HiMATE: A Hierarchical Multi-Agent Framework for Machine Translation Evaluation

Anonymous ACL submission

Abstract

The advancement of Large Language Models (LLMs) enables flexible and interpretable automatic evaluations. In the field of machine translation evaluation, utilizing LLMs with translation error annotations based on Multidimensional Quality Metrics (MQM) yields more human-aligned judgments. However, current LLM-based evaluation methods still face challenges in accurately identifying error spans and assessing their severity. In this paper, we propose HiMATE, a **Hi**erarchical **M**ulti-**A**gent Framework for Machine **T**ranslation **E**valuation. We argue that existing approaches inadequately exploit the fine-grained structural and semantic information within the MQM hierarchy. To address this, we develop a hierarchical multi-agent system grounded in the MQM error typology, enabling granular evaluation of subtype errors. Two key strategies are incorporated to further mitigate systemic hallucinations within the framework: the utilization of the model's self-reflection capability and the facilitation of agent discussion involving asymmetric information. Empirically, HiMATE outperforms competitive baselines across different datasets in conducting human-aligned evaluations. Further analyses underscore its significant advantage in error span detection and severity assessment, achieving an average F1-score improvement of 89% over the best-performing baseline. We make our code and data publicly available at `https://anonymous.4open.science/r/HiMATE-Anony`.

1 Introduction

Translation capability reflects the cross-lingual comprehension and generation quality of Large Language Models (LLMs) (Hendy et al., 2023; Yang et al., 2024; Dubey et al., 2024). Compared to traditional machine translation models, applying LLMs significantly enhances the translation capabilities of machine translation systems and introduces new challenges to Machine Translation Evaluation (MTE) (Kocmi et al., 2024).

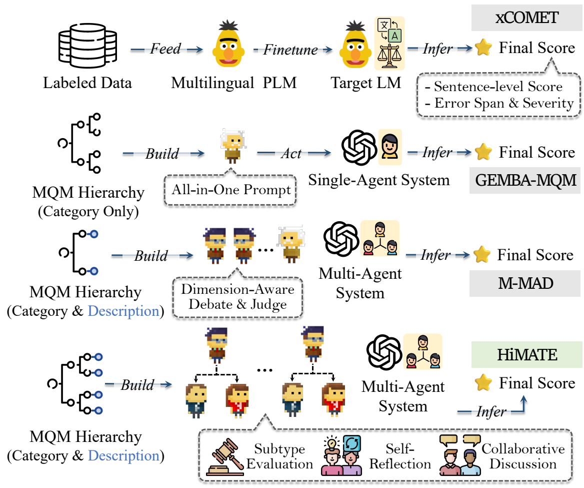

Figure 1: A comparison of HiMATE and other representative MQM-based metrics for MTE.

Previously, statistics-based MTE metrics such as BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005) are utilized for their efficiency, despite inherent limitations in accurately capturing semantic similarity. To address this gap, learned metrics such as COMET-20 (Rei et al., 2020) and RoBLEURT (Wan et al., 2021) have been developed to enhance evaluations. However, these approaches primarily yield numeric scores and cannot identify the specific error spans in translations (Fernandes et al., 2023; Leiter et al., 2022).

The machine translation data, annotated based on the guidelines of Multidimensional Quality Metrics (MQM),[1] includes detailed information regarding error spans, categories, and severity for each translated sentence, which facilitates more precise finetuning and thorough analysis of MTE models. The xCOMET model (Guerreiro et al., 2024) is finetuned on such data for interpretable scores, while it necessitates resource-intensive multi-stage training and shows a lack of generalization abil-

[1]themqm.org.

ity. The LLM-based single-agent evaluation systems, including EAPrompt (Lu et al., 2024) and GEMBA-MQM (Kocmi and Federmann, 2023a), utilize MQM-informed prompts to guide LLMs in error detection. Nevertheless, the hallucination and position bias inherent within LLM evaluators affect their performance (Wang et al., 2024). In response to these challenges, the recently proposed method M-MAD (Feng et al., 2024) employs a multi-agent system to decompose complex MTE tasks into simpler assessments, showcasing the potential of this paradigm. However, its effectiveness primarily stems from the self-consistency of agents operating within identical contexts, which may limit nuanced, multi-level error analysis. We argue that the rich semantic and hierarchical structure of the MQM framework has been largely underutilized in current methodological designs. By developing a more fine-grained, hierarchical agent-based system explicitly tailored to leverage the untapped potential of MQM's structured hierarchy, error detection accuracy could be significantly enhanced, thereby narrowing the gap with human expert performance.

In this paper, we propose HiMATE, a **Hi**erarchical **M**ulti-**A**gent Framework for Machine **T**ranslation **E**valuation. Specifically, we establish the agent topology based on the MQM error typology, facilitating hierarchical agent nodes to exchange information in alignment with human error classification methodology. Error descriptions associated with each node are then employed to initialize multi-agent evaluators, allowing them to focus on detailed, simplified discrimination tasks rather than evaluations against broad, comprehensive criteria. This design facilitates fine-grained subtype error evaluations through multi-agent collaboration. Furthermore, we propose two post-stages to enhance the accuracy of the system in error detection and severity assessment. Concretely, the subtype evaluator first revises the original translations based on its error detection results, then re-evaluates the validity of the initial judgments through comparison. Cases with low confidence during this self-reflection phase are forwarded to the next stage, where agents from different tiers collaboratively determine the presence and severity of translation errors through iterative discussion. In contrast to prior methods, the proposed method leverages MQM information more effectively for decomposing complex tasks and constructing hierarchical multi-agent systems. By guiding these agents through self-reflection and collaboration, we achieve more human-aligned automatic evaluations. Figure 1 highlights the key differences between HiMATE and other representative MTE metrics.

We summarize our contributions as follows:

- We propose HiMATE, a novel multi-agent machine translation evaluation framework that leverages MQM hierarchy-derived structural and semantic information to define agent topology, resulting in reliable, human-aligned judgments.
- We develop a three-stage strategy to enhance error detection and severity assessment within the framework, facilitated by self-reflection and collaborative discussions among agents.
- Empirical results across various models and datasets demonstrate the effectiveness of HiMATE. Further analyses highlight its superiority in both error span identification and severity evaluation, achieving an average 89% improvement in F1-score and 95% enhancement in recall over the best-performing baseline.

2 Related Work

2.1 Machine Translation Evaluation

The objective of MTE is to assess the quality of machine-translated text. Traditional statistics-based metrics (Papineni et al., 2002; Lin, 2004; Zhang et al., 2019) often inadequately capture semantic similarity (Rei et al., 2020). Certain learned metrics (Rei et al., 2022a,b; Guerreiro et al., 2024) have shown improvement, while high iteration costs and limited generalization ability constrain these methods. Recent advancements harness the power of LLMs for translation evaluation. GEMBA-DA (Kocmi and Federmann, 2023b) implements direct scoring through prompt tuning, whereas EAPrompt (Lu et al., 2024) identifies errors without numerical scoring to mitigate bias. AUTOMQM (Fernandes et al., 2023) and GEMBA-MQM (Kocmi and Federmann, 2023a) prompt models to produce a score within the MQM framework. However, these methods insufficiently utilize the hierarchical information of MQM framework.

2.2 Evaluation based on Multi-Agent

Multi-agent systems have demonstrated their potential in enhancing the accuracy of evaluation within Nature Language Generation (NLG). ChatEval (Chan et al., 2023) employs LLMs in a framework similar to human group discussions. MATEval (Li et al., 2024) introduces architectures that

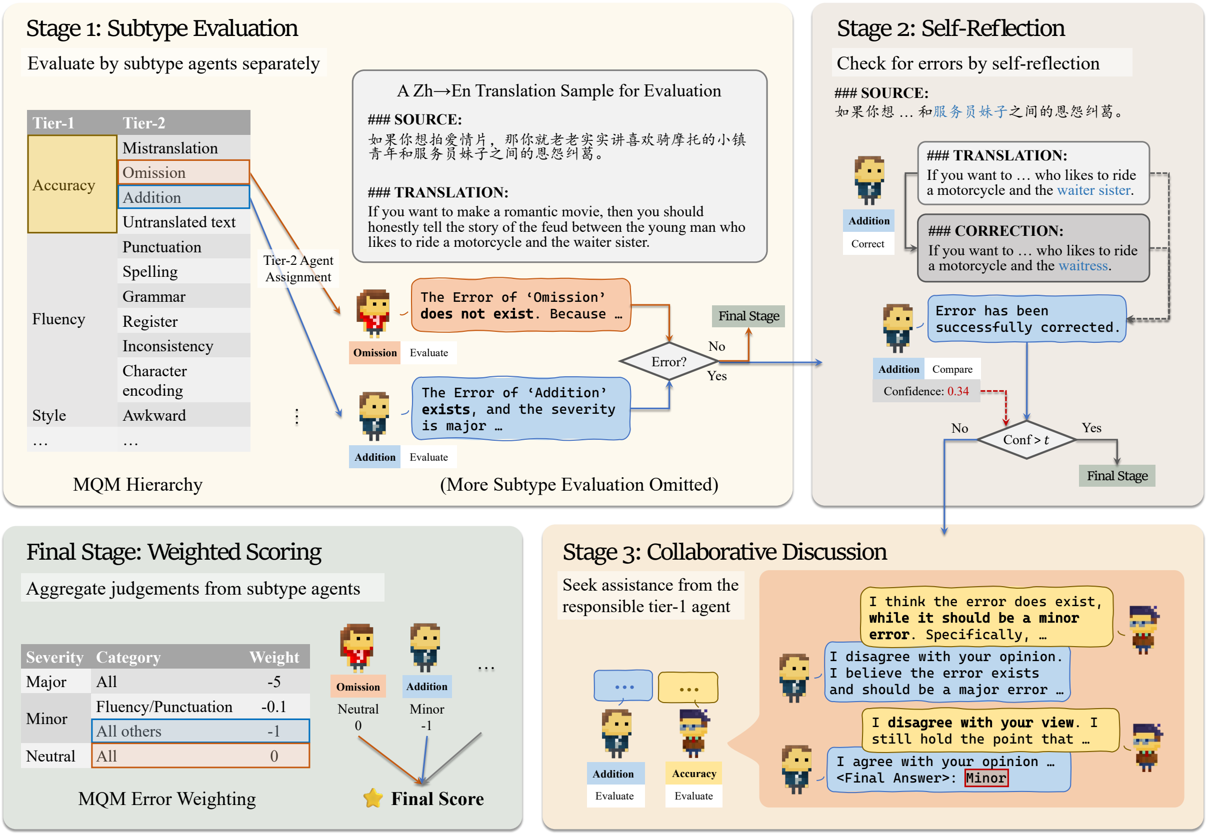

Figure 2: An overview of the proposed framework HiMATE where translation errors are categorized into tier-1 high-level errors and tier-2 fine-grained subtype errors. The tier-2 agent first evaluates a subtype error. Then the judgment is refined through the self-reflection stage. If needed, the tier-2 agent discusses with its upper-level tier-1 agent for further confirmation. The final score is computed by summing the weighted values of errors.

rely on multi-round discussion and summarization for evaluation. Similarly, a courtroom-inspired architecture has been proposed, utilizing LLMs as adversarial judges engaging in multi-round debates for NLG assessment (Bandi and Harrasse, 2024). M-MAD (Feng et al., 2024) integrates multi-agent systems into MTE, leveraging the collaborative reasoning capabilities of LLMs. Nevertheless, existing approaches for MTE, such as M-MAD, insufficiently leverage fine-grained error definitions and semantic information within the MQM hierarchy.

3 Methodology

In this section, we introduce our proposed framework HiMATE for conducting human-aligned machine translation evaluations. Adhering to the hierarchical structure of the MQM error typology, HiMATE organizes translation errors into two tiers: high-level error categories (tier-1) and fine-grained error subtypes (tier-2). Two distinct types of agents are assigned to manage these tiers, with each agent specializing in assessing specific error categories, as depicted in Figure 2. Initially, tier-2 agents individually perform evaluations on their corresponding error subtypes. This preliminary assessment is then refined by harnessing the self-reflection capabilities of LLMs. Finally, results exhibiting low confidence are subjected to collaborative discussion involving tier-1 and tier-2 agents. The entire evaluation process operates without reliance on reference, ensuring broad applicability for evaluation.

3.1 Agent Topology

The hierarchical error structure of the evaluation framework comprises high-level core error types e_i (e.g., Accuracy), which are allocated to tier-1 agents $A_i^{[1]}$, and fine-grained error subtypes e_j (e.g., Omission, Addition) that belong under these core errors and are assigned to tier-2 agents $A_j^{[2]}$. Specific descriptions are provided in Appendix A.

3.2 Subtype Evaluation

Initially, the tier-2 agents are required to seek the specific errors individually within the translations. For a given source text x, translation y, and system prompt p_{SE}, each agent $A_j^{[2]}$ is provided with distinct error definitions corresponding to tier-2, enabling them to leverage detailed information effectively. All the agents adhere to a uniform definition of severity when evaluating identified errors.

Upon completion of this stage, a preliminary assessment concerning the error and the corresponding severity is produced. Evaluators also provide explanations supporting their judgments. The initial evaluation information I_{SE} generated above is then passed on to the subsequent stage:

$$\text{I}_{\text{SE}} \leftarrow A_j^{[2]}(x, y, p_{\text{SE}}) \tag{1}$$

If no errors are detected, results proceed directly to the final stage; otherwise, sentences flagged with errors advance to the subsequent stage.

3.3 Self-Reflection

To enhance the robustness of the preliminary evaluation, we further validate the initial assessment by harnessing the self-reflection capabilities of LLMs. The prompt p_{SR} consists of two components, p_{c} for error correction and p_{v} for comparative verification. Initially, the tier-2 agents scrutinize all error annotations and justifications from the preliminary assessment. Utilizing the information I_{SE}, tier-2 agents generate the corrected versions y_{c}:

$$y_{\text{c}} \leftarrow A_j^{[2]}(x, y, p_{\text{c}}, \text{I}_{\text{SE}}) \tag{2}$$

Subsequently, the evaluators compare the original translation y with its corrected version y_{c}. Suppose an error is accurately identified during the initial phase, the correction should either rectify the issue or mitigate its severity, thereby validating the initial judgment. Otherwise, this may indicate a false positive in the previous evaluation, suggesting that the purported error might not exist. This verification mechanism yields more precise error judgments, diminishing workload concerning error type and severity in subsequent stages. Following the stage, additional confirmation regarding error information I_{SR} can be procured:

$$\text{I}_{\text{SR}} \leftarrow A_j^{[2]}(y, y_{\text{c}}, p_{\text{v}}), \text{I}_{\text{SE}} \tag{3}$$

Responses with confidence scores, calculated by summing token logarithmic probabilities, below a predefined stage-transition threshold advance to the subsequent phase, whereas those surpassing this threshold proceed directly to the final stage.

3.4 Collaborative Discussion

The final stage is designed to validate the ambiguous assessment through collaborative discussion. This methodology stems from the observation that judgments with low confidence often indicate potential inaccuracies requiring expert validation. By focusing specifically on these instances, HiMATE achieves enhanced precision without compromising computational efficiency.

In this phase, the tier-1 agents $A_i^{[1]}$, responsible for high-level error categorization, collaborate and engage in discussions with the tier-2 agents. These tier-1 agents possess comprehensive knowledge of all tier-2 subcategory error definitions within their respective domains, enabling them to evaluate translation from a broader perspective.

Throughout the discussion period, tier-1 agents first analyze the information, initially set as I_{SR} at the start of the discussion, from tier-2 agents. Subsequently, tier-1 agents provide expert evaluations, integrated with the prior chat history to form an updated discussion record. Tier-2 agents then contribute their perspectives, indicating agreement or disagreement with the assessments provided by tier-1 agents, thereby further updating the discussion record. The discussion continues iteratively until consensus is reached or the predefined maximum number of discussion rounds is exceeded. The final evaluation outcome is derived from the history of the concluding conversation.

3.5 Weighted Scoring

Commencing with evaluation on subtype errors, HiMATE refines judgments through a self-reflection process to validate identified errors, eventually engaging in a collaborative discussion to reach consensus. After these three stages, the evaluation outcome for the translation sentence pair (x, y) regarding subtype error e_j is obtained, consisting of the severity label s_j and its corresponding weight w_j. The weights associated with each error subtype and severity level are detailed in Appendix B. The score S can be calculated as:

$$\text{S} = -\sum w_{\text{j}} \tag{4}$$

4 Experiments

4.1 Settings

Dataset We conduct our main experiments on the MQM22 dataset (Freitag et al., 2021). Due

Table 1: Results of MTE methods on the ZH-EN and EN-DE subsets of the MQM22 dataset. A higher Kendall's correlation coefficient (τ) and Spearman's correlation coefficient (s), as well as a lower Mean Absolute Error (MAE) and Mean Squared Error (MSE), indicate better alignment with human evaluations. The abbreviations *4o-mini* and *qwen2.5* refer to the gpt-4o-mini and qwen2.5-72b-instruct models, respectively. Reference-based methods are indicated with a gray background. The best result in each column is **bolded**, and the second-best is underlined.

Model	Methods	ZH-EN				EN-DE			
		$\tau\uparrow$	$s\uparrow$	$MAE\downarrow$	$MSE\downarrow$	$\tau\uparrow$	$s\uparrow$	$MAE\downarrow$	$MSE\downarrow$
	BLEU	0.172	0.233	0.6849	0.5064	0.161	0.213	0.6184	0.4322
	BERTSCORE	0.300	0.400	0.3334	0.1349	0.208	0.275	0.3936	0.1970
	COMET-22	0.369	0.489	0.1527	0.0306	<u>0.301</u>	**0.390**	0.0948	0.0144
	COMETKIWI	0.365	0.487	0.1821	0.0386	0.212	0.277	0.1351	0.0221
	xCOMET-XL	0.394	0.515	0.1210	0.0317	0.283	0.355	0.0567	0.0115
4o-mini	GEMBA-MQM	0.381	0.464	0.0755	0.0131	0.270	0.316	0.0420	0.0052
	M-MAD	0.350	0.435	0.0379	0.0033	0.282	0.327	**0.0164**	**0.0007**
	HiMATE (ours)	0.404	0.502	0.0421	0.0051	0.293	0.338	0.0250	0.0026
qwen2.5	GEMBA-MQM	0.407	0.510	0.0419	0.0039	0.229	0.279	0.0428	0.0045
	M-MAD	0.297	0.365	<u>0.0323</u>	**0.0027**	0.227	0.266	0.0286	0.0018
	HiMATE (ours)	0.413	0.498	**0.0281**	<u>0.0028</u>	0.248	0.274	0.0198	0.0016
qwen-max	GEMBA-MQM	<u>0.417</u>	<u>0.527</u>	0.0873	0.0159	0.292	0.347	0.0471	0.0071
	M-MAD	0.350	0.425	0.0417	0.0050	0.216	0.253	0.0258	0.0017
	HiMATE (ours)	**0.425**	**0.531**	0.0386	0.0045	**0.318**	<u>0.358</u>	<u>0.0187</u>	<u>0.0014</u>

to cost constraints, we randomly select the outputs from the HuaweiTSC system for Chinese-English (ZH-EN) and the comet-bestmbr system for English-German (EN-DE), including 1,875 and 1,315 annotated translations, respectively. The human-annotated MQM scores serve as the ground truth for evaluating machine translation quality.

To mitigate potential data leakage risks due to overlaps between LLM pre-training corpora and evaluation datasets, we also perform experiments on the recently released MQM24 dataset. Experimental results are provided in Appendix C.

Baseline The reference-based evaluation metrics, including BLEU (Papineni et al., 2002), BERTSCORE (Zhang et al., 2019), and COMET-22 (Rei et al., 2022a), as well as the reference-free ones, such as COMETKIWI (Rei et al., 2022b), xCOMET-XL (Guerreiro et al., 2024), GEMBA-MQM (Kocmi and Federmann, 2023a), and M-MAD (Feng et al., 2024), are adopted as baselines for comparison. Among these, GEMBA-MQM and M-MAD are LLM-based methods like ours.

Evaluation Method Four widely adopted metrics are employed to evaluate the alignment of HiMATE with human judgments. Specifically, we use Kendall's correlation coefficient τ and Spearman's correlation coefficient s to measure correlation, while using Mean Absolute Error (MAE) and Mean Squared Error (MSE) to quantify similarity.

Experiment Settings The gpt-4o-mini (Hurst et al., 2024), qwen2.5-72b-instruct, and qwen-max (Yang et al., 2024) are adopted as backbone models for LLM-based MTE methods. The error typology used in HiMATE is categorized and further refined into 5 core error types and 19 distinct subtypes. The temperature parameter is set to 0. Several instances from the MQM20 and MQM21 datasets are selected to serve as 2-shot demonstration examples for the subtype evaluation stage. Stage-transition thresholds t are adaptively configured for different models (refer to Appendix D for specific procedures). The maximum number of dialogue turns during the collaborative discussion stage is set to four, with each agent allowed at most two statements.

4.2 Main Results

Table 1 illustrates the experimental results of different MTE methods on the MQM22 dataset. According to the results, HiMATE consistently exhibits

Table 2: Ablation study of HiMATE across different backbone models. The abbreviations *4o-mini* and *qwen2.5* refer to the gpt-4o-mini and qwen2.5-72b-instruct models, respectively. SE, SR, and CD respectively denote the subtype evaluation, self-reflection, and collaborative discussion stages. SA denotes the single-agent baseline method. The best result for each metric within the same backbone model group is **bolded**.

Model	Methods	ZH-EN				EN-DE			
		$\tau\uparrow$	$s\uparrow$	$MAE\downarrow$	$MSE\downarrow$	$\tau\uparrow$	$s\uparrow$	$MAE\downarrow$	$MSE\downarrow$
4o-mini	SA	0.345	0.417	0.0427	**0.0042**	0.210	0.230	**0.0166**	**0.0008**
	HiMATE-SE	0.359	0.468	0.1614	0.0614	0.278	0.329	0.0698	0.0205
	-SE+SR	0.370	0.477	0.1178	0.0362	0.283	0.332	0.0497	0.0106
	-SE+SR+CD	**0.404**	**0.502**	**0.0421**	0.0051	**0.293**	**0.338**	0.0250	0.0026
qwen2.5	SA	0.341	0.413	0.0356	0.0031	0.183	0.200	**0.0197**	0.0017
	HiMATE-SE	0.397	0.489	0.0572	0.0172	0.240	0.267	0.0341	0.0100
	-SE+SR	0.406	0.496	0.0479	0.0110	0.245	0.271	0.0281	0.0054
	-SE+SR+CD	**0.413**	**0.498**	**0.0281**	**0.0028**	**0.248**	**0.274**	0.0198	**0.0016**
qwen-max	SA	0.395	0.480	**0.0374**	0.0049	0.286	0.325	0.0200	**0.0014**
	HiMATE-SE	0.409	0.523	0.1047	0.0405	0.292	0.333	0.0436	0.0140
	-SE+SR	0.409	0.517	0.0925	0.0326	0.290	0.329	0.0354	0.0085
	-SE+SR+CD	**0.425**	**0.531**	0.0386	**0.0045**	**0.318**	**0.358**	**0.0187**	**0.0014**

improvements across different evaluation metrics and backbone models. Stable enhancements in correlation and similarity metrics are observed for both ZH-EN and EN-DE translation tasks.

Specifically, HiMATE showcases improved alignment with human evaluations, as evidenced by correlation coefficients. For ZH-EN, when utilizing the qwen-max model, HiMATE achieves peak correlation values of 0.425 in Kendall's correlation coefficient and 0.531 in Spearman's correlation coefficient. This advantage is sustained across diverse backbone models, with correlation metrics generally outperforming LLM-based methods. Similar improvements are observed on the EN-DE dataset, where HiMATE surpasses other LLM-based methods, achieving the highest τ and second-best s. On the other hand, HiMATE achieves either the best or second-best performance in MAE and MSE, showing comparable evaluation similarity with M-MAD and markedly exceeding other baseline methods. The significant improvement of similarity with human evaluation suggests the effectiveness of aligning human-built evaluation criteria within the proposed method. In addition, HiMATE demonstrates consistently strong performance with different backbone models, highlighting its robustness to changes in model capacity. All of these observations serve as compelling evidence of the efficacy of HiMATE in conducting high-quality evaluations.

5 Ablation and Analysis

5.1 Ablation Study

To verify the contribution of each stage within HiMATE, we conduct an ablation study of the framework. Concretely, we introduce a single-agent evaluation method (denoted as SA) as a baseline, which conducts an all-in-one evaluation encompassing all subtype errors from the subtype evaluation stage in HiMATE. Evaluation results of the ablated framework on the MQM22 dataset are shown in Table 2. According to the results, HiMATE-SE consistently outperforms SA in terms of correlation, particularly improving Kendall's correlation coefficient τ from 0.341 to 0.397 when using a qwen2.5 backbone model on the ZH-EN subset, which underscores the importance of fine-grained subtype error evaluation through a multi-agent approach. The performance improvements from the self-reflection stage are relatively modest; however, this stage helps reduce the computational cost of the subsequent stage by filtering the high-confidence, reliable judgments. The collaborative discussion stage contributes more substantially, achieving the best or second-best results across all four evaluation metrics. It is worth noting that the performance improvements from collaborative discussion vary across different backbone LLMs, likely influenced by instruction-following and divergent thinking capability of the model.

Table 3: Evaluation results of different MTE methods on MQM22 ZH-EN under different sentence lengths. We use gpt-4o-mini as the backbone model. The threshold for separating sentences is set to 12 and 27.

	short			medium			long		
	τ	s	F1-score	τ	s	F1-score	τ	s	F1-score
GEMBA-MQM	**0.369**	0.433	0.085	0.382	0.460	0.093	0.344	0.427	0.100
M-MAD	0.333	0.386	0.124	0.316	0.392	0.108	0.311	0.408	0.116
HiMATE	0.368	**0.436**	**0.218**	**0.425**	**0.529**	**0.260**	**0.377**	**0.491**	**0.294**

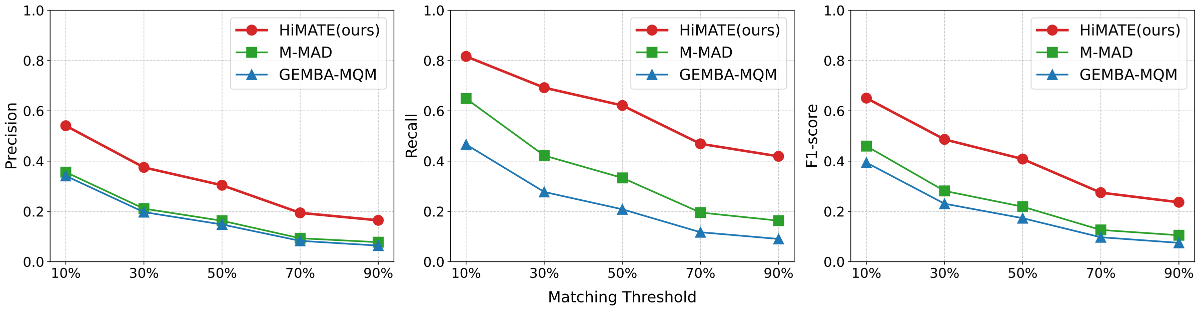

Figure 3: Error span detection results for different MTE methods on the MQM22 ZH-EN dataset, based on gpt-4o-mini. We report Precision, Recall, and F1-score across varying matching thresholds.

5.2 Error Span Detection

As mentioned earlier, current LLM-based MTE methods struggle with accurately identifying translation error spans. In this section, we aim to evaluate the error span detection performance of HiMATE and several baseline approaches. Our experiments are conducted on the MQM22 ZH-EN subset, using gpt-4o-mini as the backbone model. Following Ye et al. (2024), we utilize Precision, Recall, and F1-score for evaluation. Figure 3 illustrates the evaluation results at varying matching thresholds, which are defined by the minimum overlap ratio between detected error spans and gold-standard annotations. According to the results, HiMATE consistently surpasses GEMBA-MQM and M-MAD across all matching thresholds on all three metrics. Notably, compared to the best-performing baseline M-MAD, HiMATE achieves average improvements of approximately 89% in F1-score and 95% in Recall, demonstrating its superior effectiveness in error span detection. These results further validate the effectiveness of HiMATE's hierarchical agent design. More details regarding the error span matching algorithm are provided in Appendix E.

5.3 Performance on Distinct Sentence Length

Longer contexts present significant challenges to both machine translation quality and evaluation accuracy. To investigate how sentence length affects MTE performance, we conduct experiments on the ZH-EN subset of MQM22, grouping samples by token count. As shown in Table 3, HiMATE consistently outperforms baselines across all sentence-length groups. While baseline methods achieve comparable results on shorter samples, HiMATE maintains superior performance as sentence length increases. In contrast, GEMBA-MQM and M-MAD exhibit limited or declining accuracy with rising complexity. Notably, at a matching threshold of 50%, the F1-score gap between HiMATE and the best-performing baseline widens from 0.094 for short sentences to 0.194 for long sentences. These findings demonstrate HiMATE's robustness in accurately identifying translation errors, particularly in longer sentences.

5.4 Analysis on Domain-Specific Evaluation

The MQM22 dataset encompasses four distinct text domains: news, social, conversation, and ecommerce. We present an analysis of the performance of different MTE methods across these domains using gpt-4o-mini as the backbone model. As illustrated in Table 5, HiMATE significantly surpasses the other two LLM-based methods in the conversation, social, and ecommerce domains. Although M-MAD slightly outperforms HiMATE in the news domain, it exhibits the least effective performance in the other three domains. We also conduct experiments using other metrics and datasets, the details of which can be found in Appendix F. Over-

Table 4: A case of responses generated by gpt-4o-mini on MQM22 ZH-EN among different LLM-based methods. The parts marked in red indicate incorrect judgments, while the parts marked in green indicate correct judgments.

Source	希望你们了解一下
Translation	I hope you know about it.
Reference	I hope you can find out about it.
Annotation	Major-Mistranslation-'know about it'
GEMBA-MQM	Major-Mistranslation-'I hope you know about it', Minor-Awkward-'I hope you know about it'
M-MAD Stage1	Minor-Omission-'know about it'
M-MAD Stage2&3	Minor-Omission-'know about it'
HiMATE-SE	Minor-Inappropriate for context-'know', Major-Omission-'希望', Minor-Awkward-'I hope you know about it', Minor-Addition-'know about it', Major-Mistranslation-'know about it'
-SE+SR	Minor-Awkward-'I hope you know about it', Minor-Addition-'know about it', Major-Mistranslation-'know about it'
-SE+SR+CD	Major-Mistranslation-'know about it'

Table 5: Domain-specific results of various MTE methods on MQM22 ZH-EN, measured by Spearman's correlation coefficient s using gpt-4o-mini as backbone.

	news	conversation	social	ecommerce
GEMBA-MQM	0.333	0.367	0.471	0.538
M-MAD	**0.365**	0.347	0.430	0.444
HiMATE	0.355	**0.399**	**0.529**	**0.577**

all, these findings underscore the robustness of HiMATE across diverse textual contexts.

5.5 Case Study

Through specific case studies, we compare the evaluation results of baseline methods with our proposed framework in Table 4, highlighting the distinct behavior of HiMATE. As shown, while GEMBA-MQM correctly identifies the type and severity of one error, it inaccurately expands the error span to the entire sentence, making the assessment unreliable. Additionally, it falsely flags non-existent errors. Due to its single-agent, single-stage design, the system lacks the ability to refine or correct its judgments. In the case of M-MAD, while the system precisely pinpoints the error span, it inaccurately classifies the error type and severity, and fails to rectify the previous misjudgment in the subsequent multi-agent debate stage. In contrast, while HiMATE initially over-identifies errors during the subtype evaluation stage, its self-reflection and collaborative discussion stages effectively filter out these incorrect assessments while retaining valid judgments. This demonstrates HiMATE's effectiveness in allocating different tiers of hierarchical MQM information to distinct agents, enabling each to focus specifically on the error types under its responsibility, thereby enhancing the evaluation accuracy. More detailed case examples with staged evaluation processes across agents are provided in Appendix H.

6 Conclusion

This paper introduces HiMATE, a hierarchical multi-agent framework for LLM-based machine translation evaluation. The framework assembles a three-stage strategy to enhance error detection and severity assessment, leveraging self-reflection and collaborative discussion mechanisms after subtype evaluation. The experimental results based on correlation and similarity metrics demonstrate the superior performance of HiMATE. Ablation study further validates the contribution of each stage in aligning closely with human judgments. Detailed analyses highlight that HiMATE exhibits pronounced advantages in error span detection compared to existing LLM-based methods. Additionally, evaluations across varying sentence lengths indicate robust and consistent performance. HiMATE also achieves stable advancements across various text domains. The achievement of HiMATE emphasizes the significance of effectively utilizing the human evaluation framework in designing multi-agent collaborative evaluation strategies.

Limitations

Due to constraints in time and resources, the validation process presented in this study primarily focuses on selected models and language pairs, rather than an exhaustive evaluation across diverse models and languages. The current study does not evaluate the applicability of the proposed framework to advanced reasoning models, which have shown promising potential in comprehensive tasks. Additionally, the current experiment does not include the evaluation of the cross-lingual comprehension and generation quality of LLMs.

References

Chaithanya Bandi and Abir Harrasse. 2024. Adversarial multi-agent evaluation of large language models through iterative debates. *arXiv preprint arXiv:2410.04663*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

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

Zhaopeng Feng, Jiayuan Su, Jiamei Zheng, Jiahan Ren, Yan Zhang, Jian Wu, Hongwei Wang, and Zuozhu Liu. 2024. M-mad: Multidimensional multi-agent debate framework for fine-grained machine translation evaluation. *arXiv preprint arXiv:2412.20127*.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083, Singapore. Association for Computational Linguistics.

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

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

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

Tom Kocmi and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Christoph Leiter, Piyawat Lertvittayakumjorn, Marina Fomicheva, Wei Zhao, Yang Gao, and Steffen Eger. 2022. Towards explainable evaluation metrics for natural language generation. *arXiv preprint arXiv:2203.11131*.

Yu Li, Shenyu Zhang, Rui Wu, Xiutian Huang, Yongrui Chen, Wenhao Xu, Guilin Qi, and Dehai Min. 2024. Mateval: A multi-agent discussion framework for advancing open-ended text evaluation. In *International Conference on Database Systems for Advanced Applications*, pages 415–426. Springer.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Qingyu Lu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. 2024. Error analysis prompting enables human-like translation evaluation in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8801–8816, Bangkok, Thailand. Association for Computational Linguistics.

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

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

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

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yu Wan, Dayiheng Liu, Baosong Yang, Tianchi Bi, Haibo Zhang, Boxing Chen, Weihua Luo, Derek F. Wong, and Lidia S. Chao. 2021. RoBLEURT submission for WMT2021 metrics task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1053–1058, Online. Association for Computational Linguistics.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

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

Xiaojun Ye, Junhao Chen, Xiang Li, Haidong Xin, Chao Li, Sheng Zhou, and Jiajun Bu. 2024. MMAD:multi-modal movie audio description. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11415–11428, Torino, Italia. ELRA and ICCL.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A MQM Hierarchy

The error typology used in HiMATE is categorized and further refined into 5 core error types and 19 distinct subtypes. Considering the characteristics and practical requirements of the proposed framework, specific error definitions within MQM have been further refined without altering their original meanings. These refinements are based on the MQM-*Full Master File Instructions* and build upon the previous work (Freitag et al., 2021), as detailed in Table 8.

B MQM Error Weight

Following the architectural design and operational requirements of HiMATE, we eliminate the "non-translation" category and assign error weights that remain closely aligned with established human assessment practices (Freitag et al., 2021) for machine translation evaluation. The MQM error weights used for calculating the final score are shown in Table 6.

C Evaluation Experiment on MQM24

To eliminate potential data leakage risks arising from the overlaps between LLM pre-training corpora and evaluation datasets, we perform additional evaluation experiments on outputs from the IKUN system on the MQM24 EN-DE dataset, encompassing 487 annotated translations. We select gpt-4o-mini as the backbone model for this experiment, whose training data cut-off is prior to the release of the MQM24 dataset. The experimental results, presented in Table 7, closely align with the conclusions drawn from experiments conducted on the MQM22 dataset. On the one hand, HiMATE achieves excellent performance in terms of correlation and similarity. On the other hand, HiMATE demonstrates stable improvements across all evaluation stages. These findings are consistent with the primary experimental results, further validating the effectiveness and robustness of HiMATE in enhancing evaluation accuracy.

D Stage-Transition Threshold Acquisition

To determine an appropriate stage-transition threshold, several sentences unrelated to the primary experiment and not utilized as two-shot demonstrations are randomly selected from the MQM20 and MQM21 datasets, both for ZH-EN and EN-DE, forming the validation set. The size of this validation set is approximately one-tenth of the MQM22 dataset utilized in the main experiments.[2] Each sentence undergoes subtype evaluation and subsequent self-reflection to produce confidence scores. Subsequently, these confidence scores are ranked, and empirical analysis suggests the confidence score at the 60% percentile as the optimal stage-transition threshold. Threshold values vary across language pairs and models.

Table 6: MQM error weights used to calculate the final score in HiMATE.

Severity	Category	Weight
Major	all others	5
Minor	Fluency/Punctuation	0.1
	all others	1
Neutral	all	0

Table 7: Results of MTE methods on the EN-DE subset of the MQM24 dataset, where the backbone model of LLM-based methods is gpt-4o-mini. A higher Kendall's correlation coefficient (τ) and Spearman's correlation coefficient (s), as well as a lower Mean Absolute Error (MAE) and Mean Squared Error (MSE), indicate better alignment with human evaluations. Reference-based methods are indicated with a gray background. The best result in each column is **bolded**, and the second-best is underlined.

Methods	EN-DE			
	$\tau\uparrow$	$s\uparrow$	$MAE\downarrow$	$MSE\downarrow$
BLEU	0.050	0.071	0.1664	0.0394
BERTSCORE	0.109	0.153	0.0537	0.0095
COMET-22	0.245	0.343	0.1611	0.0376
COMETKIWI	0.190	0.263	0.0574	0.0099
xCOMET-XL	**0.344**	**0.467**	0.0940	0.0178
GEMBA-MQM	0.281	0.350	0.0563	0.0087
M-MAD	0.267	0.347	**0.0275**	**0.0029**
HiMATE-SE	0.254	0.344	0.0975	0.0322
HiMATE-SR	0.280	0.372	0.0732	0.0191
HiMATE-CD	0.321	0.424	0.0440	0.0072

[2]The validation set can be found in https://anonymous.4open.science/r/HiMATE-Anony

Algorithm 1 Matching Algorithm

Given: Gold-standard error span G, detected error span E, matching threshold $\theta \in (0, 1]$.
Tokenize $G \to G_t$, $E \to E_t$
$n \leftarrow |G_t|$, $m \leftarrow |E_t|$
$L \leftarrow \min(n, m)$
while $L \geq 1$ **do**
  for $i = 0$ **to** $n - L$ **do**
    if $\exists\, j$ s.t. $G_t[i : i+L] = E_t[j : j+L]$ **then**
      $\alpha \leftarrow L/n$, $\beta \leftarrow L/m$
      if $\alpha \geq \theta$ **and** $\beta \geq \theta$ **then**
        return true
      end if
    end if
  end for
  $L \leftarrow L - 1$
end while
return false

E Matching Algorithm of Error Span Detection

A matching algorithm is employed to validate the accuracy of error span detection. Specifically, given a gold-standard error span G and a detected error span E, the match is confirmed only if both overlap ratios satisfy $|I|/|G| \geq \theta$ and $|I|/|E| \geq \theta$, where I denotes the intersection segment $G \cap E$, $|\cdot|$ represents length measurement (e.g., measured by token or character count), and θ is a predefined matching threshold $0 < \theta \leq 1$. For instance, when $\theta = 50\%$, if the gold span G is "go back to the lab" (4 tokens) and the detected span E is "back to the lab tomorrow" (5 tokens), their intersection I ("back to the lab", 3 tokens) yields overlap ratios of 75% (3/4 for G) and 60% (3/5 for E), both exceeding the matching threshold 50%. The pseudocode is shown as Algorithm 1.

F Supplementary Analysis on Domain-Specific Performance

As illustrated in Figure 4, although HiMATE may exhibit slightly lower performance in specific domains, it demonstrates strong alignment with manual evaluation across most scenarios. Incorporating a hierarchical multi-agent structure combined with fine-grained error detection ensures robust performance of HiMATE across diverse text domains.

G Prompts of HiMATE Agents

The prompts utilized by the proposed HiMATE are presented in Tables 9 and 10. A three-stage pipeline is adopted, where the self-reflection stage comprises two sequential steps, and the collaborative discussion stage involves a dialogue procedure between two tiers of agents.

H Detailed Case Example

This section presents three illustrative cases demonstrating the complete workflow of each agent within HiMATE. These examples are generated utilizing the gpt-4o-mini model. The first example, sourced from MQM22 ZH-EN, corresponds to the case discussed in Chapter 5.5, detailed in Tables 11 and 12. The second example, also derived from MQM22 ZH-EN, is illustrated from Table 13 to Table 15. The third example, originating from MQM22 EN-DE, is presented in Tables 16 and 17.

Table 8: Core error subtypes and descriptions after refinement. Here, 'Core' refers to high-level error categories at tier-1, while 'Subtype' denotes fine-grained error classifications at tier-2. 'Description' are utilized within the prompt to characterize each error type.

Core	**Subtype**	**Description**
Accuracy	Addition	Translation includes information (including the punctuation) not present in the source.
	Omission	A paragraph present in the source (including the noun, verb, adverb, adverbial, punctuation, and so on) is missing in the translation.
	Mistranslation	Incorrect use in target content of a word, inconsistent match, and the incorrect segmentation.
	Untranslated text	Source text has been left untranslated. Not translating special symbols or placeholders is not an untranslated text error.
Terminology	Inappropriate for context	Use of a term that differs from term usage required by a specified termbase or other resource.
	Inconsistent use	Terminology is used inconsistently.
Fluency	Punctuation	Unpaired quote marks or parentheses. Missing mark from a set of paired punctuation marks, such as a missing parenthesis or quote mark. And the omission or addition of punctuation.
	Spelling	Error occurring when a word is misspelled.
	Grammar	Subject-verb disagreement, incorrect verb tenses or forms, and improper declension of nouns, pronouns, or adjectives.
	Register	Wrong grammatical register (e.g. inappropriately informal pronouns).
	Inconsistency	Internal inconsistency (not related to terminology).
	Character encoding	Error occurring when characters are garbled due to incorrect application of an encoding.
Style	Awkward	Some languages provide grammatical and stylistic features that support complex, embedded ideas, which can result in awkward style if a target text retains these features.
Locale Convention	Address format	Error involving inappropriate address format for locale. Such as the wrong form used for postal codes for target locale.
	Currency format	Incorrect currency format for its locale.
	Date format	Error involving inappropriate date format for its locale.
	Name format	Name format used in text or a data field inappropriate for its locale, such as switching the order of last and first names inappropriately.
	Telephone format	Error involving inappropriate telephone number form for locale.
	Time format	Error involving incorrect time format for its locale. For example, unless specified as using a 24-hour clock, US time formats report time after noon using 12-hour notation.

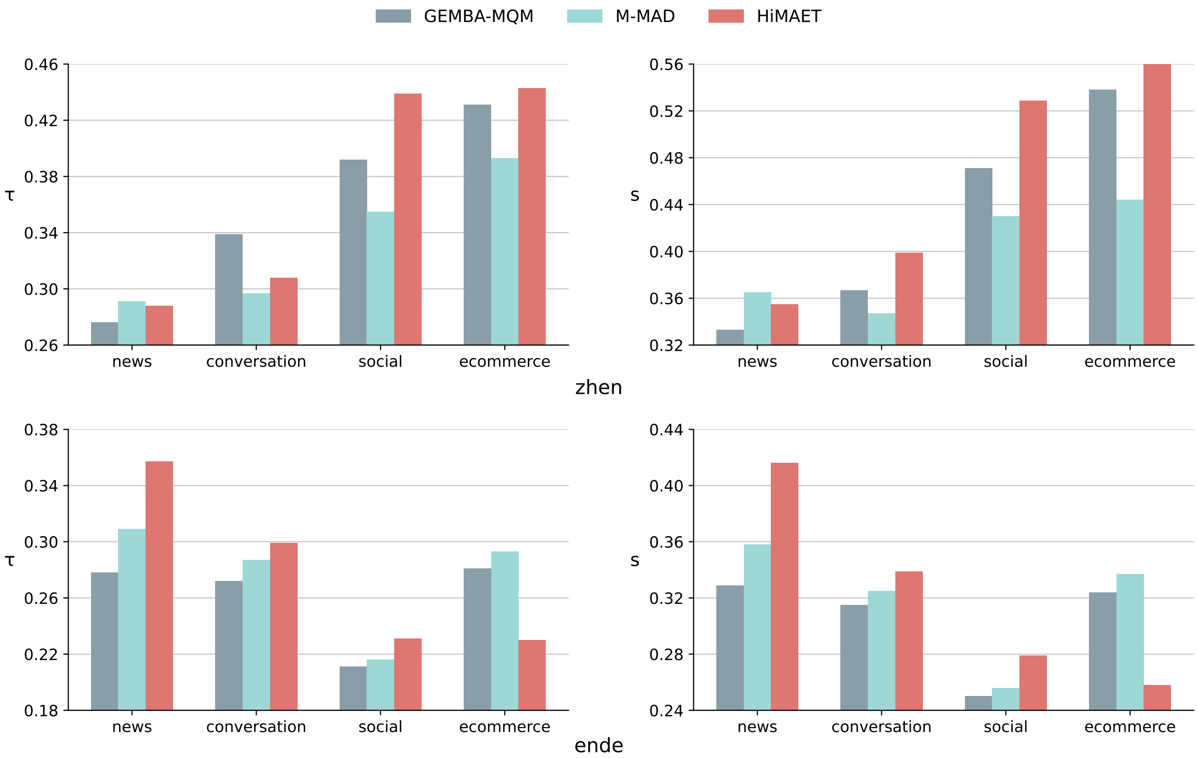

Figure 4: The Kendall's correlation coefficient τ and Spearman's correlation coefficient s across multiple text domains on MQM22 ZH-EN and EN-DE dataset using the gpt-4o-mini model.

Table 9: The prompt of Subtype Evaluation and Self-Reflection stages. Here, the Self-Reflection stage comprises two sequential steps, the error correction and comparative verification.

Subtype Evaluation	*Self-Reflection*	
Role-Play Prompt You are an evaluator conducting a manual translation evaluation based on the MQM framework, and you need to determine whether the following translation from Chinese to English below contains an error.	**Role-Play Prompt** You are an evaluator conducting a manual translation evaluation based on the MQM framework. Here you are given a pair of sentences: a source Chinese sentence and its English translation. You need to correct the errors in the translation.	**Role-Play Prompt** You are an evaluator conducting a manual translation evaluation based on the MQM framework. Here you need to compare the following original translation and the corrected translation based on the error definition.
Task Assignment Prompt {Source} {Translation} {Error Definition}	**Task Assignment Prompt** {Source} {Translation} {Error Definition} {Error Information}	**Task Assignment Prompt** {Original Translation} {Corrected Translation}
Please check if there are errors in the translation and whether they meet the definition of error types. You also need to indicate the degree of error. Error severities are assigned independently of category, and consist of Major, Minor, and Neutral levels, corresponding, respectively, to actual translation or grammatical errors, smaller imperfections, and purely subjective opinions about the translation.	Previously, you have identified specific types of errors and the severity based on the error definition. Based on the information below, please correct the errors in the translation.	Previously, you have corrected the translation. Now, determine whether the error mentioned previously in the original translation has been corrected.

Table 10: The prompt of the Collaborative Discussion stage. Two tiers of agents participate in this stage and engage in discussion.

Collaborative Discussion	
Tier-1 Agent Round	**Tier-2 Agent Round**
Role-Play Prompt You are an expert in manual translation evaluation based on the MQM framework and are responsible for a high-level error type. In previous work, an evaluator responsible for a subtype has expressed his lack of confidence in his answers. Please review his work and provide your opinions.	**Role-Play Prompt** You are an evaluator who manually evaluates translations based on the MQM framework. In your previous work, you were not sure about the accuracy of your evaluation. Therefore, you communicated with the expert in charge of coordinating the evaluation of the error subtype for which you are responsible.
Task Assignment Prompt {Source Sentence} {Translation} {Previous Error Information} {Previous Chat History} The following is the source sentence and its corresponding translation, as well as the assessment of the subcategory evaluator on the existence and severity of the error. Please provide your opinion. You need to determine whether an error exists, and then judge whether the degree of error is correct. You don't need to completely negate it.	**Task Assignment Prompt** {Source Sentence} {Translation} {Previous Error Information} {Previous Chat History} Regarding the source sentence and its translation, as well as the evaluation answer you provided, your superior evaluation expert has provided the following opinions. Please indicate whether you agree with his viewpoint.

Table 11: The workflow for each agent in the first case from the MQM22 ZH-EN dataset generated by gpt-4o-mini. The table illustrates the subtype evaluation stage, with segments highlighted in red indicating incorrect judgments and those in green indicating correct ones. Subsequent tables will consistently adopt this convention.

Source	希望你们了解一下
Translation	I hope you know about it.
Reference	I hope you can find out about it.
Annotation	Major-Mistranslation-‘know about it’
HiMATE-SE	Major-Mistranslation-‘know about it’ tier-2 agent: "The translation "I hope you know about it" does not accurately convey the meaning of the source text "希望你们了解一下" which more appropriately translates to "I hope you all understand" or "I hope you all are aware" The use of "know about it" implies a lesser degree of understanding and does not capture the intent of encouraging awareness or understanding, thus constituting a major mistranslation." Minor-Inappropriate for context-‘know’ tier-2 agent: "The translation "I hope you know about it" does not accurately convey the nuance of "希望你们了解一下," which suggests a desire for the audience to gain a better understanding or awareness of something, rather than simply knowing about it. The term "了解" implies a deeper comprehension than "know," which could be seen as inappropriate for the context. However, the overall meaning is still conveyed, hence the error is classified as minor." Major-Omission-‘希望’ tier-2 agent: "The translation omits the word "希望" ("hope") in the source, which changes the tone and intention of the message. The translation does not convey the speaker’s hope or desire for understanding, resulting in a significant difference in meaning." Minor-Awkward-‘I hope you know about it’ tier-2 agent: "The translation "I hope you know about it" is somewhat awkward because it implies a level of familiarity that may not be intended in the original Chinese phrase "希望你们了解一下" which suggests a desire for the listeners to gain some understanding or awareness rather than simply knowing about something. A more natural translation could be "I hope you can understand this" or "I hope you can learn about this" which would better convey the intent of the original message." Minor-Addition-‘know about it’ tier-2 agent: "The translation "I hope you know about it" is somewhat awkward because it implies a level of familiarity that may not be intended in the original Chinese phrase "希望你们了解一下" which suggests a desire for the listeners to gain some understanding or awareness rather than simply knowing about something. A more natural translation could be "I hope you can understand this" or "I hope you can learn about this" which would better convey the intent of the original message."

Table 12: The workflow for each agent in the first case from the MQM22 ZH-EN dataset generated by gpt-4o-mini. The table illustrates the self-reflection and collaborative discussion stages.

Source	希望你们了解一下
Translation	I hope you know about it.
Reference	I hope you can find out about it.
Annotation	Major-Mistranslation-'know about it'
HiMATE-SR-Correction	Major-Mistranslation-'know about it' tier-2 agent: "I hope you all understand." Minor-Inappropriate for context-'know' tier-2 agent: "I hope you can understand it better." Major-Omission-'希望' tier-2 agent: "I hope you understand this." Minor-Awkward-'I hope you know about it' tier-2 agent: "I hope you can understand this." Minor-Addition-'know about it' tier-2 agent: "I hope you understand a bit about it."
HiMATE-SR-Verification	Major-Mistranslation-'know about it' tier-2 agent: "Error has been corrected." Minor-Inappropriate for context-'know' tier-2 agent: "No significant difference between the original translation and the corrected Translation." Major-Omission-'希望' tier-2 agent: "No significant difference between the original translation and the corrected Translation." Minor-Awkward-'I hope you know about it' tier-2 agent: "Error has been corrected." Minor-Addition-'know about it' tier-2 agent: "Error has been corrected."
HiMATE-CD-Round1	Major-Mistranslation-'know about it' tier-1 agent: "Error Exist: Yes Error Severity: Major. The evaluator's assessment is correct. The translation "I hope you know about it" does not accurately capture the meaning of the source sentence "希望你们了解一下" which conveys a sense of encouraging understanding or familiarity rather than merely hoping that the audience knows something. This misrepresentation of intent in the translation qualifies as a major error, as it alters the original message significantly." Minor-Awkward-'I hope you know about it' tier-1 agent: "Error Exist: No. Error Severity: Neutral. The translation "I hope you know about it" is grammatically correct but does not fully capture the nuance of the source sentence "希望你们了解一下" which implies a desire for understanding rather than mere knowledge. The awkwardness stems from the choice of words rather than a clear error, making it more subjective. Therefore, I would classify the severity as Neutral rather than Minor." Minor-Addition-'know about it' tier-1 agent: "Error Exist: No. Error Severity: Neutral. The evaluator correctly identified the addition error, as the translation introduces a nuance that is not present in the source text. However, the severity should be considered neutral rather than minor, as the addition does not significantly distort the original meaning or create a major misunderstanding. It is more of a stylistic difference rather than a critical error."
HiMATE-CD-Round2	Major-Mistranslation-'know about it' tier-2 agent: "I agree with him." Minor-Awkward-'I hope you know about it' tier-2 agent: "I agree with him." Minor-Addition-'know about it' tier-2 agent: "I agree with him."

Table 13: The workflow for each agent in the second case from the MQM22 ZH-EN dataset generated by gpt-4o-mini. The table illustrates the subtype evaluation stage.

Source	无论从接待、化妆、服装、灯光、摄影到后期都无可挑剔。
Translation	From reception, makeup, costumes, lighting, photography to the late stage, it is impeccable.
Reference	It was perfect in reception, makeup, clothing, lamplight, photography and post-editing.
Annotation	Major-Mistranslation-'late stage'
HiMATE-SE	Major-Mistranslation-'the late stage' tier-2 agent: "The phrase "to the late stage" in the translation does not accurately convey the meaning of the original Chinese text, which implies that everything is impeccable "from reception to post-production" without a clear distinction or segmentation of "the late stage." This mistranslation leads to confusion about the intended meaning and scope of the description." Minor-Addition-'the late stage' tier-2 agent: "The translation introduces the phrase "the late stage" which is not present in the source text ("后期" could refer to a general later stage rather than a specific term). This addition alters the original meaning slightly by being more specific than the original context intended." Major-Omission-'无论从' tier-2 agent: "The translation omits the phrase "无论从" (which means "regardless of" or "no matter if") from the original Chinese sentence, altering the meaning and emphasis of the statement. This results in a significant loss of context and intent, qualifying it as a major omission." Minor-Grammar-'to the late stage' tier-2 agent: "The phrase "to the late stage" does not accurately reflect the intended meaning of "后期" which typically refers to "post-production" or "editing" in a context involving photography or film. Although the overall translation is understandable, this specific phrase could lead to ambiguity regarding the stages mentioned, making it a minor grammatical error." Minor-Inconsistency-'the late stage' tier-2 agent: "The translation uses ẗhe late stageẅhich is not a commonly used term in English for "后期" in this context. A more appropriate term would be "post-production" or "final stage." This creates a slight inconsistency in the translation, but it does not significantly alter the meaning." Minor-Awkward-'to the late stage' tier-2 agent: "The translation "From reception, makeup, costumes, lighting, photography to the late stage, it is impeccable." feels awkward due to the phrasing "to the late stage" which is not a commonly used expression in English. A more natural phrasing like "post-production" would make it smoother. However, the overall meaning is conveyed, which keeps the severity at a minor level."

Table 14: The workflow for each agent in the second case from the MQM22 ZH-EN dataset generated by gpt-4o-mini. The table illustrates the self-reflection stage.

Source	无论从接待、化妆、服装、灯光、摄影到后期都无可挑剔。
Translation	From reception, makeup, costumes, lighting, photography to the late stage, it is impeccable.
Reference	It was perfect in reception, makeup, clothing, lamplight, photography and post-editing.
Annotation	Major-Mistranslation-‘late stage’
HiMATE-SR-Correction	Major-Mistranslation-‘the late stage’ tier-2 agent: "From reception, makeup, costumes, lighting, photography to post-production, everything is impeccable." Minor-Addition-‘the late stage’ tier-2 agent: "From reception, makeup, costumes, lighting, photography to post-production, it is impeccable." Major-Omission-‘无论从’ tier-2 agent: "Regardless of reception, makeup, costumes, lighting, photography, or the late stage, it is impeccable." Minor-Grammar-‘to the late stage’ tier-2 agent: "From reception, makeup, costumes, lighting, photography to post-production, it is impeccable." Minor-Inconsistency-‘the late stage’ tier-2 agent: "From reception, makeup, costumes, lighting, photography to post-production, it is impeccable." Minor-Awkward-‘to the late stage’ tier-2 agent: "From reception, makeup, costumes, lighting, photography to post-production, it is impeccable."
HiMATE-SR-Verification	Major-Mistranslation-‘the late stage’ tier-2 agent: "Error has been corrected." Minor-Addition-‘the late stage’ tier-2 agent: "Error has been corrected." Major-Omission-‘无论从’ tier-2 agent: "No significant difference between the original translation and the corrected Translation." Minor-Grammar-‘to the late stage’ tier-2 agent: "No significant difference between the original translation and the corrected Translation." Minor-Inconsistency-‘the late stage’ tier-2 agent: "Error has been corrected." Minor-Awkward-‘to the late stage’ tier-2 agent: "Error has been corrected."

Table 15: The workflow for each agent in the second case from the MQM22 ZH-EN dataset generated by gpt-4o-mini. The table illustrates the collaborative discussion stage.

Source	无论从接待、化妆、服装、灯光、摄影到后期都无可挑剔。
Translation	From reception, makeup, costumes, lighting, photography to the late stage, it is impeccable.
Reference	It was perfect in reception, makeup, clothing, lamplight, photography and post-editing.
Annotation	Major-Mistranslation-'late stage'
HiMATE-CD-Round1	Major-Mistranslation-'the late stage' tier-1 agent: "Error Exist: Yes. Error Severity: Major. The translation "From reception, makeup, costumes, lighting, photography to the late stage, it is impeccable" misrepresents the original meaning. The term "后期" typically refers to "post-production" rather than "the late stage." This distortion significantly alters the intended message about the quality of the various aspects involved, thus justifying a major severity rating." Minor-Addition-'the late stage' tier-1 agent: "Error Exist: No. Error Severity: Neutral. The evaluator correctly identified the addition error, as the phrase "late stage" in the translation does not correspond to the source text. However, the severity should be considered neutral rather than minor, as this error does not belong to an addition error. " Minor-Inconsistency-'the late stage' tier-1 agent: "Error Exist: No. Error Severity: Neutral. The evaluator correctly identified the existence of an inconsistency in the translation, as the phrase "late stage" does not accurately reflect the original meaning of "后期" which typically refers to "post-production" in this context. However, the severity should be considered Neutral rather than Minor, as this does not significantly impact the overall understanding of the text. The translation remains comprehensible, and the error does not lead to a misunderstanding of the main message." Minor-Awkward-'to the late stage' tier-2 agent: "Error Exist: No. Error Severity: Neutral. The translation retains grammatical acceptability but lacks the fluidity and coherence expected in the target language. While it does exhibit awkwardness, the severity should be assessed as neutral rather than minor, as it primarily involves a subjective opinion rather than causing a notable impact on the overall meaning or readability. The phrase "to the late stage" feels particularly awkward and could confuse readers, but it does not constitute a significant error affecting comprehension."
HiMATE-CD-Round2	Major-Mistranslation-'the late stage' tier-2 agent: "I agree with him." Minor-Addition-'the late stage' tier-1 agent: "I agree with him." Minor-Inconsistency-'the late stage' tier-1 agent: "I agree with him." Minor-Awkward-'to the late stage' tier-2 agent: "I agree with him."

Table 16: The workflow for each agent in the third case from the MQM22 EN-DE dataset generated by gpt-4o-mini. The table illustrates the subtype evaluation stage.

Source	This move is just the latest twist in the Dallas-Washington rivalry, one that got juiced up even more this week when Cowboys head coach Mike McCarthy predicted a win for his team, prompting some fireworks back and forth with Washington's Ron Rivera and players.
Translation	Dieser Schritt ist nur die jüngste Wendung in der Rivalität zwischen Dallas und Washington, die diese Woche noch stärker gemacht wurde, als Cowboys-Cheftrainer Mike McCarthy einen Sieg für sein Team vorhersagte und ein Feuerwerk zwischen Ron Rivera und Spielern aus Washington auslöste.
Reference	Diese Aktion ist nichts als ein weiteres Kapitel in der Rivalität zwischen Dallas und Washington, die diese Woche noch mehr an Fahrt aufnahm, als der leitende Coach der Cowboys Mike McCarthy einen Sieg seines Teams voraussagte und ein hitziges Wortgefecht mit Ron Rivera von Washington und seinen Spielern auslöste.
Annotation	Minor-Awkward-'stärker gemacht', Minor-Awkward-'Feuerwerk', Minor-Mistranslation-'Feuerwerk'
HiMATE-SE	Minor-Awkward-'stärker gemacht wurde' tier-2 agent: "The phrase "stärker gemacht wurde" is somewhat awkward in German. A more natural phrasing could be "intensiver wurde" or "mehr an Intensität gewann." The current translation retains the original structure but does not flow as smoothly in German, leading to a minor awkwardness in style." Major-Mistranslation-'Feuerwerk' tier-2 agent: "The translation uses "Feuerwerk" which is more commonly associated with fireworks in a literal sense, while the context implies a figurative sense of heated exchanges or confrontations. A more appropriate term might be "Auseinandersetzungen" or "Konflikte" to convey the intended meaning of verbal disputes." Minor-Inconsistency-'die diese Woche noch stärker gemacht wurde' tier-2 agent: "The phrase "die diese Woche noch stärker gemacht wurde" implies that the rivalry was made stronger, whereas the original phrase suggests a progression in intensity due to recent events. A more accurate translation could clarify that the rivalry was intensified rather than simply "made stronger." This is a minor inconsistency that doesn't convey the precise meaning of the source text."

Table 17: The workflow for each agent in the third case from the MQM22 EN-DE dataset generated by gpt-4o-mini. The table illustrates the self-refleciton and collaborative discussion stages.

Source	This move is just the latest twist in the Dallas-Washington rivalry, one that got juiced up even more this week when Cowboys head coach Mike McCarthy predicted a win for his team, prompting some fireworks back and forth with Washington's Ron Rivera and players.
Translation	Dieser Schritt ist nur die jüngste Wendung in der Rivalität zwischen Dallas und Washington, die diese Woche noch stärker gemacht wurde, als Cowboys-Cheftrainer Mike McCarthy einen Sieg für sein Team vorhersagte und ein Feuerwerk zwischen Ron Rivera und Spielern aus Washington auslöste.
Reference	Diese Aktion ist nichts als ein weiteres Kapitel in der Rivalität zwischen Dallas und Washington, die diese Woche noch mehr an Fahrt aufnahm, als der leitende Coach der Cowboys Mike McCarthy einen Sieg seines Teams voraussagte und ein hitziges Wortgefecht mit Ron Rivera von Washington und seinen Spielern auslöste.
Annotation	Minor-Awkward-'stärker gemacht', Minor-Awkward-'Feuerwerk', Minor-Mistranslation-'Feuerwerk'
HiMATE-SR-Correction	Minor-Awkward-'stärker gemacht wurde' tier-2 agent: "Dieser Schritt ist nur die jüngste Wendung in der Rivalität zwischen Dallas und Washington, die diese Woche noch intensiver wurde, als Cowboys-Cheftrainer Mike McCarthy einen Sieg für sein Team vorhersagte und ein Feuerwerk zwischen Ron Rivera und Spielern aus Washington auslöste." Major-Mistranslation-'Feuerwerk' tier-2 agent: "Dieser Schritt ist nur die jüngste Wendung in der Rivalität zwischen Dallas und Washington, die diese Woche noch stärker gemacht wurde, als Cowboys-Cheftrainer Mike McCarthy einen Sieg für sein Team vorhersagte und Auseinandersetzungen zwischen Ron Rivera und Spielern aus Washington auslöste." Minor-Inconsistency-'die diese Woche noch stärker gemacht wurde' tier-2 agent: "Dieser Schritt ist nur die jüngste Wendung in der Rivalität zwischen Dallas und Washington, die diese Woche noch intensiver wurde, als Cowboys-Cheftrainer Mike McCarthy einen Sieg für sein Team vorhersagte und ein Feuerwerk zwischen Ron Rivera und Spielern aus Washington auslöste."
HiMATE-SR-Verification	Minor-Awkward-'stärker gemacht wurde' tier-2 agent: "Error has been corrected." Major-Mistranslation-'Feuerwerk' tier-2 agent: "Error has been corrected." Minor-Inconsistency-'die diese Woche noch stärker gemacht wurde' tier-2 agent: "No significant difference between the original translation and the corrected Translation."
HiMATE-CD-Round1	Minor-Awkward-'stärker gemacht wurde' tier-1 agent: "No explanation due to exceeding the stage-transition threshold." Minor-Mistranslation-'Feuerwerk' tier-1 agent: "Error Exist: Yes. Error Severity: Minor. While there is a mistranslation present, the errors noted do not significantly alter the overall meaning of the source text. The translation captures the core sentiment and information though with some imprecision. Therefore, although there is an error, it does not warrant a major severity classification, as it does not impede comprehension or alter the message's intent in a substantial way. A minor severity classification is more appropriate here."
HiMATE-CD-Round2	Minor-Mistranslation-'Feuerwerk' tier-2 agent: "I agree with him."