Answers to Reviewer Questions

A. Answer to R3Q1: What are the contributions of GPT to the accuracy of TermMT?

Motivation and Approach. First, we need to emphasize that the core idea of our approach is to design metamorphic rules (MRs). ChatGPT is only used to improve the accuracy of translation alignment as one step to implement our metamorphic rules: as Section 3.4.4 explains, some false positives in the error reports arise from factors like alignment tool errors when extracting term translations. We use GPT to filter the detected error reports and address this.

To evaluate GPT's contribution to TermMT's accuracy, we designed an ablation experiment comparing the accuracy and number of detected errors with and without GPT. We removed the GPT error filtering step in TermMT and detected errors in translation systems and the original dataset. To check the accuracy of our method TermMT without GPT, we randomly selected 100 errors from Google Translate output under each MR (totaling 300 errors across three MRs), which were manually inspected by two annotators, resolving any conflicts through discussion.

Results. Table I illustrates the impact on error detection after removing the GPT filtering step. TermMT represents the original method, while TermMT_woGPT is the method **without** the GPT filtering step. From Table I, we observe that the number of errors detected on Google Translate, Bing Microsoft Translator, and mBART increased from 3,765, 2,351, and 6,011 to 5,467, 3,879, and 9,211, respectively, i.e., TermMT_woGPT could **increase the number of detected errors**.

TABLE I
THE ERRORS DETECTED ACROSS DIFFERENT TRANSLATION SYSTEMS OF TERMMT AND TERMMT WITHOUT GPT FILTRATION

	Method	MR1	MR2	MR3	Total
Google	TermMT	441	239	3169	3765
	TermMT_woGPT	2127	330	3191	5467
Bing	TermMT	413	155	1837	2351
	TermMT_woGPT	1956	200	1857	3879
mBART	TermMT	2513	1596	2557	6011
	TermMT_woGPT	5491	2198	2659	9211

Table II compares the precision before and after ablation with a Cohen's kappa value of 71.69% for TermMT_woGPT. The precision of TermMT_woGPT decreases to 65.67% from the original 83.00%, indicating that while the GPT filtering step effectively improves accuracy, TermMT still maintains a reasonable level of accuracy without it. We conclude that GPT is not the primary contributor to TermMT's accuracy.

TABLE II
THE PRECISION VALUES (%) OF TERMMT AND TERMMT WITHOUT GPT FILTRATION FROM THE MANUAL CHECK ON 300 OUTPUTS OF GOOGLE TRANSLATE

	MR1	MR2	MR3	Total
TermMT	83.00	95.00	71.00	83.00
TermMT_woGPT	43.00	92.00	62.00	65.67

> Answer to **R3Q1**. The accuracy of TermMT can be enhanced with GPT filtering, but even without this step, TermMT maintains a high level of accuracy. Thus, GPT is not the primary contributor to TermMT's accuracy.

B. Answer to R3Q2: How much does TermMT rely on the language model for similarity computation, and can a better language model improve the performance?

Motivation and Approach. As mentioned in Section 3.4, after obtaining the translation results, we detect errors by calculating the term similarity between the original sentence translations and the mutated sentence translations. Thus, using a language model for similarity computation is a required part of our error detection process. The reviewer suggests that the language model could be directly applied to the original sentences and their translations, while in TermMT, it is used specifically for detecting errors through MR.

To evaluate TermMT's reliance on the language model for similarity calculations, we reproduce the approach mentioned by reviewer 3, which directly calculates the similarity between the original sentence and the translation results, termed Language Model Detect (LMD), and compare it with TermMT. For implementation, we choose the multilingual version of the Sentence-BERT[1] language model. Since the comparison requires manual annotation and due to the limited experimental time during the rebuttal period, we conduct experiments on the Google Translate dataset used in Section 5.2 (i.e., 300 original-mutated sentence pairs randomly selected for 3 MRs) . It is worth noting that this dataset serves as the benchmark for error reporting on TermMT, leading to a much higher error

[1]Sentence-BERT: https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

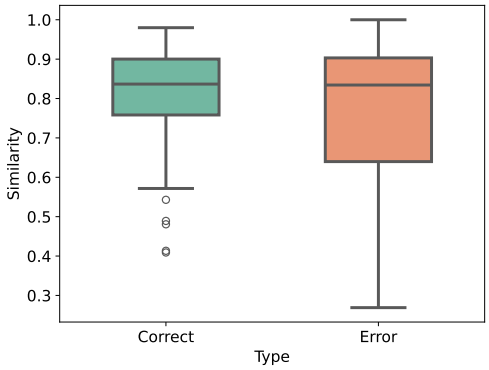

Fig. 1. Box Plot of Sentence BERT Similarity Scores for Correct and Error Translations in the Manual Dataset

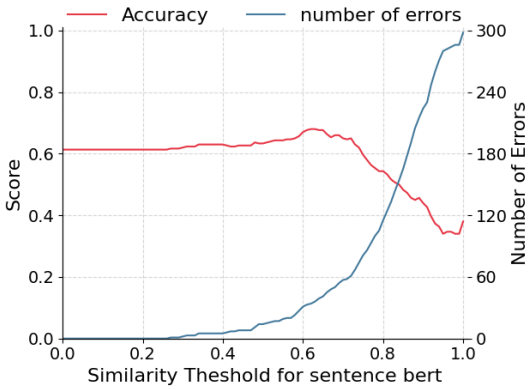

Fig. 2. The Impact of Similarity Threshold when Testing Google Translate with LMD

rate than usual. Consequently, re-labeling and calculating similarity will be more advantageous to LMD. Since LMD does not involve mutated sentences, we only consider the original sentence, its translation, and the reference translation when re-annotating the data to determine translation errors.

Furthermore, to address whether a better language model can improve performance, we use Sentence-BERT to replace the original language model BGE in TermMT, creating TermMT_SBERT, and detecte errors in the Google Translate translation results. From the view of model size, we argue that Sentence-BERT is a less complex model than BGE, which may lead to worse performance. We then sample 100 manual annotations from these errors to calculate the precision.

Results. Figure 1 presents the box plot results of LMD's similarity calculations on the artificial dataset, with a Cohen's Kappa coefficient of 76.84%. The medians for correct and incorrect data are relatively close, indicating LMD's low accuracy in identifying errors. Figure 2 illustrates the accuracy and number of detected errors by LMD as the threshold changes, i.e., the similarities less than the threshold are considered errors. Following the threshold selection strategy in Section 5.3, we set the similarity threshold at 0.63, achieving the highest accuracy 68.00% for LMD. Consequently, LMD detects 36 errors, only 12% of the 300 errors detected by TermMT, with the remaining 88% being unique errors identified by TermMT.

Table III shows the number of errors detected after replacing BGE with Sentence BERT. The "Total" column represents the total number of errors detected, "Pred" is the precision value, and "TP" is the estimated number of true positives, calculated as Total×Prec. "TermMT_SBERT" refers to the method that substitutes BGE with Sentence BERT. With Sentence BERT, TermMT_SBERT detects 5,417 errors on Google Translate, with a manual verification precision of 49% and a Cohen's Kappa coefficient of 65.50%, resulting in 2,654 TPs. The results indicate that while the number of detected errors increases with Sentence BERT, the accuracy decreases slightly. However, the difference in the number of true positives is only about 470, demonstrating that changing the similarity calculation model has minimal impact on TermMT's ability to detect real errors.

TABLE III
THE ERROR DETECTION RESULT OF TERMMT AND TERMMT WITH SENTENCE BERT

	MR1	MR2	MR3	Total	Prec	TP
TermMT	441	239	3169	3765	83%	3124
TermMT_SBERT	287	234	4967	5417	49%	2654

Answer to **R3Q2**. Compared to using language models directly for error detection, TermMT can detect 88% unique errors. Additionally, TermMT does not heavily rely on specific language models for similarity calculations, and the choice of different language models has negligible impact on TermMT's ability to detect real errors.

C. Answer to R3Q3: Compare the TermMT with reference-based method.

Motivation and Approach. The reviewer mentions that we did not compare our tool with any reference-based or general testing methods. TermMT is designed for large-scale testing of MT systems' terminology translation capabilities and does **not rely on reference translations**, which is the core advantage of our methods to solve "no test oracle" problem. Therefore, a fair comparison between our method and reference-based methods is challenging.

Given the time constraints of the rebuttal period and the high annotation cost, we follow the comparison approach with LMD in R3Q2. We perform error detection on the same re-annotated Google Translate datasets with 300 items in R3Q2 using reference-based methods. For the comparison, we select the state-of-the-art reference-based

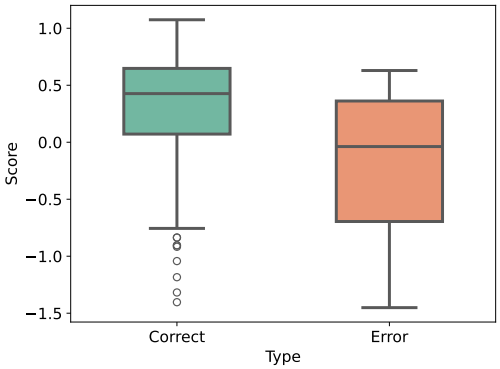

Fig. 3. Box Plot of UniTE Scores for Correct and Error Translations in the Manual Dataset

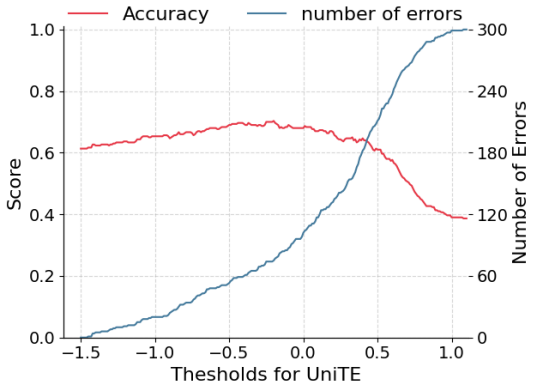

Fig. 4. The Impact of Score Threshold when Testing Google Translate with UniTE

translation evaluation method, UniTE: Unified Translation Evaluation, which was published in ACL 2022.

Results. The box plot results in Figure 3 indicate that UniTE can somewhat match the correctness of translation results. There is a noticeable gap between the median UniTE values of incorrect and correct entries; however, the main distribution of these values significantly overlaps, and there are numerous outliers among the correctly translated samples. This suggests that UniTE struggles to accurately distinguish between correct and error translations.

Figure 4 illustrates the accuracy and number of detected errors by UniTE as the threshold changes. Following the threshold selection strategy in Section 5.3, we set the similarity threshold at -0.2, achieving UniTE's highest accuracy of 73.33%. Ultimately, UniTE detects 75 errors, which is only 25% of the 300 errors identified by TermMT, leaving 75% of the errors uniquely detected by TermMT. This demonstrates that our method serves as a valuable complement to UniTE in error detection.

Answer to **R3Q3**. Compared to the state-of-the-art reference-based translation testing method, UniTE, TermMT can detect 75% unique errors, demonstrating that our method effectively complements existing reference-based testing methods.