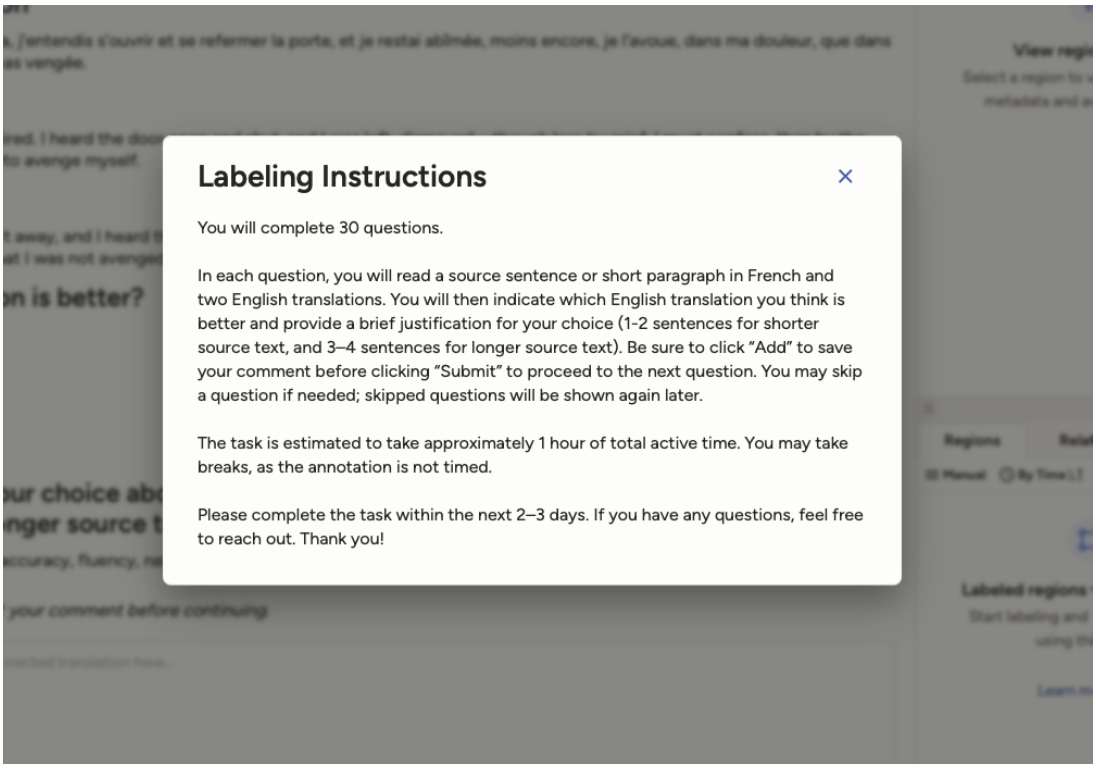

Figure 9: The instructions shown to translators before starting the annotation process. The human evaluation task was hosted on Label Studio.

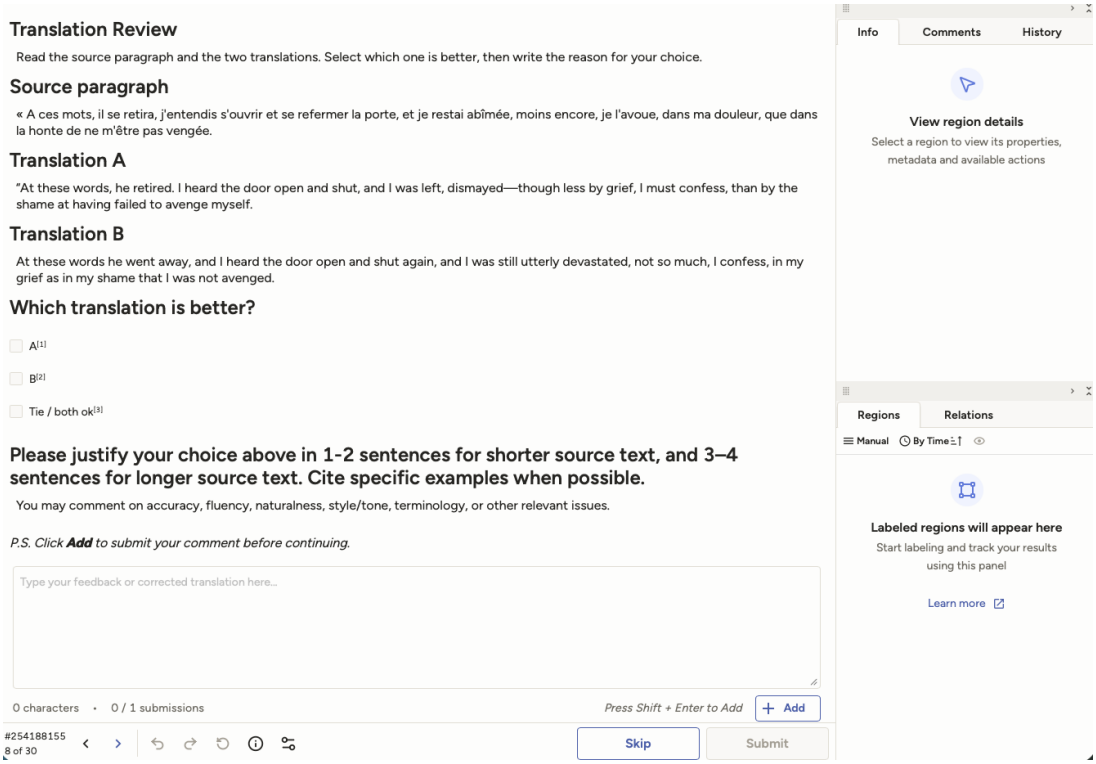

Figure 10: A screenshot of the Label Studio interface used for human evaluation, showing an example question presented to translators.

System	Output
Source (FR)	*— Ah! voilà qui est nouveau, reprit le roi. N'allez-vous pas dire que vos trois mousquetaires damnés, Athos, Porthos et Aramis et votre cadet de Béarn, ne se sont pas jetés comme des furieux sur le pauvre Bernajoux, et ne l'ont pas maltraité de telle façon qu'il est probable qu'il est en train de trépasser à cette heure! N'allez-vous pas dire qu'ensuite ils n'ont pas fait le siège de l'hôtel du duc de La Trémouille, et qu'ils n'ont point voulu le brûler! ce qui n'aurait peut-être pas été un très grand malheur en temps de guerre, vu que c'est un nid de huguenots, mais ce qui, en temps de paix, est un fâcheux exemple. Dites, n'allez-vous pas nier tout cela?*
Human 1	"Ah! That's a new one," replied the king. "Are you going to tell me that your three damned musketeers, Athos, Porthos, and Aramis, and your cadet from Béarn, haven't sprung like furies on poor Bernajoux and injured him so badly that by now he's probably dead? Are you going to tell me that they didn't lay siege to the hôtel of the Duc de La Trémouille, and that they didn't try to burn it down? Of course, that might not be such a bad thing in times of war, since it's a nest of Huguenots—but in peacetime, it sets a bad example! Speak! Isn't this what you're going to tell me?"
Human 2	"Ah! here's something new!" the king picked up. "Are you going to tell me that your three damned musketeers, Athos, Porthos, and Aramis, and your lad from Béarn, didn't throw themselves like madmen on poor Bernajoux and mistreat him so badly that he's probably on his way out of this world as I speak? Are you going to say that after that they didn't lay siege to the hôtel of the duc de La Trémouille, and that they didn't intend to burn it? Which would perhaps have been no great misfortune in time of war, seeing that it's a nest of Huguenots, but which, in time of peace, is a regrettable example. Tell me, are you going to deny all that?"
Unfiltered (all)	"Ah! this is new," resumed the king. "Do you not say that your three damned musketeers, Athens, Porthos, and Aramis, and your Cadet de Béarn, did not rush like madmen upon poor Bernajoux, and did not so badly treat him that he is probably dying at this hour? Do you not say that they did not siege the Duke de La Trémouille's house, and did not want to burn him! which would not have been a great misfortune in war, since it is a nest of Huguenots, but which in peace is a dreadful example. Say, do you not deny all this?"
Medium+high (all)	"Ah! this is something new," continued the king. "Don't you say that your three damned musket-boys, Athenos, Porthos, and Aramis, and your Cadet de Béarn, did not throw themselves like mad on poor Bernajoux, and not so ill-treated him that he is probably in the process of dying at this hour! Don't you say that they did not seize the Duke de La Trémouille's house, and that they did not want to burn him! This would perhaps not have been a great misfortune in war, since it is the nest of Huguenots, but which in peace is a nasty example. Say, will you not deny all this?"
Synthetic (Gemma3)	"Ah! that is new," resumed the king. "Don't you say that your three damned mousquetaires, Athens, Porthos, Aramis, and your cadet of Béarn, did not throw themselves like furious beasts upon poor Bernouch, and did not treat him so badly that he is probably on the point of passing through! Don't you say that they did not take siege on the Duke de La Trémouille's house, and that they did not want to burn him! which would not have been very bad in wartime, since it was a nest of Hunchbacks, but which, in peace, is a nasty example. Say, will you not deny all this?"
Synthetic (Mistral)	"Ah, this is a new thing," resumed the king. "Don't you say that your three damned musketeers, Athens, Porthos, and Aramis, and your scholar of Béarn, did not throw themselves like furious beasts upon poor Bernajoux, and did not treat him so badly that he is probably in the process of passing this hour! Don't you say that they did not seize the siege of the Duke de La Trémouille's house, and did not want to burn him! which would perhaps not have been a great misfortune in war, since it is a nest of Huguenots, but which, in peace, is a dreadful example. Say, will you not deny all this?"
Synthetic (Qwen3)	'Ah, that's nothing new,' the king went on. 'Don't you say that your three damned spies, Athenaeus, Porthos and Arimas, and your cadet of Benares, did not throw themselves like a madman upon poor Bernard, and did not treat him so badly that he is probably going to be passing through this hour! Don't you say that they did not take the Duke of La Trémoïlles' house, and did not want to burn it! What would have been a great misfortune in war, since it was a nest of Hunchbacks, but what is a shameful example in peace. Say, won't you hide all this?'

Table 12: Example system outputs for a source paragraph from *Les Trois Mousquetaires* (*The Three Musketeers*) by Alexandre Dumas.