RUBRIC-MQM : Span-Level LLM-as-judge in Machine Translation For High-End Models

Ahrii Kim
Independent Researcher
ahriikim@gmail.com

Abstract

Referred to as *LLM-as-judge*, a generative large language model (LLM) has demonstrated considerable efficacy as an evaluator in various tasks, including Machine Translation (LAJ-MT) by predicting scores or identifying error types for individual sentences. However, its dependability in practical application has yet to be demonstrated, as there is only an *approximated match* due to the task's open-ended nature. To address this problem, we introduce a straightforward and novel meta-evaluation strategy **PROMPTCUE** and evaluate cutting-edge LAJ-MT models such as GEMBA-MQM. We identify their fundamental deficits, including certain label biases and the inability to assess near-perfect translations.

To improve reliability, we investigate more trustworthy and less biased models using multidimensional prompt engineering. Our findings indicate that the combination of span-level error quantification and a rubric-style prompt tailored to the characteristics of LLMs has efficiently addressed the majority of the challenges current LAJ-MT models face. Furthermore, it demonstrates a considerably enhanced alignment with human values. Accordingly, we present **RUBRIC-MQM**, the LAJ-MT for high-end models and an updated version of GEMBA-MQM.[1]

1 Introduction

A notable strength of generative Large Language Models (LLMs) lies in their capacity to utilize user instructions to execute tasks that are both unseen and untuned, thereby demonstrating remarkable performance across various domains of natural language processing (NLP) such as Code Generation and Text Summarization (Ouyang et al., 2022; Wang et al., 2023; Dainese et al., 2024; Zhang et al., 2024). This swift expansion has prompted more scholars to initiate comprehensive investigations into their potential, including capacity for self-evaluation, referred to as *LLM-as-judge* (LAJ) (Bavaresco et al., 2024; Ashktorab et al., 2024; Ashktorab et al., 2024). This paradigm employs LLMs to evaluate model-generated outputs based on a set of predefined criteria (Li et al., 2024). What sets this approach apart from traditional evaluation metrics is its inherent flexibility. This flexibility permits LLMs to leverage their comprehensive knowledge, acquired from extensive data, to conduct evaluations in accordance with user directives.

Within this context, the domain of Machine Translation (MT) has illustrated prompt-based evaluation (LAJ-MT) models demonstrating noteworthy efficacy (Kocmi and Federmann 2023b; (Lu et al., 2024); Fernandes et al. 2023; Kocmi and Federmann 2023a). Despite their outstanding performance, their meta-evaluation relies on ***approximating error spans*** due to its open-ended nature. Furthermore, their performance is evaluated solely by juxtaposing them with pre-existing metrics, which provides limited insight into their reliability, advantages, or disadvantages in practical applications. As a result, they become just another *black-box* LLM (Fernandes et al., 2023; Kocmi and Federmann, 2023a).

To tackle these issues, we introduce a novel meta-evaluation method called **PROMPTCUE** (***Prompt**-based **C**lassification for **U**ncovering **E**rrors*), facilitating targeted error classification in MT. Eliminating error detection from the traditional evaluation process simplifies the task into a basic classification problem. We propose this approach as the first direct meta-evaluation of its kind. Our comprehensive analysis uncovers multiple critical deficiencies present within the existing LAJ-MT metrics, some of which are:

[1] This paper is the final version of our preprint, which can be found at: DR-100. All pertinent code and data are accessible at https://github.com/trotacodigos/Rubric-MQM.git.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), pages 147–165
July 28-30, 2025 ©2025 Association for Computational Linguistics

a) Biased to MISTRANSLATION and MAJOR
b) Systematic failure in NO-ERROR
c) Hallucinating error category

We enhance existing LAJs by streamlining the evaluation structure and implementing optimal prompt strategies. Our experiments explore the best prompting strategies with nine prompt types applying five strategies to GEMBA-MQM: Enumeration, Definition, Explanation, Rubric, and SQM style. We experimentally demonstrate that the rubric style yields the best performance in the current system. Thus, we present **RUBRIC-MQM**, a customized span-level MT evaluation metric that predicts MQM errors while simultaneously assigning scores out of 100 based on a detailed rubric. Our findings demonstrate that it successfully tackles two key challenges of GEMBA-MQM and considerably improves its alignment with human values in high-quality translations. This confirms its suitability for evaluating advanced MT models. Our key contributions are:

- We present PROMPTCUE, an innovative and straightforward approach for the direct meta-evaluation of LAJ-MT models.
- The traditional MQM scoring system is upgraded with RUBRIC-MQM, which assesses DA at the span level using a detailed scoring rubric. This approach enhances correlation and is especially apt for assessing top-tier models.
- We pinpoint major weaknesses of GEMBA-MQM and rectify them by examining fundamental prompt structures.

2 Background

MQM Framework

The MQM framework for Translation Quality Evaluation (TQE) is initially developed to perform a comprehensive analysis of translations produced both by human translators and machine-generated systems (Lommel et al., 2014). In this framework, an evaluator detects sentence errors and categorizes them by predefined category and severity criteria. For category, a hierarchical error typology includes seven meta-level errors with multiple sub-levels. The typology is customizable for specific linguistic features or uses. Conversely, the severity is divided into four types: NEUTRAL, MINOR, MAJOR, and CRITICAL. When scoring, the default weight for categories is set to 1, whereas severity is assigned weights of [0, -1, -5, -25], respectively. See Lommel et al. (2014) for scoring details.

This prominent evaluation framework, well-regarded in the field, has gained significant interest from MT researchers and is integrated into the Workshop on Machine Translation (WMT) with a few modifications. The hierarchy of the labels is simplified to 22 categories and three severities (NEUTRAL, MINOR, MAJOR) with a weight scheme of [0, -1, -5] (Kocmi et al., 2022). A single category type affecting the score is FLUENCY/PUNCTUATION, with a value of -0.1. The sentence-level score is calculated by summing all identified errors, with a cap of -25, which corresponds to either five instances of MAJOR or a single NON-TRANSLATION.

Defining Evaluation Function

Applying the evaluation process of LAJ as defined by Li et al. (2024), we have reconceptualized the MQM process with the following equation:

$$\mathcal{Y} = E(\mathcal{T}, \mathcal{C}, \mathcal{X}, \mathcal{R}) \tag{1}$$

where the evaluation *function* (E) is executed using evaluation inputs of *type* ($\mathcal{T}$), *criteria* ($\mathcal{C}$), *items* ($\mathcal{X}$), and an optional *reference* ($\mathcal{R}$), subsequently producing *outputs* ($\mathcal{Y}$) in the format of a numerical score or categorical label.

Current MT evaluation tasks often involve utilizing a single LLM ($\mathcal{T}$) to determine the severity and category ($\mathcal{C}$) of translation errors, based on the source and target sentences ($\mathcal{X}$), with or without reference ($\mathcal{R}$). Within this context, the existing prompt-based models have two criteria —severity ($\mathcal{C}_{sev}$) and category ($\mathcal{C}_{cat}$) —, wherein each criterion independently yields results in the form of a categorical label, as in Equation 2.

$$\begin{aligned} \mathcal{Y}_{cat} &= E(\mathcal{T}, \mathcal{C}_{cat}, \mathcal{X}, \mathcal{R}) \\ \mathcal{Y}_{sev} &= E(\mathcal{T}, \mathcal{C}_{sev}, \mathcal{X}, \mathcal{R}) \end{aligned} \tag{2}$$

We consolidate this redundant procedure into a singular task by transforming severity as criteria ($\mathcal{C}_{sev}$) into a numerical output of category ($\mathcal{Y}_{cat}$), as illustrated in Equation 3. Note that our framework is without reference ($\mathcal{R}$). For clarity, we designate category and severity as C_{cat} and Y_{sev} respectively throughout this paper.

$$\mathcal{Y}'_{sev} = E(\mathcal{T}, \mathcal{C}_{cat}, \mathcal{X}) \tag{3}$$

```
English source: ```I do apologise
about this, (...) from <v>the
account holder</v> to discuss an
order (...) holders permission.```

German translation: ```Ich
entschuldige mich dafür, (...)
geschehen <v>wäre</v>, aber ohne
die Erlaubnis des Kontoinhabers
wäre ich nicht in der Lage, dies mit
<v>dir</v> <v>involvement</v>.```


{prompt} They are enclosed with <v>
and </v> tags.
```

Table 1: Components of the PROMPTCUE strategy, as delineated in blue words, are applicable to any LAJs.

3 PROMPTCUE

3.1 Design

The fundamental concept involves the precise delineation of error spans with <v> and </v> tags within the translation process (Table 1). The model is responsible for the correct allocation of labels for C_{cat} and Y_{sev} to the designated span, as specified in Table 1. Removing the initial step thus turns the task into a straightforward classification problem. Defining the error range ourselves and treating it as a finite task has three benefits. Firstly, we evaluate the model's grasp of criteria C_{cat} and Y_{sev}. Secondly, the core framework of all LAJ-MTs is consistent in our evaluation environment, which guarantees our approach's universal applicability. Finally, and most importantly, calculating a match ratio becomes simple and clear.

3.2 Match Ratio

The estimated match ratio quantifies how closely model predictions (A) align with gold judgments (B) by calculating $|A \cap B|/|B|$ (Kocmi et al., 2021). Fortunately, PROMPTCUE enables a straightforward comparison of A and B. We define a match per criterion: span, category, and severity.

Span Match The successful response rate is calculated when the expected answer for the span is given. If there is no response, we label it as `none`, employing a One-vs-Rest classification. Noisy responses with multiple entries are treated as error margins.

Category Match It pertains to the precise correspondence of the C_{cat} label. Our predefined error typology is detailed in the Appendix E.1.

Severity Match It refers to an exact match with MAJOR/MINOR. If a method lacks a binary system or produces numerical value, we calculate the optimal threshold for the best match ratio. Appendix E.2 provides the details.

3.3 Metrics

The primary metrics utilized for PROMPTCUE are Accuracy and Macro-F1 scores. Accuracy represents the count of correct classifications, encompassing both positive and negative matches. The Macro-F1 score is computed by the unweighted mean of class-wise F1 scores.

4 Experiment Setup

4.1 Data Construction

We use the MQM 2023 test set (Freitag et al., 2023) for Chinese-to-English translation, anticipating that this high-resource language pair will facilitate broader generalization of the results obtained from our novelty evaluation. We create three benchmarks, GEN, PTB, and MIS, each with 1,000 segments for label-centric evaluation. Details of the dataset are in Appendix B.

GEN set It evaluates the general performance by C_{cat} of 10 labels and Y_{sev} of two labels, evenly distributed across the benchmark.

PTB set It evaluates the ability to distinguish perfect (NO-ERROR) from imperfect (MAJOR) translations.[2] Flawed synthetic sentences are created using perturbation techniques.

MIS set It evaluates the model's peak performance in C_{cat} classification using MISTRANSLATION labels only.

4.2 Prompting Strategies

GEMBA-MQM is the default prompt setup, using a reference-free three-shot method. Subsequently, five distinct prompt strategies are mix-matched to form diverse slot scenarios, as in Table 2. We also test scales 4, 8, and 100 to find the optimal scale for strategies *Rubric* and *Continuous*. This results in nine slot scenarios named: `DeepCat`, `DeepShot`, `DeepCatShot`, `DeepRubric-`n, and `DeepQ-`n ($n = [4, 8, 100]$). `DeepQ-n` is inspired by the GEMBA-SQM fashion (Kocmi and Federmann, 2023a). Table 2 provides their detailed design features. Detailed prompt templates and lines are described in Appendix F.

[2] Our framework does not include the original NON-TRANSLATION label.

Strategy	Abbr.	About	Slot Scenarios Base	 DC	 DS	 DCS	 DR	 DQ
Enumeration	ENUM	A list provides the types of C_{cat} labels.	✓	✓	✓	✓	✓	✓
Definition	DEF	A definition per C_{cat} label is given.		✓		✓		
Example	EXP	Each C_{cat} label is elucidated using ICL examples.			✓	✓		
Rubric	-R	The scale for Y_{sev} is described with a scoring rubric.					✓	
Continuous	-Q	A continuous statement is used to describe the scale of Y_{sev}.						✓

Table 2: Prompting strategies and their applicability across slot scenarios.

		GEN ↑	PTB	MIS
Y_{sev}	Major	**63.59**	**74.80**	**79.20**
	Minor	18.57	17.10	14.90
	None	13.28	8.10	5.50
	No-error	4.56	-	0.40
C_{cat}	Mistranslation	**42.63**	**55.50**	**76.50**
	Omission	9.75	2.70	3.50
	Punctuation	9.54	3.60	0.60
	Terminology	5.91	0.50	6.80
	Addition	4.36	10.80	1.70
	Word order	3.01	10.30	1.10
	Grammar	2.80	3.60	1.80
	Untranslated	2.28	2.60	0.10
	Inconsistency	1.76	2.30	2.00
	Source issue	0.10	-	-

Table 3: Label distribution of GEMBA's prediction (unit: %). NONE, signifying no response, is included as a Y_{sev} label.

4.3 Judge Model

The SOTA LAJ-MT models referenced in §1 have learned from one another, leading to similar prompt lines, particularly concerning our goal. Therefore, in our study, we utilize GEMBA-MQM, referred to as GEMBA, as the base metric, representing the current SOTA models. We employ the proprietary GPT-4o (`gpt-4o-2024-11-20`) (OpenAI et al., 2024) as the foundational model, although the model specifications are unclear. To ensure reproducibility, the temperature is initially set to 0 and is increased only if there is no response.

5 Result: GEMBA

GEMBA effectively identifies errors but systematically fails to discern perfect translations, often mislabeling them as MAJOR or MISTRANSLATION. It fails to respond 8.95% of the cases, suggesting a relatively low match rate in real-world scenarios. This section discusses further details.

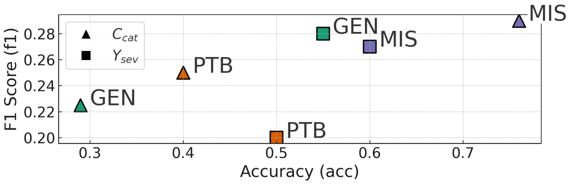


Figure 1: Varying outcomes for GEMBA performance across different datasets.

Different label distributions tell different stories.

Figure 1 illustrates the performance variation of GEMBA across different dataset types. The MIS set, which consists solely of MISTRANSLATION, exhibits the highest C_{cat} performance of the model ($acc = 0.765$, $f1 = 0.288$), which is a notable exaggeration in comparison to other datasets. Conversely, the PTB set, primarily comprising NO-ERROR, highlights its deficiency through a low F1 score as depicted in Figure 1. An in-depth analysis is conducted to examine the model's bias toward particular labels.

Biased towards MAJOR and MISTRANSLATION

As depicted in Table 3, the model predicts most errors for Y_{sev} as MAJOR and C_{cat} as MISTRANSLATION throughout the dataset. In PTB, 49.8% of the total NO-ERROR (74.8%) is a misclassification to MAJOR, while in MIS, 42.6% are wrongly labeled as MISTRANSLATION despite the fact that they make up merely 10% of the actual total. This problem, termed ***Overconfidence bias*** by Li et al. (2024), occurs primarily due to the uneven distribution of training sets.

NO-ERROR is consistently unacknowledged.

Although GEMBA demonstrates robustness on MISTRANSLATION, which serves as the gold stan-

Scenario	GEN				PTB				MIS				Win
	C_{cat}		Y_{sev}		C_{cat}		Y_{sev}		C_{cat}		Y_{sev}		
	acc	**f1**	**acc**	**f1**	**acc**	**f1**	**acc**	**f1**	**acc**	**f1**	**acc**	**f1**	**Win**
GEMBA	0.282	0.223	0.534	0.280	0.399	0.252	0.499	0.200	**0.765**	**0.289**	0.584	0.272	2
DQ-100	0.117	0.123	0.340	0.229	0.565↑	0.433↑	0.694↑	0.459↑	0.696	0.274	0.556	0.344↑	5
DQ-4	0.250	**0.236**↑	0.440	0.258	**0.644**↑	**0.474**↑	**0.736**↑	0.394↑	0.711	0.277	0.637↑	0.337↑	7
DQ-8	0.134	0.133	0.336	0.211	0.627↑	0.465↑	0.715↑	0.355↑	0.729	0.281	0.530	0.234	4
avg.	*0.167*	*0.164*	*0.372*	*0.232*	*0.612*	*0.457*	*0.715*	*0.403*	*0.712*	*0.277*	*0.574*	*0.305*	*5.3*
DC	0.285↑	0.216	**0.534**↑	0.282↑	0.427↑	0.271↑	0.498	0.203↑	0.706	0.276	0.587↑	0.276↑	8
DS	0.283↑	0.225↑	0.501	0.258	0.429↑	0.274↑	0.502↑	0.204↑	0.713	0.277	0.545	0.246	6
DCS	**0.303**↑	0.233↑	0.524	0.276	0.438↑	0.286↑	0.507↑	0.210↑	0.648	0.262	0.565	0.266	6
avg.	*0.290*	*0.225*	*0.520*	*0.272*	*0.431*	*0.277*	*0.502*	*0.206*	*0.689*	*0.272*	*0.566*	*0.263*	*6.7*
DR-100	0.259	0.230↑	0.506	0.282↑	0.574↑	0.429↑	0.674↑	**0.478**↑	0.732	0.282	0.654↑	0.345↑	8
DR-4	0.272	**0.236**↑	0.510	0.283↑	0.549↑	0.409↑	0.648↑	0.346↑	0.741	0.284	**0.661**↑	**0.347**↑	8
DR-8	0.263	0.227↑	0.509	**0.285**↑	0.597↑	0.428↑	0.667↑	0.357↑	0.756	0.287	0.643↑	0.340↑	8
avg.	*0.265*	*0.231*	*0.509*	*0.284*	*0.573*	*0.422*	*0.663*	*0.394*	*0.743*	*0.284*	*0.653*	*0.344*	*8*★

Table 4: Performance of all slot scenarios. The mean scores for each cluster are illustrated in blue line. ↑ denoes improvement over the baseline.

dard for half of the PTB set, its performance is the poorest within PTB in Figure 1. Table 3 indicates that the model allocates a total of 4.96% to NO-ERROR, yet it is absent in the PTB where it is expected. This leads to notably poor performance within this dataset. We suspect that a likely reason for GEMBA unacknowledging NO-ERROR could be the exclusion of it as a valid option for Y_{sev} in the prompt. This issue will be elucidated in §6.

Clearly hallucinating error category

Regardless of dataset organization, the model preserves the distribution of C_{cat} in Figure 3, indicating its inability to distinguish this criterion. The pattern becomes clearer when focusing on NO-ERROR segments. Figure 2 illustrates an overly varied spread of C_{cat} labels for flawless sentences, indicating its lack of reasoning ability on error categories and potential hallucinations. Further study is needed.

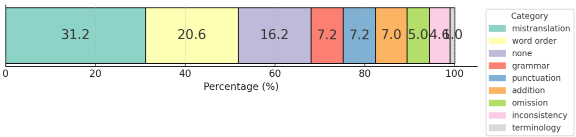


Figure 2: GEMBA's C_{cat} prediction for NO-ERROR segments in PTB (unit: %).

6 Result: Prompting Variations

RQ1: Has the general performance been improved? "Yes." Table 4 shows that no method is robust universally. When assessing the win rate against the baseline, the DR cluster consistently achieves favorable results, winning 8 out of 12 cases (67%).

RQ2: Is the Overconfidence bias alleviated? "Yes and No." The distribution of labels illustrated in Table 9 in the Appendix indicates that this bias is an intrinsic issue present across all models. However, the advantage is that the MAJOR bias is reduced by increasing MINOR in DR or having NO-ERROR in DR and DQ. To facilitate a clearer comparison, the Precision score (p) for MAJOR and MISTRANSLATION and Recall (r) for NO-ERROR are calculated by converting the predicted labels into a binary format. Table 5 indicates that while most variations have higher precision, DQ and DR effectively address issues in Y_{sev}. We propose that this enhancement results from using distinct criteria that circumvent reliance on the MAJOR / MINOR division. Conversely, the MISTRANSLATION bias is slightly reduced in some cases, but the changes are trivial.

RQ3: Is NO-ERROR discernible? "Yes." All scenarios win over GEMBA in PTB in Table 4. For instance, DQ-4 achieves 0.644 in C_{cat} and 0.736 in Y_{sev}, compared to 0.399 and 0.499 of the baseline. Table 9 in the Appendix illustrates that DR and DQ series cover a larger portion of NO-ERROR, though DQ series overestimate it in GEN, falsely labeling up to 61.83% of the cases (DQ-100). We demonstrate that the inability of GEMBA to generate NO-ERROR is closely linked to the Y_{sev} criteria, and the DR cluster effectively resolves this issue.

RQ4: Does it hallucinate less error typology? "No." Regarding C_{cat}, all scenarios demonstrate inconsistent performance by proposing a varied set of labels in PTB, as illustrated in Table 9 in

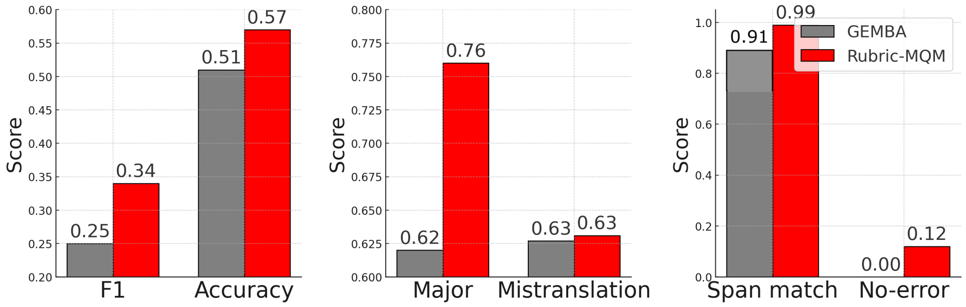


Figure 3: Six advantages of RUBRIC-MQM, addressing existing challenges of GEMBA. *Major* and *Mistranslation* indicate precision, while *No-error* refers to recall score.

	MAJOR (p)	MIST (p)	NO-ERROR (r)
GEMBA	0.616	0.627	0.000
DQ-4	**0.772**⋆	0.621 ↓	**0.158**⋆
DQ-8	0.692	0.643	0.143
DQ-100	0.750	0.637	0.129
DC	0.619	0.652	0.000 ↓
DS	0.611 ↓	0.645	0.002
DCS	0.622	**0.663**⋆	0.005
DR-4	0.759	0.632	0.099
DR-8	0.762	0.626 ↓	0.113
DR-100	0.760	0.631	0.116

Table 5: The precision and recall scores for specific labels across various scenarios. ↓ suggests a negative result, whereas ⋆ suggests the most positive.

the Appendix, and MISTRANSLATION is the most frequently chosen label. The classification capability seems largely independent of the instruction, indicating that in-depth research is required.

7 Further Study: RUBRIC-MQM

Figure 3 provides a concise overview of how RUBRIC-MQM addresses all identified challenges of GEMBA through PROMPTCUE: **it is more robust 1) in real-world scenarios with a higher match rate, 2) for high-quality translation evaluation, and 3) for MAJOR bias.** Additionally, it generates a span-level score that contributes to forming a continuous sentence-level score, thus confirming its status as the superior method.

7.1 Experiment Setting

We conduct a thorough assessment of RUBRIC-MQM to determine its efficacy in assessing advanced translation models. The model is tasked with evaluating reference translation (`ref A`) of the WMT 2023 Chinese-to-English translation (Kocmi et al., 2023).[3] Pearson (r), Spearman (p), and Kendall-Tau (τ) correlations with the gold standard (DA+SQM and MQM) are calculated at the sentence level. Additionally, other lightweight base models, beyond GPT-4o, such as GPT-3.5 Turbo (`gpt-3.5-turbo-0125`) and GPT-4o mini (`gpt-4o-mini-2024-07-18`), are examined. The parameters are uniformly set to max_token= 1024 and temperature= 0 across all cases. Given the novel nature of this trial, a standardized scoring scheme has yet to be established. Consequently, we investigate both the average and the aggregate of span-level scores.

7.2 Result

RUBRIC-MQM exhibits significant superiority over GEMBA, as well as the gold MQM, as illustrated in Figure 4. `4o-mini/avg` achieves the highest Pearson correlation with $r = 0.351$ against GEMBA ($r = 0.099$) or MQM ($r = 0.16$), while `4o/sum` excels in Spearman ($p = 0.352$ vs. 0.109) and Kendall ($\tau = 0.244$ vs. 0.08) correlations. Rankings change with scoring methods and models, though GPT-4 markedly outperforms GPT-3.5-Turbo. These results indicate that RUBRIC-MQM not only address existing issues but also significantly improve alignment with human values.

A portion of these advancements can be attributed to our method's continuous scoring system. Figure 5 illustrates that RUBRIC-MQM effectively mirrors SQM, with scores that are not clustered around 100. This is crucial as the existing gold score tends to skew toward zero (Freitag et al., 2023).

8 Conclusion

We have conducted a meta-evaluation of SOTA LAJ-MT models, utilizing a novel and streamlined strategy termed PROMPTCUE. By simplifying this

[3] Refer to Appendix B for detailed data information.

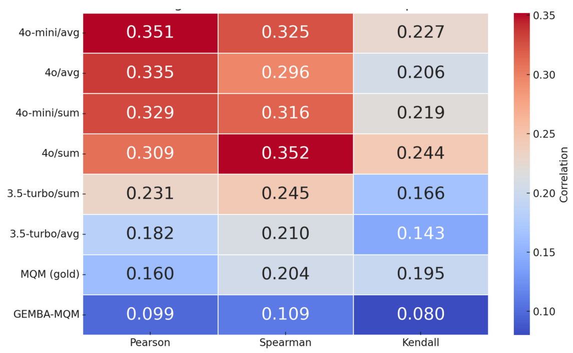
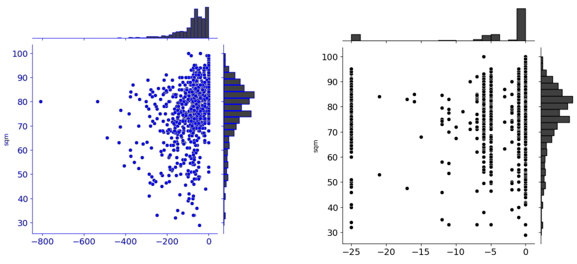


Figure 4: Three segment-level correlations to SQM, comparing GEMBA, diverse base models of RUBRIC-MQM, and gold MQM.



a) Rubric-MQM (4o-mini/avg) b) GEMBA-MQM

Figure 5: Score distribution. The x-axis represents DA+SQM shared across the two models.

process, significant issues within the MT evaluation framework are highlighted:

1) GEMBA shows biases toward MAJOR and MISTRANSLATION error types, so datasets focused on these errors will be advantageous to models of such nature.
2) Current LAJ-MT models cannot distinguish error types, a difficult task to accomplish via prompt engineering.

RUBRIC-MQM tackles most of the challenges GEMBA is facing by substituting the rigid label categorization with a scoring rubric. While emphasizing the system's exceptional performance in evaluating high-quality translations, it is evident that these achievements are facilitated by the LLM's capability to 'reason' and 'make decisions.' It is imperative to note that the capability to furnish the appropriate environment for each specific task lies within us, at least for the present moment.

Limitation and Future Work

The scope of this study is limited to a singular high-resourced language pair, analyzed unidirectionally. Given the proven effectiveness of the PROMPTCUE, future research will focus on exploring more language directions to uncover specific challenges and compare the multilingual capabilities of RUBRIC-MQM and GEMBA-MQM. The dataset is limited to a subset of WMT 23, and system-level human correlation for RUBRIC-MQM remains uninvestigated. While the metric seems effective, its reliability needs validation with broader datasets both at the segment and system levels. A further concern regarding the data is that the metrics within this study, pertaining to LAJ, are derived from proprietary models, which may possess pre-existing knowledge in GEN or MIS. Therefore, testing publicly available models like the Llama series, as recommended by Kocmi and Federmann, is a top priority.

Despite its remarkable features, RUBRIC-MQM continues to face challenges. Performance in both GEN and MIS mirrors that in GEMBA, with hallucinatory error categories persisting.There is a pressing need for a human assessment to confirm the current status and clarify the elements leading to reduced bias and better alignment with human values. Finally, researching its optimal scoring system is crucial for our future agenda.

Acknowledgement

I extend my appreciation to Teresa Min for the insightful discussions.

References

Zahra Ashktorab, Michael Desmond, Qian Pan, James M. Johnson, Martin Santillan Cooper, Elizabeth M. Daly, Rahul Nair, Tejaswini Pedapati, Swapnaja Achintalwar, and Werner Geyer. 2024. Aligning human and llm judgments: Insights from evalassist on task-specific evaluations and ai-assisted assessment strategy preferences. *Preprint*, arXiv:2410.00873.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, André F. T. Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2024. Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks. *Preprint*, arXiv:2406.18403.

Nicola Dainese, Alexander Ilin, and Pekka Marttinen. 2024. Can docstring reformulation with an LLM improve code generation? In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 296–312, St. Julian's, Malta. Association for Computational Linguistics.

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, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, and Ondrej Bojar. 2022. Findings of the 2022 Conference on Machine Translation (WMT22). *WMT*.

Tom Kocmi and Christian Federmann. GEMBA-MQM: Detecting Translation Quality Error Spans with GPT-4. Technical report.

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. *Preprint*, arXiv:2302.14520.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *Preprint*, arXiv:2412.05579.

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

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.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang Sandhini Agarwal Katarina Slama Alex Ray John Schulman Jacob Hilton Fraser Kelton Luke Miller Maddie Simens Amanda Askell, Peter Welinder Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *NeurIPS*.

Xingyao Wang, Sha Li, and Heng Ji. 2023. Code4Struct: Code generation for few-shot event structure prediction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3640–3663, Toronto, Canada. Association for Computational Linguistics.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. *Preprint*, arXiv:2301.07069.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

A New Evaluation Findings

Each category is assessed on a 100-point scale, allowing RUBRIC-MQM to offer richer system-level feedback by pinpointing the types and magnitudes of the committed errors. As the score from our evaluation naturally indicates the extent of errors, it can ultimately be used as a metric for ranking systems. As depicted in Figure 6, the ultimate score of `Reference A` is shown, categorized by both meta and sub-categories. The report highlights that the primary issue of this translation comes from ACCURACY. Nevertheless, it is essential to verify the outcome again after adequately tackling Overconfidence bias.

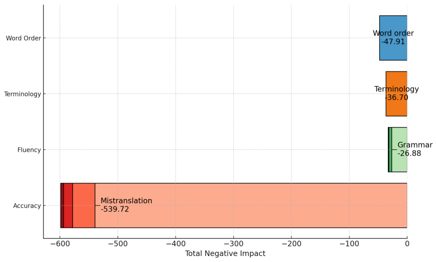


Figure 6: System-level score of Reference A (-716.54).

B Dataset

Table 7 provides comprehensive details regarding our dataset utilized in the principal experiment (GEN, PTB, and MIS) as well as in the subsequent analysis. The segments within the three benchmark datasets are distinct and unique.

B.1 GEN

The error taxonomy in this dataset encompasses 10 distinct types, as elaborated in Table 14. To ensure balanced label distribution, categories below 100 were supplemented with WMT 2022 test set. Nevertheless, the categories of PUNCTUATION, WORD ORDER, and UNTRANSLATED remain below 100, as indicated in Table 7.

B.2 PTB

The concept is derived from Quality Control of human evaluation presented in WMT 2020 (Barrault et al., 2020). We randomly select 500 sentences from the WMT 2023 Chinese-to-English evaluation set labeled with NO-ERROR by professional human evaluators. It is different from the conventional way of using a reference translation as a basis since our focus is to select error-free sentences. The construction of perturbed sentences is automatically done by selecting a random span proportional to the sentence length (in Table 6) and replacing it with phrases of the same length. As given in the example below, the green phrase from Sentence A is swapped with another phrase to make Sentence B. Considering the advanced performance of LLM, we avoid too easy options of short sentences (less than 5 words, i.e. `how are you?`). The focal point is that the phrase itself is a fluent sequence of words comprised of a high probability of tokens that will make the whole sentence significantly wrong (Barrault et al., 2020).

BP length (n)	#. replaced words in BP
6 - 8	3
9 - 15	4
16 - 20	5
> 20	$n/4$

Table 6: The number of words to swap in a sentence for perturbation (Barrault et al., 2020). Sentences that contain fewer than five words are excluded.

Example

Original Could you help follow up on it because I'm in a hurry, thank you.

Perturbed Could you help follow up on it because in the inspection shafts, thank you.

In such a setting, we expect that the model tags NO-ERROR for near-perfect sentences (*original*) and MAJOR for perturbed ones, given that NON-TRANSLATION is not an option in our task. While the primary focus is on the classification of Y_{sev}, we also elaborate on the model's selection of C_{cat}. A significant advantage of utilizing synthetic data is that it remains completely unexposed to the training dataset.

B.3 MIS

In the early phase of our study, most C_{cat} labels identified by the models were MISTRANSLATION. Thus, we attempted to understand the models' peak performance with this label by curating a dataset full of it.

B.4 Reference A

The initial dataset of the WMT 2023 consists of 1,996 sentences spanning 16 systems, not accounting for synthetic references. Upon the exclusion of sentences lacking human scores, the dataset is reduced to 884 sentences.

	GEN	PTB	MIS	Reference A
# Segment	964	1000	1000	884
Source length (avg.)	62.57	35.54	59.04	41.30
Source length (min)	3	4	7	1
Source length (max)	299	157	275	275
Target length (avg.)	39.20	22.53	37.81	25.92
Target length (min)	2	6	6	1
Target length (max)	177	125	129	127
# System	11	6	13	-
# Rater	8	8	8	-
Severity Type	Major, Minor	No-error, Major	Major, Minor	-
Size per label	500 / 464	500 / 500	500 / 500	-
Category Type	Omission, Mistranslation, Grammar, Addition, Source, Terminology, Inconsistency, Punctuation, Word Order, Untranslated	No-error, Mistranslation	Mistranslation	-
Size per label (detail)	Punctuation (96), Word Order (85), Untranslated (83)	500	1000	-

Table 7: Dataset overview.

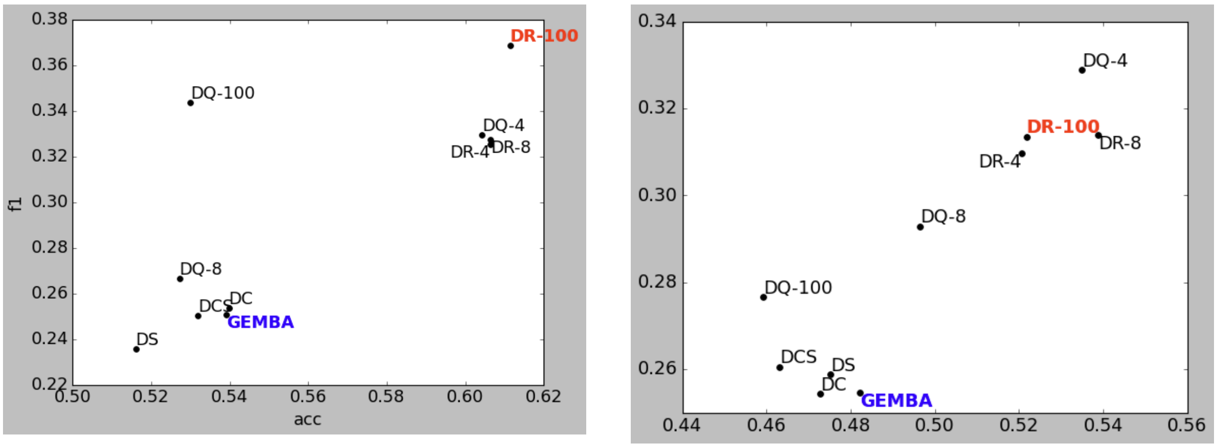
C Related Works

The GPT Estimation Metric Based Assessment (GEMBA) (Kocmi and Federmann, 2023a) was a pioneering initiative in employing LAJ in MT, offering an optimistic perspective for MT evaluation through the utilization of LLMs. The researchers posited that a model capable of translation could effectively discern between translations of varying quality. Based on this hypothesis, they investigated four distinct prompt designs. GEMBA-DA requested a score within the range of 0 to 100. GEMBA-SQM employed the same numerical scale but included continuous descriptive labels with each score. GEMBA-Stars implemented a star rating system to evaluate quality. GEMBA-Classes used labels without descriptions. GEMBA-DA, employing GPT-4 in a zero-shot context with a reference, exhibited superior accuracy when compared to SOTA metrics of WMT 22. This approach focuses on quality as the main evaluation criterion, presenting results as numerical scores.

In light of these significant results, AutoMQM (Fernandes et al., 2023) employed reasoning and ICL methodologies within the GEMBA-SQM prompting framework to augment interpretability throughout the evaluation process. Utilizing a predefined severity classification of MINOR/MAJOR, the model was asked to identify errors and assess their severity. Notably, the prompt lacked detailed categorization options, with guidance only available in a few-shot context. They removed unnecessary categories based on their criteria and computed MQM scores. The study revealed that specific zero-shot models derived from PaLM-2 attained the highest accuracy at the system level when references were incorporated. However, at the sentence level, achieving either accuracy or the Pearson correlation of the SOTA metrics required additional fine-tuning.

EAPrompt (Lu et al., 2024) employed the CoT prompting strategy within the AutoMQM framework, leveraging one-shot learning. The ICL example included source, reference, and translation segments with errors shown as per the specified format `{severity}: {error span} - {category}`. The task was subsequently divided into two distinct stages: (1) the identification of errors, guided by the AutoMQM instruction, and (2) the quantification of severity labels. This approach exhibited superior performance relative to GEMBA-DA in respect to sentence- and system-level accuracy. They reported that the task separation enhanced the model's focus on individual tasks.

The recently initiated project, designated GEMBA-MQM (Kocmi and Federmann), implemented a more stringent methodology concerning the skill set by enumerating a thorough list of valid error categories alongside their corresponding severities. These severity classifications were augmented to encompass CRITICAL, MAJOR, and



a) Y_{sev}

b) C_{cat}

Figure 7: Performance comparison of all slot scenarios. `DR-100` wins over all scenarios in Y_{sev}, and `DR` series outperform in C_{cat}.

MINOR, each briefly defined in the prompt. A core aspect of this method was using three ICL examples with different language pairs, enabling the model to attain results comparable to those of existing metrics across 15 high-resource languages.

Our study, closely aligned with several groundbreaking works, examines the use of LAJ-MT in MQM with the intention of surpassing the current SOTA evaluation metrics. **While they focus on a methodology-driven strategy for prompt engineering using advanced techniques like ICL or CoT, we intentionally shift focus to highlight the perceptual complexities of the evaluation context.** We are particularly focused on the prompts influencing critical skills in Severity and Category classification. Other elements, including ICL or detailed prompt parts such as indicating the source or target language (Zhang et al., 2023), fall beyond our scope. We aim to maximize the general-purpose LLM's effectiveness in MT evaluation within the defined limits of prompt engineering.

D Reasoning for MultiScale

Lu et al. (2024) highlights the subjectivity and unreliability in assigning a single score to a sentence. Consequently, MQM emerges as a viable alternative to DA by suggesting evidence through error spans and aggregating partial scores. Notwithstanding, we presume that challenges occur when fixed weights are used for predefined label sets. Indeed, MQM is hindered by its discrete scoring framework, possibly resulting in low correlations at the sentence level. Furthermore, the result of GEMBA-DA has demonstrated that predicting a single score often leads to outputs in multiples of five (Kocmi and Federmann, 2023a). To address these issues, we come to propose the application of the DA score scheme at the span level. We seek the best scale, starting from 4, reflecting the original MQM-TQE scheme and which is widely favored in assessment, to 8, which has recently gained popularity in human evaluations such as GEMBA-SQM, and further to 100, which is deemed the most intuitive and capable of encompassing more extensive ranges.

E Adjustment of Labels

MQM	Ours
Accuracy/Mistranslation	Mistranslation
Accuracy/Addition	Addition
Accuracy/Omission	Omission
Accuracy/Source language fragment	Untranslated
Fluency/Punctuation	Punctuation
Fluency/Grammar	Grammar
Fluency/Inconsistency	Inconsistency
Source Issue	Source Issue
Source Error	Source Issue
Style/Bad sentence structure	Word Order
Terminology/Inappropriate for context	Terminology
Terminology/Inconsistent	Terminology
No-error	No-error

E.1 Category Match

The MQM error typology is adaptable based on the evaluation context, including the language pair and evaluation purpose (Lommel et al., 2014). Consequently, we have organized our own set of error types that are broadly employed and can provide informative insights into the evaluation process. These types are predominantly sourced from MQM, although some have been removed or consolidated due to their rarity in the dataset. The category for our experiment thus comprises 10 items: Omission, Addition, Mistranslation, Grammar, Untranslated, Punctuation, Inconsistency, Source Issue, Word Order, and Terminology. Their definitions are detailed in Table 14.

The main feature of our labels is that most categories are language- and model-agnostic, found throughout the WMT dataset over many years. We have also excluded meta-category labels from the ICL examples, moving from Accuracy/Mistranslation to Mistranslation, since our preliminary study indicates they impair the perception of LLMs, outputting Accuracy/Punctuation, Style/Mistranslation, or Fluency/Accuracy, etc.. Finally, No-error is defined with the other terms, allowing the model to produce it separately.

E.2 Severity Match

We match all Severity labels to the original MQM dataset that has a binary division of Major/Minor. As elucidated in Table 8, when the predicted labels are discrete, Critical is regarded as Major. Otherwise, an optimal threshold is searched for each method that produces the highest accuracy in the given datasets. Severity criteria are compared in Figure 8 using a 100-point scale. GEMBA considers Major when the score is above 67, and `DR-100` sets 52/100 as its threshold.

Method	Threshold	
MQM	Major	Minor
GEMBA, DC, DS, DCS	Critical, Major	Minor
DR-4, DQ-4	$n \geq 3$	$n < 3$
DR-8, DQ-8	$n \geq 5$	$n < 5$
DR-100	$n \geq 52$	$n < 52$
DQ-100	$n \geq 34$	$n < 34$

Table 8: Ideal threshold of Major and Minor for each scenario.

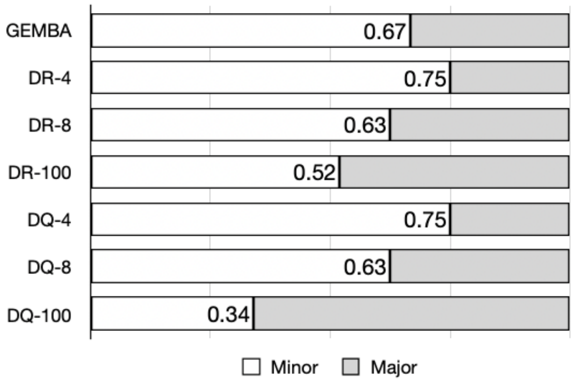


Figure 8: Numerical threshold delineating Major from Minor per scenario within the 0 to 1 interval.

		GEN										**PTB**										**MIS**									
		GEMBA	DC	DS	DCS	DR-4	DR-8	DR-100	DQ-4	DQ-8	DQ-100	GEMBA	DC	DS	DCS	DR-4	DR-8	DR-100	DQ-4	DQ-8	DQ-100	GEMBA	DC	DS	DCS	DR-4	DR-8	DR-100	DQ-4	DQ-8	DQ-100
Y_{sev}	Major	63.59	62.55	65.15	61.83	40.77	40.98	41.91	36.62	28.39	25.03	74.80	72.40	74.40	73.10	57.70	58.00	58.80	56.90	62.20	62.20	79.20	77.90	80.10	76.50	53.50	49.50	51.10	50.30	59.90	60.80
	Minor	18.57	20.75	15.66	20.33	39.52	37.45	37.66	34.13	14.83	12.93	17.10	18.30	13.70	17.20	27.00	24.80	23.80	19.20	16.10	18.40	14.90	16.00	9.20	14.70	35.20	38.50	36.70	37.20	27.90	25.90
	None	13.28	13.17	14.11	12.14	2.70	1.14	0.62	0.93	0.84	0.21	8.10	9.30	11.60	8.90	0.50	0.20	-	0.20	0.20	-	5.50	5.60	9.30	6.60	2.10	0.60	0.20	0.50	1.00	-
	No-error	4.56	3.53	5.08	5.71	17.01	20.44	19.81	28.32	55.94	61.83	-	-	0.30	0.80	14.80	17.00	17.40	23.70	21.50	19.40	0.40	0.50	1.40	2.20	9.20	11.40	12.00	12.00	11.20	13.30
C_{cat}	Mistranslation	42.63	38.69	38.49	32.78	42.43	43.05	45.54	35.89	24.19	21.45	55.50	55.20	55.20	53.70	54.10	60.00	53.80	54.60	54.00	50.20	76.50	70.60	71.30	64.80	74.10	75.60	73.20	71.10	72.90	69.60
	Omission	9.75	9.44	7.16	9.13	8.82	10.17	8.71	8.51	3.58	2.84	2.70	2.10	2.30	2.60	1.30	1.30	1.30	0.90	2.50	1.50	3.50	4.80	2.70	2.70	1.40	2.50	2.00	1.40	2.00	1.80
	Punctuation	9.54	11.00	10.68	11.72	6.02	5.29	5.08	5.81	3.79	2.31	3.60	3.70	2.90	3.80	0.40	0.30	0.40	0.40	0.30	0.30	0.60	0.50	0.20	1.00	0.10	0.10	-	0.40	0.20	0.10
	Terminology	5.91	7.16	5.71	8.30	6.54	4.56	4.36	6.33	6.62	4.31	0.50	1.70	0.40	2.10	1.10	0.70	1.20	0.20	0.70	0.70	6.80	10.10	6.70	10.30	6.80	4.10	5.90	7.30	5.70	8.70
	Addition	4.36	3.84	4.25	3.73	3.84	3.73	3.94	3.84	2.31	2.21	10.80	5.90	7.30	6.10	8.10	6.00	6.90	8.80	8.70	12.30	1.70	1.50	1.60	2.10	2.40	2.20	2.10	2.20	2.80	2.10
	Word order	3.01	5.08	6.85	7.88	5.60	4.46	4.56	4.56	1.05	1.47	10.30	14.60	16.50	16.20	13.90	9.90	12.10	8.00	8.00	9.00	1.10	1.40	3.00	5.90	1.60	1.70	1.80	1.90	2.00	1.20
	Grammar	2.80	2.90	1.87	3.42	2.59	3.32	3.94	2.59	1.68	1.79	3.60	3.30	1.60	2.00	2.00	2.10	2.20	1.60	2.40	3.50	1.80	3.20	2.10	2.90	2.20	0.90	1.90	2.50	2.00	2.40
	Untranslated	2.28	4.15	4.46	3.63	3.63	2.90	2.90	2.59	0.32	1.05	2.60	4.00	1.60	3.50	3.20	2.20	3.90	1.50	1.30	2.10	0.10	0.70	0.40	0.50	0.30	0.30	0.10	0.20	0.10	0.10
	Inconsistency	1.76	0.73	1.04	0.83	0.41	0.41	0.10	0.21	-	-	2.30	0.20	0.30	0.20	-	-	0.10		0.20	-	2.00	0.90	0.90	0.70	0.10	0.10	-	-	0.20	0.20
	Source issue	0.10	0.31	0.31	0.73	-	-	-	-	-	-	-	-	-	0.10	-	-	-	-	-	-	-	0.20	0.40	0.30	-	-	-	-	-	-

Table 9: Label distribution of all slot scenarios.

F Slot Scenarios

````
{source lang} source: ```{source sentence}```

{target lang} translation: ```{target sentence}```


Based on the source and machine translation segments surrounded with triple
backticks, identify error types in the segment and classify them.  The
categories of errors are:  accuracy (addition, mistranslation, omission,
untranslated text), fluency (grammar, inconsistency, punctuation), source
issue, incorrect word order, terminology inappropriate for context,
inconsistent use), or no-error.

Each error is classified as one of three categories:  critical, major,
and minor.  Critical errors inhibit comprehension of the text.  Major
errors disrupt the flow, but what the text is trying to say is still
understandable.  Minor errors are technically errors, but do not disrupt
the flow or hinder comprehension.


[ICL Examples]

{examples}


[Assistant’s Answer]
````

Table 10: Prompt template: GEMBA-MQM and DeepShot. The ICL examples vary between them.

````
Outlined below are the definition of translation errors across 12
categories including no-error.

[Error Category]

{definition}


[Instruction]

{source lang} source: ```{source sentence}```

{target lang} translation: ```{target sentence}```


Based on the source and machine translation segments surrounded with triple
backticks, identify error types in the segment and classify them.  We
would like you to classify the errors in the translation into addition,
mistranslation, omission, untranslated text, grammar, inconsistency,
punctuation, source issue, incorrect word order, terminology, or no-error,
according to the following definition:

Each error is classified as one of three categories:  critical, major,
and minor.  Critical errors inhibit comprehension of the text.  Major
errors disrupt the flow, but what the text is trying to say is still
understandable.  Minor errors are technically errors, but do not disrupt
the flow or hinder comprehension.


[ICL Examples]

{extended examples}


[Assistant’s Answer]
````

Table 11: Prompt template: DeepCat and DeepCatShot. The ICL examples vary between them.

````
Outlined below are the definition of a scale of severity of translation
errors.
[Scale of Error Severity]
{rubric}


[Instruction]
{source lang} source: ```{source sentence}```
{target lang} translation: ```{target sentence}```


Based on the source and machine translation segments surrounded with triple
backticks, identify error types in the segment and classify them.  The
categories of errors are:  addition, mistranslation, omission, untranslated
text, grammar, inconsistency, punctuation, source issue, incorrect word
order, terminology (inappropriate for context, inconsistent use), or
no-error.
Evaluate the severity of each error on a scale from 1 to {n} according to
the given rubric.


[ICL Examples]
{examples}


[Assistant's Answer]
````

Table 12: Prompt template: DeepRubric.

````
{source lang} source: ```{source sentence}```
{target lang} translation: ```{target sentence}```


Based on the source and machine translation segments surrounded with triple
backticks, identify error types in the segment and classify them.  The
categories of errors are:  addition, mistranslation, omission, untranslated
text, grammar, inconsistency, punctuation, source issue, incorrect word
order, terminology (inappropriate for context, inconsistent use), or
no-error.
Evaluate the severity of each error on a scale from 1 to {n}
{continuous line}.
[ICL Examples]
{examples}


[Assistant's Answer]
````

Table 13: Prompt template: DeepSQM.

Addition: This error occurs when extra content not in the original text leads to repetition, unnecessary details, or redundancy, distorting the message and potentially confusing readers or diverging from the original intent.

Mistranslation: This error involves inaccurate translation or interpretation, often due to poor word choice, leading to a message that strays from the original content's meaning and intent.

Omission: This error occurs when essential elements from the original text are missing in the translation, resulting in incomplete meaning and loss of critical information or nuances needed for full understanding.

Untranslated text: This error refers to parts of the source language that remain in the translation without being converted, resulting in an incomplete or inaccurate translation.

Grammar: This error involves incorrect grammar, such as tense, verb form, pronouns, agreement, articles, or gender, disrupting fluency and coherence and risking misunderstandings or credibility loss.

Inconsistency: It refers to variations in style or structure that undermine the fluency and readability of the translated text.

Punctuation: This error stems from incorrect punctuation, prepositions, quotation marks, or hyphenation, disrupting clarity and reading flow, and potentially causing misunderstandings.

Source issue: It refers to any problematic elements originating from the source text (i.e., ambiguities, grammatical errors, of unclear phrasing) that hinder accurate translation and lead to misunderstandings.

Incorrect word order: This error occurs when the translation fails to keep the original structure, order, or phrasing, which can alter the meaning, clarity, or emphasis, leading to awkward or confusing text.

Terminology: This error occurs when a term or word choice is contextually inappropriate or inconsistent, leading to misaligned meaning or intent and potentially causing confusion or lack of clarity, especially with technical or specialized terms.

No-error: This category denotes a flawless translation, accurately conveying the source text's meaning, tone, nuances, consistency, and style with clarity, cultural appropriateness, and grammatical accuracy in the target language.

Table 14: Definition for the **DEF** strategy. Comprehensive guidelines for categorization are essential, as baseline models frequently fail to incorporate this aspect. The objective is to test whether general-purpose models are capable of utilizing the information.

```
English source:  "`I do apologise about this, we must gain permission from <v>the account
holder</v> to discuss an order with another person, I apologise if this was done previously,
however, I would not be able to discuss this with yourself without the account holders
permission."`
German translation:  "`Ich entschuldige mich dafür, wir müssen die Erlaubnis einholen,
um eine Bestellung mit einer anderen Person zu besprechen.  Ich entschuldige mich<v>;</v>
falls dies zuvor geschehen <v>wäre</v>, aber ohne die Erlaubnis des Kontoinhabers wäre ich
nicht in der Lage, dies mit <v>dir</v> <v>involvement</v> <v>permission</v>."`
MQM annotations:
Critical:
no-error
Major:
mistranslation - "involvement"
punctuation - ";"
omission - "the account holder"
untranslated text - "permission"
Minor:
grammar - "wäre"

English source:  "`Talks have resumed in Vienna to <v>trying</v> to revive the nuclear pact,
with both sides trying to gauge the prospects of success after the latest exchanges in
<v>the stop-start</v> negotiations."`
Czech translation:  "`Ve Vídni se <v>ve Vídni</v> obnovily rozhovory o oživení jaderného
paktu, přičemž obě <v>partaje</v> se snaží posoudit vyhlídky na úspěch po posledních
výměnách v jednáních."`
MQM annotations:
Critical:
source issue - "trying"
Major:
addition - "ve Vídni"
omission - "the stop-start¨
Minor:
terminology - "partaje"

Chinese source: “‘大众点评乌鲁木齐家居卖场频道为您提供高铁居然之家地址，电话，营业时间等
 最新商户信息，找装修公司，就上大众点评“‘
English translation:  "`Urumqi Home Furnishing Store Channel provides <v>with you</v> the
latest business information such as the address, telephone number, business hours, <v>etc.,
</v> <v>of high-speed rail</v>, and find a decoration <v>incorporation</v>, and <v>go to
the reviews</v>."`
MQM annotations:
Critical:
addition - "of high-speed rail"
Major:
mistranslation - "go to the reviews"
Minor:
incorrect word order - "with you"
inconsistency - "incorporation"
```

Table 15: Extended ICL examples for **EXP** strategy, applied to DeepShot and DeepCatShot. The blue lines and simulated errors in the segments have been attached to the current lines.

DR-4

```
Evaluate the severity of each error on a scale from 1 to 4 according to the given rubric.
Scale 1:  The error slightly changes in wording with has no impact on message clarity or intent.
Scale 2:  The error makes some alteration of wording, but the overall message and intent remain
mostly clear.
Scale 3:  The error has noticeable impact on comprehension and may slightly distort the intended
message.
Scale 4:  The error substantially distorts the message, making the translation unfaithful and
potentially misleading.
```

DR-8

```
Evaluate the severity of each error on a scale from 1 to 8 according to the given rubric.
Scale 1:  The error has no impact on comprehension or intent.
Scale 2:  The error slightly alters wording but not the overall message.
Scale 3:  The error is somewhat affecting clarity but intent remains clear.
Scale 4:  The error impacts clarity and slightly distorts the message.
Scale 5:  The error affects understanding and partially alters intent.
Scale 6:  The error distorts meaning and message clarity is compromised.
Scale 7:  The error substantially misinterprets the message and intent.
Scale 8:  The error makes the translation unfaithful and misleading.
```

DR-100

```
Evaluate the severity of each error on a scale from 1 to 100 according to the given rubric.
Scale 10:  The error has negligible impact; the message and intent are unaffected.
Scale 20:  The error is tweaking some wording but leaving the overall message intact.
Scale 30:  The error has minimal effect on clarity; the intent remains clear.
Scale 40:  The error could lead to minor misunderstandings but overall message is still graspable.
Scale 50:  The error is affecting clarity; the message may require some interpretation.
Scale 60:  The error is distorting part of the message and intent can be ambiguous.
Scale 70:  The error is leading to misunderstandings and altering the message substantially.
Scale 80:  The error makes the core parts of the message misinterpretable, affecting communication.
Scale 90:  The error is causing serious miscommunication and loss of original intent.
Scale 100:  The error makes the translation completely unfaithful and misleading.
```

Table 16: Score rubric for **-R** strategy.

DQ-4

```
Evaluate the severity of each error on a scale from 1 to 4, where 1 starts on "minimal error
with no impact on clarity", goes to "minor alterations" and "noticeably impact comprehension",
up to 4, indicating "significant error substantially distort the message".
```

DQ-8

```
Evaluate the severity of each error on a scale from 1 to 8, that progresses from 1, where
the error has no impact on comprehension or intent, to 3, where it somewhat affects clarity
while intent remains clear, to 5, where it affects understanding and partially alters intent,
and finally to 8, where it makes the translation unfaithful and misleading.
```

DQ-100

```
Evaluate the severity of each error on a continuous scale from 1 to 100, that progresses from 10,
with negligible impact and the message intact, to 100, where the translation is completely
unfaithful and misleading, with intermediate levels introducing increasing challenges:  30 has minimal
clarity impact, 50 affects clarity and requires interpretation, 70 leads to substantial
and 90 results in serious miscommunication and intent loss.
```

Table 17: Continuous lines for **-Q** strategoy.