A Benchmark for Translations Across Styles and Language Variants

Xin Tan and **Bowei Zou** and **Ai Ti Aw**
Institute for Infocomm Research (I^{2}R), A*STAR, Singapore
{tan_xin,zou_bowei,aaiti}@i2r.a-star.edu.sg

Abstract

As machine translation (MT) rapidly advances in bridging global communication gaps, there is growing interest in *variety*-targeted translation for fine-grained language variants and specific translation styles. This translation variant aims to generate target outputs that are not only contextually accurate but also culturally sensitive. However, the lack of comprehensive evaluation benchmarks has hindered progress in this field. To bridge this gap, this work focuses on the translation across styles and language variants, aiming to establish a robust foundation for the automatic evaluation of fine-grained cultural and stylistic nuances, thereby fostering innovation in culturally sensitive translations. Specifically, we evaluate translations across four key dimensions: semantic preservation, cultural and regional specificity, expression style, and fluency at both the word and sentence levels. Through detailed human evaluations, we validate the high reliability of the proposed evaluation framework. On this basis, we thoroughly assess translations of state-of-the-art large language models (LLMs) for this task, highlighting their strengths and identifying areas for future improvement.

1 Introduction

Machine Translation (MT) has made significant strides in breaking down communication barriers around the world, particularly for widely spoken languages like Chinese and English at a broad level. As MT technologies continue to advance, there is growing interest in *variety*-targeted translation, targeting fine-grained language variants such as regional dialects (Kumar et al., 2021; Riley et al., 2023), and specialized stylistic adaptations, including formality-aware MT (Niu et al., 2017, 2018; Wang et al., 2019) and personalized MT (Michel and Neubig, 2018; Vincent, 2021). This evolution in MT aims to ensure that translations are not only contextually accurate but also culturally sensitive, thereby facilitating cross-cultural communication (Yao et al., 2024). The emphasis on integrating translations with different regions, cultural contexts, and specific styles highlights the unique challenges of this task compared to general machine translation. As a result, traditional evaluation metrics such as BLEU are no longer adequate to measure the quality of these fine-grained translations (Riley et al., 2023). Progress in this area has been hampered by the lack of comprehensive, high-quality evaluation benchmarks to assess stylistic and cultural variations in translations.

To bridge this gap, this work explores automatic evaluation metrics for translations across styles and language variants. Specifically, we focus on the translation scenario from English to Chinese variants, targeting social media translations in Mainland Mandarin (zh_CN), Taiwanese Mandarin (zh_TW), and the web-minority Singaporean Mandarin (zh_SG). To comprehensively capture cultural and regional nuances as well as the desired expression style in translations, we assess translations at both word and sentence levels across four key dimensions: semantic preservation, cultural and regional specificity, expression style, and fluency. At the word level, we evaluate lexical terms that explicitly reflect regional and cultural nuances, focusing on: 1) models' ability to accurately understand and translate region-specific vocabulary; 2) the alignment of lexical choices in models' translations with local references, showcasing its grasp of domain- or culture-specific expression patterns. At the sentence level, we leverage implicit linguistic expression features to evaluate the model's overall performance in meaning preservation, regional cultural adaptation, and expression style transfer.

In summary, the key contributions of this work are three-fold:

- We develop and release a benchmark for the translation across styles and language variants,

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 2389–2402

featuring several automatic evaluation metrics from linguistic perspectives, along with test sets that are manually annotated with region- and style-specific words.[1]

- We conduct detailed human evaluation across multiple evaluation dimensions, verifying the strong consistency between human judgments and the automatic metrics, thereby ensuring the high reliability of the proposed evaluation framework.

- Using the proposed evaluation framework, we provide a comprehensive assessment of predictions generated by several state-of-the-art large language models (LLMs), highlighting their strengths in this task and identifying directions for future improvement.

2 Related Work

2.1 Variety-Targeted Machine Translation

Nowadays, variety-targeted MT work mainly focuses on regions and styles. Among these, region-aware MT targets specific regions or dialects (Zbib et al., 2012; Baniata et al., 2018; Costa-jussà et al., 2018; Honnet et al., 2018; Chakraborty et al., 2018; Lakew et al., 2018; Sajjad et al., 2020; Wan et al., 2020; Kumar et al., 2021). Style-targeted MT has explored several subtypes such as formality-aware MT (Niu et al., 2017, 2018; Wang et al., 2019), which focuses on different levels of formality, and personalized MT (Michel and Neubig, 2018; Vincent, 2021), which aims to match an individual's specific style. These efforts contribute to more contextually appropriate and user-centric translations.

2.2 Cross-Cultural and Stylistic Evaluation

Evaluation on translations across cultural and stylistic boundaries remains underexplored. Yao et al. (2024) address cultural evaluation by focusing on culture-specific items, Riley et al. (2023) examine regional lexical and terminological variations. However, they focus on vocabulary-level differences and overlook finer-grained cultural, regional, and stylistic nuances embedded in discourse patterns, idiomatic expressions. Besides, research in text style transfer (TST), which aims to modify the stylistic properties (such as formality, politeness, and sentiment) of a sentence while preserving its core meaning, sharing important parallels with cross-cultural and -stylistic translation. Despite its contribution in evaluating content preservation, fluency, and style transfer (Li et al., 2018; Mir et al., 2019; Pryzant et al., 2020; Briakou et al., 2021), current TST evaluation remains limited in capturing cultural nuances.

To address these limitations, this work uniquely focuses on evaluating sensitivity to cross-cultural expressive styles, moving beyond superficial vocabulary differences. By capturing these nuances, our work introduces a comprehensive evaluation framework that goes beyond traditional MT metrics such as BLEU, providing a deeper assessment of the cultural adaptability and stylistic appropriateness of translations.

2.3 LLMs on Machine Translation

Large language models (LLMs), with billions of parameters and training on massive multilingual datasets, have shown promising results in the domain of MT. In addition to LLMs with strong multilingual translation capabilities, such as GPT-4o[2] and models designed specifically for translation-related tasks like TowerInstruct[3], there is a growing body of work exploring the translation capabilities of LLMs, particularly through techniques like fine-tuning, prompt engineering, and domain adaptation (Zhang et al., 2023; Bawden and Yvon, 2023; Vilar et al., 2023; Hendy et al., 2023; Lu et al., 2024; Zhu et al., 2024a; Zeng et al., 2024; Zhu et al., 2024b). The field of MT has undergone a dramatic transformation, achieving remarkable improvements in both fluency and contextual accuracy, steadily breaking down language barriers.

In contrast, traditional NMT systems lag behind LLMs, especially in variety-targeted MT, where the scarcity of large-scale training data limits their performance. Given this gap, this work focuses exclusively on LLMs, analyzing their relative strengths and limitations in facing linguistic diversity.

3 Variety-Targeted MT across Styles and Language Variants

3.1 Task Definition

General MT translates between coarse-grained language sentences. Given a source sentence $X = (x_1, x_2, .., x_n)$, a translation model generates the

[1] https://github.com/txAnnie/Evaluation-on-Variety-Targeted-MT.

[2] https://openai.com/index/hello-gpt-4o/

[3] https://huggingface.co/Unbabel/TowerInstruct-7B-v0.2

	General MT	Variety-Targeted MT across Styles and Languages
Translation Language	Coarse-grained languages. E.g., Chinese, English	Fine-grained language variants (regional dialects). E.g., Singaporean Mandarin, Taiwanese Mandarin
Translation Style	Remain source style	Specific style different from Source
Translation Focus	Word by word translation	Semantic translation

Table 1: A comparison of general and variety-targeted MT.

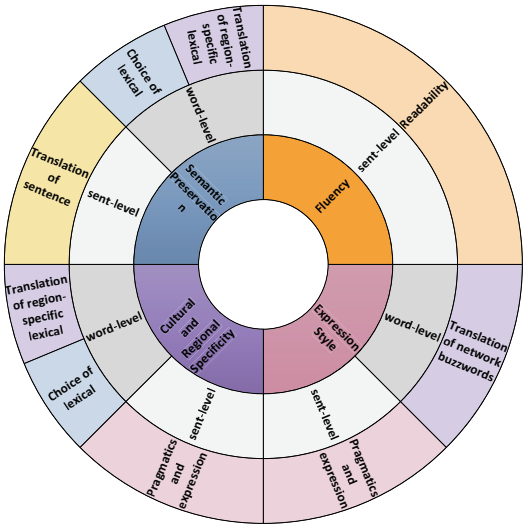

Figure 1: Four evaluation dimensions and their manifests at the word and sentence levels.

corresponding target sentence $\hat{Y} = (\hat{y}_1, \hat{y}_2, .., \hat{y}_m)$, prioritizing the semantic accuracy of the words.

In contrast, Variety-targeted MT goes beyond content preservation, adapting the source sentence $X = (x_1, x_2, .., x_n)$ into a target sentence $Y_T^{ES} = (y_1, y_2, ..., y_k)$ that retains the same semantic meaning while incorporating a distinct style ES suited to regional dialects or fine-grained language variants. Table 1 outlines the core differences. While general MT emphasizes literal or meaning-preserving translation between standard languages, variety-targeted MT demands context-sensitive adaptation at both the lexical and stylistic levels. This distinction makes it more challenging: the model must infer implicit style and variant cues and produce outputs that satisfy both semantic fidelity and stylistic conformity. This paper focuses on Chinese variants in social media scenarios, where style transformation involves: a) using appropriate slang and colloquialisms; b) adopting typical social media discourse patterns; and c) reflecting the cultural norms and sensitivities.

3.2 Evaluation Criteria

To evaluate whether a translation aligns with the intended cultural context, regional variation, and stylistic requirements, we assess outputs across four key dimensions: 1) Semantic Preservation. How well the core meaning of the source sentence is retained in the translation. 2) Cultural and Regional Specificity. Whether the translation reflects the appropriate regional dialect and culturally relevant expressions. 3) Expression Style. The degree to which the translation adopts target style, particularly social media discourse patterns and informal tone. 4) Fluency. The overall naturalness, grammaticality, and readability of the translation. These dimensions are assessed at both the word and sentence levels, as illustrated in Figure 1. Specifically: At the word level, we evaluate:

- Region-specific lexical term translation. The ability of a model to correctly translate region-specific vocabulary.
- Vocabulary similarity. The alignment of lexical choices with culturally preferred or regionally conventional terms.

At the sentence level, we assess:

- Semantic preservation. The extent to which the sentence meaning is retained.
- Cultural and style adaptation. The implicit adaptation of tone, idiomatic usage, and cultural references.
- Fluency. The sentence's coherence and grammatical correctness.

The dual-level evaluation provides a holistic view of both explicit lexical choices and implicit contextual appropriateness, to ensure that translations are not only accurate but also stylistically and culturally resonant.

3.3 Evaluation Metrics

To operationalize the five evaluation dimensions introduced above, we propose a set of automatic metrics.

Region-Specific Lexical Term Translation. Certain regions use unique lexical terms influenced by local culture. For example, in Singaporean Mandarin, the term "多多" refers to a lottery gaming activity. To assess whether the model correctly translates culturally or regionally distinctive terms, we annotate region-specific terms in the reference translations (refer to 3.5 for details) and calculate the match ratio between model output and reference. It allows for partial matches in semantically equivalent variants. For example, "多多" (ToTo) and "多多彩票" (ToTo lottery) share the same meaning, we allow partial matches to ensure evaluation flexibility.

$$score_{WR} = \frac{N_{L\_match}}{N_{L\_match}+N_{L\_mismatch}}, \quad (1)$$

where N_{L_match} and $N_{L\_mismatch}$ are the numbers of correctly and incorrectly translated annotated terms, respectively.

Vocabulary Similarity. Beyond marked terms, we assess how well the model aligns with region-preferred vocabulary. For instance, the expressions "一杯烧咖啡" in Singaporean Mandarin and "一杯热咖啡" in Mainland Mandarin both convey "a cup of hot coffee", but the terms "烧" and "热" are contextually fixed to their respective regions, reflecting distinct linguistic conventions. Key content words in the reference r_i and hypothesis h_i are identified using TF-IDF vectors [4], and a weighted match score is calculated as:

$$Match(h_i, r_i) = \frac{N_{V\_match}}{N_{V\_match}+N_{V\_mismatch}}, \quad (2)$$

where N_{V_match} and $N_{V\_mismatch}$ denote the number of key content words in the reference that are matched and unmatched in the hypothesis, respectively. While vocabulary similarity (e.g., word overlap) is useful, it may fail to capture semantically equivalent expressions. To mitigate this limitation, we incorporate semantic similarity, measured by TF-IDF vector cosine similarity (sim)[5], as a penalty weight to adjust the lexical match score. After empirical experiments, a threshold of 0.7 (very similar) is used:

$$sent_{score} = \begin{cases} Match(h_i, r_i), \text{if } sim \geq 0.7 \\ sim \cdot Match(h_i, r_i), \text{otherwise} \end{cases} \quad (3)$$

The final score is averaged at the sentence level across the corpus:

$$score_{WV} = (\sum sent_{score})/N \quad (4)$$

Semantic Preservation. Semantic preservation measures the similarity in content between reference translations and system-generated outputs. In general MT tasks, where high word-level overlaps are often required, BLEU (Papineni et al., 2002) is commonly employed as it evaluates *n*-gram overlaps between system outputs and reference translations. However, variety-targeted MT frequently involves variations in word choice and word order while preserving semantic meaning, which limits BLEU's effectiveness due to its inability to account for reordered words. In contrast, *chrF* (Popović, 2015), which evaluates character *n*-gram F-scores, has demonstrated a strong correlation with human judgments in the TST tasks (Briakou et al., 2021). Its ability to capture nuanced linguistic differences makes it well-suited for evaluating semantic preservation.

$$score_{SS} = (\sum chrF(r_i, h_i))/N \quad (5)$$

Cultural and Style Adaptation. Beyond explicit lexical elements, implicit features within contextual sentences play a key role in shaping subtle cultural nuances and stylistic traits. To automatically extract these features for assessing Cultural and Style Adaptation, we leverage a language model (LM) to classify whether translations satisfy the expected cultural and expressive style, inspired by the success of TST (Rao and Tetreault, 2018; Briakou et al., 2021). We fine-tune XLM-R[6] (Conneau et al., 2020), a multilingual pre-trained language model, using both human-written news and social media sentences in zh_CN, zh_SG, and zh_TW language variants (see Appendix A.1 for fine-tuning details). The fine-tuned XLM-R serves as a classifier C, which predicts the accuracy of model-generated translations r_i aligning with the desired language variant and expression style ES, as follows:

$$score_{SC} = (\sum N_{C(r_i)=ES})/N \quad (6)$$

Fluency. Fluency, also referred to as grammaticality, readability, and naturalness of a sentence (Mir et al., 2019), plays a crucial role in evaluating translation quality. Previous work on

[4] https://scikit-learn.org/

[5] We train TF-IDF vectors using the reference translations of our contributed benchmark dataset.

[6] https://huggingface.co/docs/transformers/model_doc/xlm-roberta

TST has validated fluency evaluation by measuring perplexity and likelihood scores (PPL) based on the probability distributions of language models (LMs) applied to model-generated outputs (Pang and Gimpel, 2019). In particular, (Briakou et al., 2021) demonstrated strong correlations with human judgments using pseudo-likelihood scores (PSEUDO-LL) derived from pre-trained masked XLM-R models[7]. Inspired by this, we adopt PSEUDO-LL for fluency evaluation of translations. Given PSEUDO-LL score P_i for each translation, we employ min-max normalization to obtain the corpus-level score:

$$Score_{SF} = (\sum \frac{P_i - min(P)}{max(P) - min(P)})/N \quad (7)$$

3.4 Evaluation Scenarios

Overall Assessment. The metrics described above reflect distinct aspects of the translations individually. To comprehensively evaluate the model's performance, it is essential to consider these metrics collectively, integrating their insights to provide a holistic assessment. To achieve this, we propose a combination method that rewards consistency across individual scores while penalizing substantial imbalances among them. Specifically, we first normalize the individual scores using min-max scaling to ensure all metrics are scaled to the same range and thus directly comparable.[8] Additionally, we introduce a penalty term p_o for the fusion of metrics from different perspectives. It is calculated as the mean absolute deviation (MAD) of the individual normalized scores $\hat{S}core_i$ ($i \in \{WR, WV, SS, SC, SF\}$) from their mean value $\overline{S}core$:

$$p_o = (\sum \left| \hat{S}core_i - \overline{S}core \right|)/5 \quad (8)$$

This penalty term highlights discrepancies between the metrics, ensuring a balanced and fair evaluation across different dimensions of translation quality. With the penalty term, we define the final overall score F_o as:

$$F_o = (\sum \hat{S}core_i - \omega \cdot p_o)/5 \quad (9)$$

where ω is a penalty weight[9].

While we encourage using the overall score F_o for a comprehensive assessment of translation quality, we also recognize that *variety*-targeted translation tasks may have varying requirements and that test sets in other languages may present unique challenges. Therefore, we provide additional assessments tailored to specific needs as following.

[7]XLM-R is trained in 100 languages, including zh_yue, zh, zh_min_nan, and zh_classical.

[8]The minimum and maximum scores are set to 0 and 100.

[9]The penalty weight is set to 0.1 to provide a moderate penalty for inconsistencies among individual metrics.

Language	Sent Num.	Avg Ref Len.	Lexical Num.
zh_CN	200	36.83	240
zh_TW	200	28.93	209
zh_SG	200	52.42	254

Table 2: Statistic on test sets. "Lexical Num." refers to the number of annotated region-specific lexical terms.

Word-Level Assessment. Evaluation metrics for Region-Specific Lexical Term Translation ($Score_{WR}$) and Vocabulary Similarity ($Score_{WV}$) provide detailed insights into translation quality at the lexical level. Together, these metrics offer complementary perspectives on the lexical fidelity and appropriateness of the translations, enabling a thorough word-level evaluation. Similar to overall assessment, to mitigate large discrepancies among the individual scores, we introduce the penalty term p_w, computed among normalized scores $\hat{S}core_w \in \{\hat{S}core_{WR}, \hat{S}core_{WV}\}$. And the word-level score is then calculated as:

$$F_w = (\sum \hat{S}core_w - \omega \cdot p_w)/2 \quad (10)$$

Sentence-Level Assessment. Evaluation metrics for Semantic Preservation ($Score_{SS}$), Cultural and Style Adaptation ($Score_{SC}$), and Fluency ($Score_{SF}$) together provide a comprehensive evaluation of sentence-level quality, reflecting both accuracy of the translation and the appropriateness of the cultural and style. Therefore, sentence-level score is computed based on the normalized individual scores $\hat{S}core_s \in \{\hat{S}core_{SS}, \hat{S}core_{SC}, \hat{S}core_{SF}\}$ and the penalty term p_s, calculated to account for discrepancies among these scores:

$$F_s = (\sum \hat{S}core_s - \omega \cdot p_s)/3 \quad (11)$$

Content Preservation Assessment. Beyond word- and sentence-level assessments, we also evaluate the preservation of overall content. This is achieved by combining the normalized Semantic Preservation score $\hat{S}core_{SS}$ and Region-Specific Lexical Term Translation score $\hat{S}core_{WR}$, capturing meaning preservation at both the sentence and word levels:

$$F_c = \text{avg}(\hat{S}core_{SS}, \hat{S}core_{WR}) \quad (12)$$

Prompt	{0}
Please perform region-aware formality-controlled translation on the following input by translating it into the style of {0}. Output translation only. **Input:** en_src **Output:** ref »»» **Input:** en_src **Output:** »»»	Informal Mainland Mandarin, i.e., speak Chinese on social media like people in Mainland China. Informal Taiwan Mandarin, i.e., speak Chinese on social media like people in Taiwan area. Informal Singaporean Mandarin, i.e., speak Chinese on social media like Singaporeans.

Table 3: Prompt used for translation generation.

3.5 Evaluation Sets

Social media language varies widely in different platforms, showcasing different dialects, slang, and idiomatic expressions that are unique to various cultural groups. To evaluate the sensitivity of translations across language variants and styles, we construct test sets for translation scenarios from English to social media style Mainland Mandarin (zh_CN), Taiwanese Mandarin (zh_TW), and Singaporean Mandarin (zh_SG) (mainly involves gossip and daily life domains). Specifically, we collect locally written sentences from social media platforms: zh_CN samples are sourced from Zhihu[10], zh_TW samples from PTT[11], and zh_SG samples from Facebook[12]. Two paid professional translators are hired to translate the social media sentences into English, creating corresponding en-zh_* sentence pairs[13]. To ensure the validity of word-level evaluation, region-specific lexical terms differing across regions are annotated based on online resources[14] and the expertise of the translators.

As a result, we construct three test sets, with detailed statistics provided in Table 2.

3.6 Human Judgments

To verify the alignment between human judgments and each of automatic evaluation metrics, we collect human ratings as follows:

- For *Semantic Preservation*, we adopt the Semantic Textual Similarity (STS) annotation scheme (Agirre et al., 2016). Model outputs are rated on a scale from 1 to 6 based on their degree of semantic similarity to the reference. The levels are: Completely dissimilar, Not equivalent but on same topic, Not equivalent but share some details, Roughly equivalent, Mostly equivalent, Completely equivalent.

- For *Cultural and Style Adaptation*, translations are annotated with both the language variant (zh_CN, zh_TW, zh_SG) and the level of style (news or social media).

- For *Fluency*, model outputs are rated on a discrete scale from 1 to 5 to indicate fluency degree (Heilman et al., 2014). The levels are: Other, Incomprehensible, Somewhat comprehensible, Comprehensible, Perfect.

- For *Region-Specific Lexical Term Translation*, binary labels (0 and 1) are used to indicate whether the marked lexical term in the translation matches the reference.

- For *Vocabulary Similarity*, we rate the model outputs on a discrete scale from 1 to 5 based on the degree of lexical similarity with the reference. The levels are: Completely dissimilar, Slightly similar, Moderately similar, Very similar, Identical.

The alignment between human judgments and automatic metrics is reported in Section 4.2.

4 Experimentation

4.1 Experimental Settings

Models. We evaluate several LLMs to verify the consistency between automatic metrics and human judgments. The selected models include the most advanced GPT-4o (2024-05-13) (OpenAI, 2024), open Llama Family (Llama3, 2024): Llama-3-8B-Instruct and Llama-3.2-3B-Instruct, Chinese and MT oriented LLMs: TowerInstruct-7b-v0.2 (Alves et al., 2024), QWen2.5-7B-Instruct (Qwen, 2025),

[10] https://www.zhihu.com/explore

[11] https://www.ptt.cc/index.html

[12] https://www.facebook.com/facebook/

[13] During annotation, all potential personal information disclosures and offensive content were manually removed to uphold ethical standards and ensure data integrity.

[14] https://www.languagecouncils.sg/mandarin/ch/, https://www.moedict.tw/ and https://www.digitaling.com/articles/381430.html

	Semantic Preservation	Vocabulary Similarity	Fluency	Region-Specific Lexical Term Translation	Culture and Style Adaptation
Spearman's ρ	0.57	0.61	0.60	-	-
Cohen's κ	-	-	-	0.90	0.79

Table 4: Correlation between human judgments and automatic evaluation metrics. Spearman's ρ is used to measure discrete human ratings and continuous metric scores; Cohen's κ is used to measure discrete human and metric ratings.

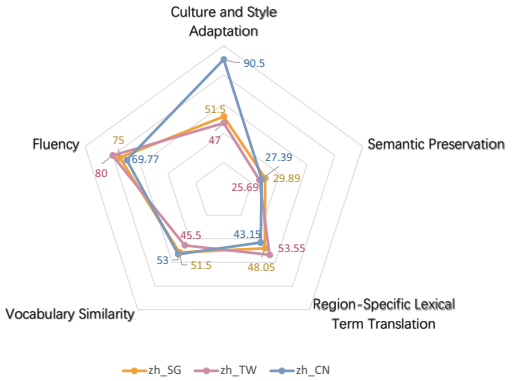

Figure 2: Comparison of individual evaluation metrics across three translation scenarios.

gemma-2-9b-it (Gemma, 2024), aya-expanse-8b (Aya, 2024), and Llama3-Chinese-8B-Instruct-v3 (Cui et al., 2024).

Parameters. For all the LLMs, cutoff_len=256 and do_sample=False during generation to reduce hallucinations and ensure deterministic outputs.

Prompts. We generate translations with 1-shot in-context learning. Table 3 lists the prompt used for this task.

4.2 Correlation Evaluation

We recruit three paid annotators, all familiar with both English and the Chinese variants, to evaluate the translation outputs of the aforementioned LLMs. The evaluation is conducted across three scenarios: en-zh_CN, en-zh_TW, and en-zh_SG. Each annotator assesses 50 randomly selected translations for each scenario, as described in Section 3.6. The annotations exhibit moderate inter-annotator agreement, ensuring the reliability of the human evaluation process. Table 4 reports the average correlation scores across annotators and the automatic metrics for a total of 150 selected translations.

For *Semantic Preservation*, *Vocabulary Similarity*, and *Fluency* metrics, we calculate the Spearman's ρ between human-annotated discrete scale labels and metrics-generated continuous scores. The correlation scores for these metrics all exceed 0.55, demonstrating a positive relationship between human and automatic evaluations. Additionally, a heatmap illustrating these correlation scores for each region is provided in Appendix A.2. For *Region-Specific Lexical Term Translation* and *Cultural and Style Adaptation* metrics, we compute Cohen's κ between human and metric-annotated discrete labels. The results indicate that the Kappa score for *Cultural and Style Adaptation* falls within substantial agreement (0.61-0.80). Notably, the correlation between human and metric evaluations for *Region-Specific Lexical Term Translation* achieves near-perfect agreement. Additionally, for *Cultural and Style Adaptation* indicator, we further assess correlations separately for language variant classification and expression style classification. The model's scores on F_1 for these classifications reach 93.24 and 91.70, respectively. Moreover, we analyze the translations with GEMBA-MQM (Kocmi and Federmann, 2023) and provide analysis examples in Appendix A.3.

All in all, these results highlight a strong alignment between human evaluations and automatic metrics, verifying the reliability of the proposed evaluation framework.

Moreover, we examine the independence and complementarity of the proposed metrics through the cross-metric Pearson correlation. The analysis in Appendix A.4 shows that these metrics are distinct yet correlated within a hierarchical assessment framework for translation quality, reflecting their ability to independently assess different aspects of translation while jointly contributing to the overall quality.

4.3 Analysis of LLM Gap in Cultural Language Understanding and Generation

We evaluate several recent LLMs on this task, grouping them into three categories for performance comparison in Table 5.

Comparing results across the three translation scenarios, LLMs generally perform better on en-zh_CN translations (average F_o =

	Model	Overall (F_o)	Sentence-Level (F_s)	Word-Level (F_w)	Content Preservation (F_c)
en-zh_CN	GPT-4o	**51.66**	**60.21**	**47.58**	**35.27**
	Llama3	33.75	52.08	23.29	16.57
	Llama3.2	24.87	42.97	14.68	10.23
	TowerInstruct-v0.2	31.16	48.68	20.56	14.82
	Qwen2.5	40.05	53.30	30.99	21.07
	Gemma2	44.58	55.62	39.19	27.40
	Aya	35.34	50.59	25.76	17.01
	Llama3-Chinese	36.88	55.83	25.79	18.45
en-zh_TW	GPT-4o	**42.07**	48.96	**49.12**	**39.62**
	Llama3	21.90	39.14	23.04	15.88
	Llama3.2	22.50	45.17	16.28	9.61
	TowerInstruct0.2	19.40	37.02	19.61	12.15
	Qwen2.5	25.49	39.69	28.19	18.74
	Gemma2	41.72	**52.68**	42.07	35.56
	Aya	21.98	35.78	26.52	17.70
	Llama3-Chinese	26.56	40.99	29.71	22.10
en-zh_SG	GPT-4o	**44.47**	50.61	**49.60**	**38.97**
	Llama3	27.62	47.26	19.50	14.64
	Llama3.2	25.25	**56.06**	13.82	9.75
	TowerInstruct0.2	28.77	54.69	20.93	14.27
	Qwen2.5	33.51	48.45	29.56	20.64
	Gemma2	32.92	50.67	24.50	17.56
	Aya	27.47	41.68	26.46	17.01
	Llama3-Chinese	28.20	44.09	23.76	16.29

Table 5: Results of evaluation metrics on diverse evaluation scenarios. All p-values (paired t-test) ≤ 0.05.

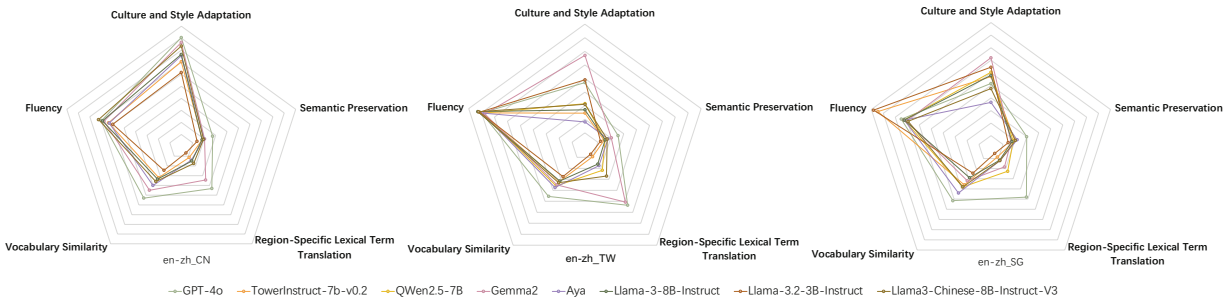

Figure 3: Comparison of individual metrics within each translation scenario.

37.29, F_s = 52.41) than on en-zh_TW (average F_o = 27.20, F_s = 42.43) and en-zh_SG (average F_o = 31.03, F_s = 49.18). Given GPT-4o's consistently strong performance across scenarios, we visualize its individual metric results in Figure 2 to examine its strengths and limitations. The figure shows that GPT-4o notably excels in sentence-level *Cultural and style Adaptation* for en-zh_CN translations, explaining its higher overall and sentence-level scores compared to en-zh_SG and -zh_TW. This advantages likely stems from training data predominantly composed of Mainland Mandarin, with limited exposure to Singaporean and Taiwanese Mandarin varieties. Meanwhile, GPT-4o's performance on other metrics remains relatively modest and consistent across all scenarios, revealing a key limitation in handling evolving slang and localized discourse practices across diverse cultural settings.

Comparing results within each translation scenario, we find that beyond GPT-4o's strong performance, Chinese and MT oriented LLMs (third group in each scenario) exhibit a clear advantage over general open models (Llama3 and Llama3.2) in capturing cross-cultural nuances, with Gemma2 being particularly notable. To further reveal the challenges faced by LLMs in this task, we visualize their performance across individual evaluation metrics in Figure 3[15]. While a few models show promise in identify cross-cultural discourse patten and idiomatic expressions (*Cultural and style Adaptation*), most struggle with word-level cultural nuances (*vocabulary Similarity*, *Region-Specific Vocabulary Term Translation*), reflecting insufficient background knowledge of LLMs. More importantly, Figure 3 reveals a fundamental and ongoing challenge: achieving cultural and stylistic adaptation without compromising semantic adequacy in

[15]Detailed results are listed in Appendix A.5.

cross-cultural and style-sensitive MT. This imbalance underscores the need for future work to effectively balance meaning preservation and culturally-aware adaptation to advance the development of translations across style and culture.

5 Conclusion

To fill the gap in a thorough evaluation of *variety*-targeted machine translation, this work proposes a benchmark for automatically assessing machine translation across language variants and styles. A detailed human assessment validates the high reliability of the proposed evaluation framework. Leveraging the proposed metrics, we perform a comprehensive evaluation of recent LLMs on this task and highlight key challenges for future research.

6 Limitations

We identify four main limitations of the proposed metrics:

Firstly, this study proposes an evaluation framework and test sets covering three Chinese variants: abundant Mainland Mandarin, few-shot Singaporean Mandarin, and Taiwanese Mandarin. These Chinese variants provide a rich testbed due to their distinct lexical, stylistic, and cultural differences. By establishing this comprehensive evaluation framework, we aim to lay the foundation for adapting the metric to other language pairs in the future. In particular, we plan to explore diverse language families, such as European Portuguese vs. Brazilian Portuguese, Canadian French vs. European French, which exhibit structural and cultural distinctions different from Chinese, thereby broadening the applicability of the metric. To achieve that, we plan to implement word-level metrics in a human-in-the-loop workflow: 1) Leveraging large region-specific corpora to automatically identify candidate dialectal terms using statistical methods such as PMI to detect words strongly associated with a specific region and 2) Automatically generating candidate lists for human annotators for efficient validation and refinement to maintain high-quality standards. Additionally, while the current test set is carefully curated with an emphasis on quality and detailed annotations (Sections 3.5) capture subtle phenomena like cultural and stylistic adaptation, we acknowledge the importance of scaling it further. Moving forward, we will continue to expand the test set and advance this line of research.

Secondly, despite our careful selection of source texts from local social media content and professional translation efforts to preserve style, cultural context, and dialectal features, translating already translated texts may still pose limitations in fidelity and naturalness. However, this also implies that although LLMs may have seen the original Chinese posts from Zhihu, PTT, or Facebook in their training data, it is highly unlikely that they were exposed to the professionally translated English source sentences we specifically created for the benchmark, which minimizes the risk of data contamination and helps ensure the reliability of the experimental results.

Thirdly, while the framework focuses on cultural and expression style transfer, *variety*-targeted machine translation encompasses a broader spectrum of styles, such as politeness and personalized tones. The current approach does not account for all these styles, limiting its ability to evaluate customized translations comprehensively.

Fourthly, we rely on in-context learning to assess large language models (LLMs) rather than fine-tuned models specifically optimized for this task. As a result, the LLMs' potential performance may not be fully reflected in the evaluation.

Acknowledgments

This research is supported by the National Research Foundation, Singapore under its National Large Language Models Funding Initiative. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not reflect the views of the National Research Foundation, Singapore.

References

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California. Association for Computational Linguistics.

Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, et al. 2024. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.

Aya. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Laith H. Baniata, Se-Young Park, and Seong-Bae Park. 2018. A neural machine translation model for arabic dialects that utilises multitask learning (mtl). *Computational Intelligence and Neuroscience*, 2018.

Rachel Bawden and François Yvon. 2023. Investigating the translation performance of a large multilingual language model: the case of BLOOM. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 157–170, Tampere, Finland. European Association for Machine Translation.

Eleftheria Briakou, Sweta Agrawal, Joel Tetreault, and Marine Carpuat. 2021. Evaluating the evaluation metrics for style transfer: A case study in multilingual formality transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1321–1336, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Saurav Chakraborty, Anup Sinha, and Sanghamitra Nath. 2018. A bengali-sylheti rule-based dialect translation system: Proposal and preliminary system. In *Proceedings of the International Conference on Computing and Communication Systems: I3CS 2016, NEHU, Shillong, India*, pages 451–460. Springer.

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

Marta R. Costa-jussà, Marcos Zampieri, and Santanu Pal. 2018. A neural approach to language variety translation. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 275–282, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yiming Cui, Ziqing Yang, and Xin Yao. 2024. Efficient and effective text encoding for chinese llama and alpaca. *Preprint*, arXiv:2304.08177.

Gemma. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Michael Heilman, Aoife Cahill, Nitin Madnani, Melissa Lopez, Matthew Mulholland, and Joel Tetreault. 2014. Predicting grammaticality on an ordinal scale. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 174–180, Baltimore, Maryland. Association for Computational Linguistics.

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*.

Pierre-Edouard Honnet, Andrei Popescu-Belis, Claudiu Musat, and Michael Baeriswyl. 2018. Machine translation of low-resource spoken dialects: Strategies for normalizing Swiss German. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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

Sachin Kumar, Antonios Anastasopoulos, Shuly Wintner, and Yulia Tsvetkov. 2021. Machine translation into low-resource language varieties. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 110–121, Online. Association for Computational Linguistics.

Surafel Melaku Lakew, Aliia Erofeeva, and Marcello Federico. 2018. Neural machine translation into language varieties. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 156–164, Brussels, Belgium. Association for Computational Linguistics.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874, New Orleans, Louisiana. Association for Computational Linguistics.

Llama3. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Hongyuan Lu, Haoran Yang, Haoyang Huang, Dongdong Zhang, Wai Lam, and Furu Wei. 2024. Chain-of-dictionary prompting elicits translation in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 958–976, Miami, Florida, USA. Association for Computational Linguistics.

Paul Michel and Graham Neubig. 2018. Extreme adaptation for personalized neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 312–318, Melbourne, Australia. Association for Computational Linguistics.

Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 2019. Evaluating style transfer for text. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 495–504, Minneapolis, Minnesota. Association for Computational Linguistics.

Xing Niu, Marianna Martindale, and Marine Carpuat. 2017. A study of style in machine translation: Controlling the formality of machine translation output. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2814–2819, Copenhagen, Denmark. Association for Computational Linguistics.

Xing Niu, Sudha Rao, and Marine Carpuat. 2018. Multi-task neural models for translating between styles within and across languages. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1008–1021, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Richard Yuanzhe Pang and Kevin Gimpel. 2019. Unsupervised evaluation metrics and learning criteria for non-parallel textual transfer. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 138–147, Hong Kong. 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.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. In *Proceedings of the aaai conference on artificial intelligence*, volume 34, pages 480–489.

Qwen. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

Parker Riley, Timothy Dozat, Jan A. Botha, Xavier Garcia, Dan Garrette, Jason Riesa, Orhan Firat, and Noah Constant. 2023. FRMT: A benchmark for few-shot region-aware machine translation. *Transactions of the Association for Computational Linguistics*, 11:671–685.

Hassan Sajjad, Ahmed Abdelali, Nadir Durrani, and Fahim Dalvi. 2020. AraBench: Benchmarking dialectal Arabic-English machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5094–5107, Barcelona, Spain (Online). International Committee on Computational Linguistics.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.

Sebastian Vincent. 2021. Towards personalised and document-level machine translation of dialogue. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 137–147, Online. Association for Computational Linguistics.

Yu Wan, Baosong Yang, Derek F Wong, Lidia S Chao, Haihua Du, and Ben CH Ao. 2020. Unsupervised neural dialect translation with commonality and diversity modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9130–9137.

Yunli Wang, Yu Wu, Lili Mou, Zhoujun Li, and Wenhan Chao. 2019. Harnessing pre-trained neural networks with rules for formality style transfer. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3573–3578, Hong Kong, China. Association for Computational Linguistics.

Binwei Yao, Ming Jiang, Tara Bobinac, Diyi Yang, and Junjie Hu. 2024. Benchmarking machine translation with cultural awareness. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13078–13096, Miami, Florida, USA. Association for Computational Linguistics.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan, and Chris Callison-Burch. 2012. Machine translation of Arabic dialects. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 49–59, Montréal, Canada. Association for Computational Linguistics.

Jiali Zeng, Fandong Meng, Yongjing Yin, and Jie Zhou. 2024. Improving machine translation with large language models: A preliminary study with cooperative decoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13275–13288, Bangkok, Thailand. Association for Computational Linguistics.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning*, pages 41092–41110. PMLR.

Shaolin Zhu, Leiyu Pan, Bo Li, and Deyi Xiong. 2024a. LANDeRMT: Dectecting and routing language-aware neurons for selectively finetuning LLMs to machine translation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12135–12148, Bangkok, Thailand. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024b. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

A Appendix

A.1 Fine-Tune XLM-R for Cultural and Style Adaptation Evaluation

To enable XLM-R to identify cultural and stylistic diversities, we employ LoRA fine-tuning on XLM-R for 5 epochs (learning_rate=5×10^{-5}, batch_size=32, shuffle_seed=42, max_seq_length=128) using a dataset of total 10,000 examples with the following labels:

> Label 0: zh_CN social media comments from Zhihu (`https://www.zhihu.com/explore`);
>
> Label 1: zh_SG social media comments from Facebook (`https://www.facebook.com/facebook/`);
>
> Label 2: zh_TW social media comments from PTT (`https://www.ptt.cc/index.html`);
>
> Label 3: zh_CN news sentences from voachinese (`https://www.voachinese.com/China`);
>
> Label 4: zh_SG news sentences from zaobao (`https://www.zaobao.com.sg/`);
>
> Label 5: zh_TW news sentences from twreporter (`https://www.twreporter.org/`)

The fine-tuned XLM-R achieves an accuracy of 97.07% on a dev set consisting of 6,000 sentences (each label 1,000 sentences).

A.2 Spearman's ρ on Each Translation Scenario

Detailed Spearman's ρ between human-annotated discrete scale labels and metrics-generated continuous scores for each translation scenario is shown in Figure 4.

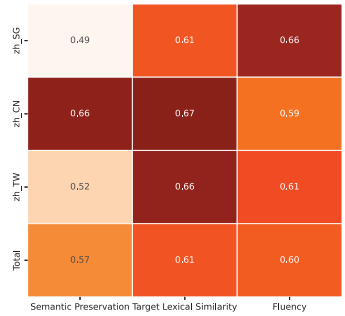

Figure 4: Spearman's ρ between human judgments and automatic metrics on three translation scenarios.

A.3 Analysis with GEMBA-MQM

We analyze the translations using GEMBA-MQM (Kocmi and Federmann, 2023). To adapt GEMBA-MQM for this task, we modify the prompt as follows:

> ***source_lang*** *source:* ***source_seg***
>
> ***target_lang*** *translation:* ***target_seg***
>
> *Based on the source segment and machine translation surrounded with triple backticks, identify error types in the translation and classify them. The categories of errors are: accuracy (addition, mistranslation, omission, untranslated text), fluency (character encoding, grammar, inconsistency, punctuation, register, spelling), style (non-informal Mainland Mandarin expressions), terminology (inappropriate for context, inconsistent use), non-translation, other, 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.*

	Source	Translation	GEMBA-MQM Analysis
zh_CN	I was working in Singapore in 1969. Back then, I often saw him in 7th month getai. He is getting old now.	我1969年在新加坡工作过。那个时候我经常在七月中秋看到他。他现在都老了。	Critical:no-error Major:accuracy/mistranslation- "七月中秋" should be "七月歌台" or "七月戏台" as "getai" refers to a live stage performance held during the Hungry Ghost Festival in the seventh lunar month, not "Mid-Autumn" which is "中秋". Minor:no-error
zh_TW	Your values have changed. Also, all-you-can-eat buffets are not the trend at all, okay?	妳的設定改變了喔！而且吃到飽根本不是流行趨勢啦，好嗎？	Critical: no-error Major: accuracy/mistranslation - "妳的設定改變了喔" should be something like "妳的價值觀改變了" as "values" in this context refers to "values" not "settings". Minor: no-error The mistranslation of "values" to "settings" (設定) could potentially lead to confusion about the meaning of the sentence, as it changes the subject from personal beliefs or principles to something more akin to preferences or configurations. This is a major error because it could significantly alter the intended message of the source text.
zh_SG	Captured a highly intelligent, hilarious CJ7 that speaks human languages. Please do not fight with me over it. Thanks!	抓到了一只超聪明、会说人话的搞笑长江七号。别跟我抢哦，谢谢！	Critical: no-error Major: accuracy/mistranslation - "搞笑" should be "可笑的" or "滑稽的" to better match "hilarious" instead of "funny" which "搞笑" implies. Minor: no-error

Table 6: Analysis with GEMBA-MQM.

As shown in Table 6, we found that GEMBA-MQM has certain ability to identify translation errors, along with region-specific lexical choices and expressions, although some of them are classified as translation inaccuracies.

A.4 Analysis on Independence and Complementarity of Metrics

We conduct a cross-metric Pearson correlation analysis. As shown in Table 7, word-level metrics (Region-Specific Lexical Term Translation and Vocabulary Similarity) are strongly correlated with the sentence-level Semantic Preservation metric ($r = 0.74$ and $r = 0.75$), reflecting the interconnected nature of translation quality. This suggests that while these word-level metrics independently assess explict lexicla choices, they also contribute substantially to the evaluation of overall sentence-level contextual adequacy. Moreover, Culture and Style Adaptation shows moderate correlations with meaning-oriented metrics: Region-Specific Lexical Term Translation, Vocabulary Similarity, and Semantic Preservation ($r = 0.41$ to 0.67), indicating an added cultural dimension beyond semantics and vocabulary. By contrast, Fluency exhibits negative correlations with the other metrics ($r = -0.27$ to -0.59), highlighting it as a distinct and sometimes competing quality dimension.

Overall, these metrics are independent yet complementary, collectively providing a comprehensive assessment of translation quality.

A.5 Results on Individual Evaluation Metrics

Detailed results of LLMs on individual evaluation metrics are presented in Table 8.

	Culture and Style Adaptation	Semantic Preservation	Region-Specific Lexical Term Translation	Vocabulary Similarity	Fluency
Culture and Style Adaptation	1.00	0.67	0.41	0.51	-0.59
Semantic Preservation	0.67	1.00	0.74	0.75	-0.46
Region-Specific Lexical Term Translation	0.41	0.74	1.00	0.60	-0.27
Vocabulary Similarity	0.51	0.75	0.60	1.00	-0.27
Fluency	-0.59	-0.46	-0.27	-0.27	1.00

Table 7: Cross-Metric Pearson Correlation Results.

Translation Task	Model	Word-Level Metric		Sentence-Level Metric		
		Region-Specific Lexical Term Translation	Vocabulary Similarity	Semantic Preservation	Culture and Style Adaptation	Fluency
en-zh_CN	GPT-4o	**43.15**	**53.00**	**27.39**	**90.50**	69.77
	Llama3	14.94	33.50	18.19	76.50	68.81
	Llama3.2	6.64	24.50	13.82	61.50	59.85
	TowerInstruct-v0.2	11.20	32.00	18.44	70.50	63.59
	Qwen2.5	21.58	42.50	20.56	84.00	62.35
	Gemma2	34.44	45.00	20.36	86.50	67.55
	Aya	14.11	40.00	19.91	75.50	62.94
	Llama3-Chinese	17.43	36.00	19.46	83.50	**72.33**
en-zh_TW	GPT-4o	**53.55**	**45.50**	**25.69**	47.00	80.00
	Llama3	16.11	31.50	15.64	27.00	**83.01**
	Llama3.2	7.11	27.50	12.11	49.00	81.49
	TowerInstruct-v0.2	9.48	32.00	14.82	24.50	79.76
	Qwen2.5	21.80	36.00	15.67	31.50	79.34
	Gemma2	50.71	35.00	20.40	**67.00**	77.56
	Aya	17.54	37.50	17.86	18.00	79.71
	Llama3-Chinese	27.01	33.00	17.19	31.00	82.58
en-zh_SG	GPT-4o	**48.05**	**51.50**	**29.89**	51.50	75.00
	Llama3	11.72	29.00	17.56	58.00	72.59
	Llama3.2	5.08	24.50	14.42	64.50	**98.18**
	TowerInstruct-v0.2	8.59	36.00	19.95	57.00	94.60
	Qwen2.5	22.66	38.00	18.62	60.50	72.62
	Gemma2	18.36	32.00	16.76	**72.00**	70.50
	Aya	12.11	44.00	21.90	36.50	72.39
	Llama3-Chinese	12.11	38.00	20.47	47.50	69.34

Table 8: Results of individual evaluation metrics.