Culture-Aware Machine Translation in Large Language Models: Benchmarking and Investigation

Anonymous ACL submission

Abstract

Large language models (LLMs) have achieved strong performance in general machine translation, yet their ability in culture-aware scenarios remain poorly understood. To bridge this gap, we introduce **CanMT**, a **C**ulture-**A**ware **N**ovel-Driven Parallel Dataset for **M**achine **T**ranslation, together with a theoretically grounded, multi-dimensional evaluation framework for assessing cultural translation quality. Leveraging **CanMT**, we systematically evaluate a wide range of LLMs and translation systems under different translation strategy constraints. Our findings reveal substantial performance disparities across models and demonstrate that translation strategies exert a systematic influence on model behavior. Further analysis show that translation difficulty varies across types of culture-specific items, and that a persistent gap remains between models' recognition of culture-specific knowledge and their ability to correctly operationalize it in translation outputs. In addition, incorporating reference translations is shown to substantially improve evaluation reliability in *LLM-as-a-judge*, underscoring their essential role in assessing culture-aware translation quality.

1 Introduction

In the era of globalization, Machine Translation (MT) has become a cornerstone technology for cross-lingual communication and global information exchange (Bahdanau et al., 2014; Ye et al., 2024a; Gain et al., 2025). Recent advances in large language models (LLMs) (OpenAI et al., 2024) have further reshaped the MT landscape, positioning LLM-based translation as an increasingly dominant paradigm (Hendy et al., 2023; Jiao et al., 2023; Zhu et al., 2024; Huang et al., 2024).

Despite the impressive progress in general-purpose translation, existing studies on LLM-based MT primarily focus on literal-level translation quality, such as adequacy and fluency. However, language is deeply intertwined with culture, and effective translation often requires more than lexical or syntactic equivalence.

Prior work has advanced cultural translation evaluation, but most approaches focusing on narrowly defined domains such as cooking recipes (Cao et al., 2023; Hu et al., 2024; Zhang et al., 2024), proverbs (Wang et al., 2025), idioms (Li et al., 2023) and poetry (Chen et al., 2025). In these datasets, cultural cues are concentrated in a small set of genre-specific features, limiting the evaluation of models in broader, unstructured contexts. Yao et al. (2024) construct parallel corpora from Wikipedia, but the resulting data are largely informational with homogeneous pragmatic and syntactic patterns, limiting the evaluation of models in more diverse, naturalistic contexts. More recently, Zhang et al. (2025) introduces parallel data from bilingual web novels; however, their corpus is limited to Zh→En, which prevents a thorough evaluation of cultural transfer across diverse cultures (See Table 1).

To bridge this gap, we propose a **C**ulture-**A**ware **N**ovel-Driven Parallel Dataset for **M**achine **T**ranslation (**CanMT**), a parallel corpus constructed from a diverse set of novels and their professional translations. For evaluation, we define assessment dimensions grounded in established translation studies theories, capturing translation quality from multiple perspectives, including contextual accuracy (Halliday, 1978), cultural adaptation (Venuti, 2008), functional equivalence (Nida, 1964), fidelity (Newmark, 1988), and naturalness (Newmark, 1988).

Building on **CanMT**, we conduct a series of experiments to systematically study culture-aware translation. First, we evaluate and compare the translation performance of a range of models and translation systems. Second, we investigate how different translation strategies (Communicative Translation and Semantic Translation) affect trans-

Benchmark	Evaluation Focus	Data Source	# Translation Directions
CulturalRecipes (Cao et al., 2023)	Cultural Adaptation in Recipes	Cooking Recipes	2
MAPS (Wang et al., 2025)	Proverb Translation (dialogue context)	Proverbs	8
PoetMT (Chen et al., 2025)	Poetry Translation	Poems	1
DITING (Zhang et al., 2025)	Novel Translation	Novels	1
CAMT (Yao et al., 2024)	Cultural-Specific Items (word / phrase-level)	Wikipedia	12
CanMT (Ours)	Sentences in Diverse Cultural Scenarios	Novels	12

Table 1: Comparison of related translation benchmarks.

lation quality (Newmark, 1981). Our results indicate that culture-aware translation performance generally improves with model scaling up. Furthermore, "Test-time Scaling" Reasoning enables consistent and incremental gains. In terms of translation strategies, communicative strategy tend to enhance fluency and functional adequacy, whereas semantic strategy prioritize accurate meaning transfer during translation.

In addition to the main experiments, we conducted several analysis. First, regarding strategy preference, similarity analysis reveals a systemic bias toward semantic translation in default translation settings, suggesting that LLMs tend to adopt a translation strategy that is closer to a semantic-oriented approach; Second, across culture-specific items (CSIs) (Newmark, 1988) categories, we identify a clear difficulty hierarchy: models excel at geographic and ecological terms but struggle significantly with nuanced linguistic-symbolic items; Third, by probing the relationship between knowledge and performance, we find that while correct cultural knowledge generally aids translation, a persistent "knowledge-application gap" remains; Finally, we demonstrate that reference translations are vital for calibrating automatic evaluators, as their inclusion consistently improves alignment with human judgment across cultural dimensions.

Overall, this work introduces the **CanMT** benchmark for investigating culture-aware translation in LLMs, enabling a systematic evaluation of model behavior in culturally grounded translation settings. Our empirical analysis provide detailed insights into models' behavior in culture-aware translation, and establish a reference point for future research on cultural adaptability.

2 Related Works

Culture-aware Machine Translation. Despite the rapid progress of LLMs in general multilingual capabilities (Qin et al., 2024; Ye et al., 2025b,a), their ability to accurately convey culture-specific meanings remains insufficiently understood. Culture-aware machine translation aims to preserve culturally grounded meanings across languages. Existing work can be broadly grouped into two categories. The first focuses on specific culturally relevant linguistic phenomena, such as recipes (Cao et al., 2023; Hu et al., 2024; Zhang et al., 2024), proverbs (Wang et al., 2025), idioms (Li et al., 2023), novels (Zhang et al., 2025), and poetry (Chen et al., 2025). The second utilizes parallel corpora enriched with cultural information to construct benchmarks (Yao et al., 2024; Singh et al., 2024) and Ye et al. (2024b) proposed $\mathcal{X}$Transplant for cross-lingual complementarity in culture scenarios. Among them, Yao et al. (2024) introduced the CAMT dataset, which is primarily derived from Wikipedia. While Wikipedia provides well-aligned parallel data, its predominantly factual content offers limited coverage of the narrative structures and stylistic variation characteristic of literary texts. In contrast, **CanMT** leverages novels originating from the target culture, encompassing richer cultural expressions and more diverse linguistic styles, and thereby enabling a more comprehensive evaluation.

LLM-as-a-Judge. Due to the complexity and diversity of cultural translation, we employ an LLM-based approach to evaluate models' translation across multiple dimensions.Recently, the use of large language models for machine translation evaluation has gained popularity. Kocmi and Federmann (2023b) proposed GEMBA, demonstrating the potential of GPT-4 in assessing MT quality. Building on this, Kocmi and Federmann (2023a) combined GEMBA with the MQM evaluation framework, using error span detection to evaluate MT outputs. Feng et al. (2025) leverages a multi-agent debate mechanism to assess translation quality from multiple dimensions using error detection strategies. Our approach adopts a simpler, single-judge, multi-dimensional evaluation

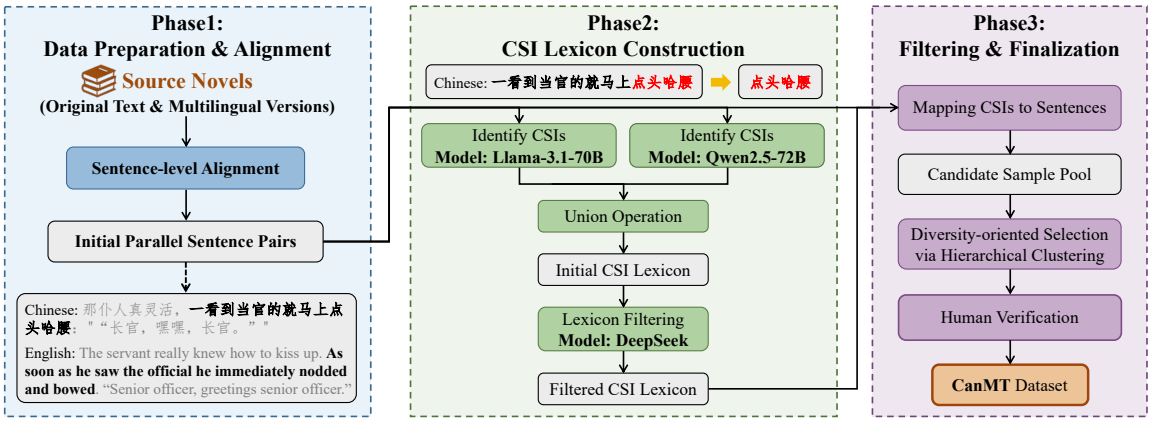

Figure 1: Overview of the data processing pipeline, including text extraction, preprocessing, and preparation of parallel text pairs for evaluation.

Country	Language	Name of Novels
America	English	*"Adventures of Huckleberry Finn"*
	English	*"The Great Gatsby"*
China	Chinese	《活着》 *(English: "Live")*
Russia	Russian	« Анна Каренина » *(English: "Anna Karenina")*
Japan	Japanese	『古都』 *(English: "The Old Capital")*
Spain	Spanish	« Don Quijote de la Mancha » *(English: "Don Quixote")*

Figure 2: Representative novels from diverse cultural backgrounds used as sources for building our novel-based culture-aware translation dataset.

Direction	En→Es	En→Ja	En→Zh	Zh→Ja	Zh→Ru	Zh→Es
# Samples	106	103	111	137	95	146
# CSIs	143	138	148	258	168	272
Direction	Es→En	Ja→En	Zh→En	Ja→Zh	Ru→Zh	Es→Zh
# Samples	134	116	125	106	125	166
# CSIs	300	255	228	221	226	368

Table 2: Statistics of the **CanMT** across translation directions, reporting the number of sample pairs and CSIs.

setup that is cost-efficient and more controlled: we provide the LLM with a detailed scoring rubric in the prompt, which allows a focused analysis of evaluation dimensions and translation objectives in cultural translation.

3 Benchmark Construction

3.1 CanMT Dataset

We construct a parallel corpus based on literary novels, which exhibit intrinsic cultural specificity often absent in general-domain corpora. To ensure cultural representativeness, we select classic novels from different countries that have multiple translations, facilitating the extraction of parallel sentences. The novels and their corresponding language are listed in Figure 2.

During data processing, we first perform sentence-level alignment on each chapter using the Vecalign tool (Thompson and Koehn, 2019) to obtain an initial set of parallel sentence pairs. From these aligned pairs, we employ LLaMA-3.1-70B-Instruct (Dubey et al., 2024) and Qwen2.5-72B-Instruct (Qwen et al., 2025) to identify CSIs, taking the union of their outputs to construct an initial CSI lexicon. To ensure its reliability, the lexicon is filtered by DeepSeek (Liu et al., 2024).

To ensure cultural diversity, we first collect all sentences containing each CSI and limit the candidate pool to at most five sentences per CSI. We then encode these sentences into embeddings using LaBSE (Feng et al., 2022) and apply hierarchical clustering on the embeddings to identify representative sentences, selecting 200 in total from the candidate pool. Finally, human annotators review these selected sentences to remove non-parallel or low-quality translations, resulting in the final dataset. The dataset statistics are summarized in Table 2. Details of the manual filtering process are provided in Appendix B.1.

3.2 Evaluation via Multi-Dimensions

To systematically evaluate culture-aware translation, we propose a multi-dimensional assessment framework grounded in classical translation theories. Our framework measures translation quality along five dimensions: contextual accuracy (Halliday, 1978), cultural adaptation (Venuti, 2008),

Dimensions	Definition	Source	Better Translation	Worse Translation
Contextual Accuracy	Preserves the author-intended meaning of CSIs.	她在弹琵琶。	She is playing the pipa.	She is picking at her banjo.
Cultural Adaptation	Culture adaptation of CSIs.	真是"**说曹操，曹操到**"。	Well, speak of the devil.	It is truly speaking Cao Cao, Cao Cao arrives.
Functional Equivalence	Preservation of the source text's communicative function.	我是不是还得给你颁个奖？	What do you want, a medal?	Do I have to give you an award?
Fidelity	Literal meaning and content preservation.	他走了十里路。	He walked about five kilometers.	He ran ten miles.
Naturalness	Target-language fluency and naturalness.	我非常喜欢这本书。	I like this book very much.	I very like this book.

Figure 3: Overview of the evaluation dimensions adopted in this study, including their definitions, illustrative sources, and representative examples of better and worse translations for each dimension.

Dimension	H–H τ	M–H τ
Contextual Accuracy	0.4536	0.4455
Cultural Adaptation	0.4202	0.3891
Functional Equivalence	0.4416	0.4503
Fidelity	0.4937	0.4625
Naturalness	0.4168	0.4716

Table 3: Evaluation consistency across dimensions. H–H τ denotes inter-annotator agreement measured by Kendall's τ, while M–H τ measures the rank correlation between machine predictions and human judgments.

functional equivalence (Nida, 1964), fidelity (Newmark, 1988), and naturalness (Newmark, 1988); The introduction of these evaluation dimensions is provided in Figure 3.

For each dimension, we employ a 7-point Likert scale. Detailed scoring rubrics for all dimensions are provided in Appendix A. The overall translation score is calculated as the arithmetic mean of the individual dimension scores, allowing for a unified yet fine-grained assessment.

$$S = \frac{1}{|D|} \sum_{d}^{D} s_d \quad (1)$$

where D denotes the total number of evaluation dimensions,s_d represents the score assigned to the d-th dimension.

For scalability and consistency, we employ GPT-5-nano as the automatic evaluator to assign scores for each dimension, with the full evaluation prompts provided in the Appendix C. To validate the scoring procedure, we conduct a controlled human evaluation and report both inter-annotator agreement and the deviation of GPT scores from human judgments.

Specifically, for each translation direction, we select 100 translation pairs and recruit two professional bilingual annotators to independently score across all evaluation dimensions. To quantify human–human agreement, we compute Kendall's τ between the two sets of human scores. For model–human comparison, we retain only those instances for which the absolute difference between the two human scores does not exceed 2. For the retained instances, we use the averaged human score to compute Kendall's τ between GPT's predictions and human judgments. Table 3 reports evaluation consistency across different dimensions.

4 Experimental Setup

To fully evaluate the effectiveness of MT systems on culture-aware translation and the influence of different translation paradigms on model outputs, we both compare multiple MT systems under our evaluation framework (§ 4.1) and investigate how explicit semantic and communicative translation constraints affect their translation behavior (§ 4.2).

4.1 Models & Systems

Our experiments cover following a wide range of models or systems:

- **Open-source LLMs**: We evaluate a set of representative open-source large language models, including the LLaMA3 (Dubey et al., 2024), Qwen2.5 (Qwen et al., 2025), Qwen3 (Yang et al., 2025), and Mixtral (Jiang et al., 2024). Generation for these models uses greedy decoding. In addition, we consider more recent models, namely DeepSeek-V3.2 (DeepSeek-AI, 2025b), DeepSeek-R1 (DeepSeek-AI, 2025a).
- **Proprietary Models**: We evaluate a set of widely used proprietary models, including GPT-4o, GPT-4 (Achiam et al., 2023), Gemini-2.5-flash (Team et al., 2023), and Grok-4.1.
- **Specialized MT Models**: We evaluate dedicated machine translation systems, including NLLB-200 (Costa-Jussà et al., 2022), Seed-X (Cheng et al., 2025) and LLaMAX3 (Lu et al., 2024) which are specifically designed for translation.

- **Production Systems**: We further incorporate widely deployed industrial translation engines, namely Google Translate and Youdao Translate, as real-world production-level baselines.

All open-source models and machine translation systems, except for DeepSeek series, are decoded using greedy decoding.

4.2 Translation Strategies

In addition, we explore the impact of translation paradigm on culture-aware translation quality. Inspired by classical translation theory (Newmark, 1988), we design two translation strategies constraint-based prompts and evaluate model performance under these paradigms.

Semantic Translation Constraint. Emphasizing preserving the meaning and core informational content of the source text, discouraging paraphrasing or the introduction of unstated information.

Communicative Translation Constraint. Emphasizing producing translations that are natural and culturally appropriate for target-language readers, while fulfilling the intended communicative function of the source text.

5 Results and Analysis

5.1 Translate ability across Models

Table 4 presents the overall culture-aware translation performance of all evaluated models across 12 language directions. Detailed results across dimensions are reported in Appendix D.

Model-wise Comparison. Proprietary LLMs and state-of-the-art open-source models consistently outperform traditional production systems across the benchmark. Notably, the performance gap between top-tier open-source and proprietary models has marginally narrowed. Furthermore, the specialized Seed-X exhibits striking efficiency; despite its smaller parameter footprint, it competitively rivals much larger general-purpose models.

Scaling Effects in Open-source Models. Among open-source systems, translation performance exhibits a clear positive correlation with model scale. Taking the Qwen2.5 series as an example, scores increase monotonically from 7B to 14B, 32B, and 72B in nearly all language directions. This trend indicates that larger open-source models consistently achieve stronger culture-aware translation performance, highlighting the benefits of scaling for cross-lingual and culturally tasks.

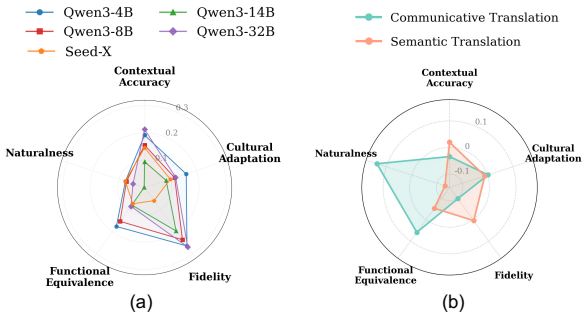

Figure 4: Comparison of improvement gains from reasoning across evaluation dimensions.

Effect of "Test-time Scaling Reasoning". As shown in Table 5, the Qwen3 variants and Seed-X exhibit consistent gains from "think" mode, indicating that incorporating reasoning mechanisms can contribute to enhanced culture-aware translation performance across multiple language pairs.

To identify the specific drivers of this performance boost, we present a dimension-wise breakdown in Figure 4 (a). While reasoning strategies generally yield positive shifts across all metrics, the distribution of gains is non-uniform. Specifically, the *think* models exhibit the most pronounced enhancement in Fidelity, indicating that the reasoning process primarily aids in strictly adhering to the source content's meaning. In contrast, the improvement in Naturalness is the least significant, suggesting that current reasoning mechanisms focus more on semantic transfer than on stylistic fluency.

5.2 Effect of Translation Strategy Constraints

This section analyzes the impact of different translation strategies constraints on translation quality, by comparing model performance under each strategy across multiple evaluation dimensions. Detailed results are reported in Appendix E.

Communicative vs. Default Translation. As shown in Table 6, compared to the default translation, the communicative constraint leads to consistent improvements in Naturalness and Functional Equivalence across most models. This indicates that the communicative constraint primarily focuses on facilitating target readers' comprehension.

Semantic vs. Default Translation. The semantic constraint leads to limited improvement across evaluation dimensions, and in some cases is associated with slight performance declines.

Model	En→Es	En→Ja	En→Zh	Es→En	Es→Zh	Ja→En	Ja→Zh	Ru→Zh	Zh→En	Zh→Es	Zh→Ja	Zh→Ru	Avg
Proprietary LLMs													
GPT-4	4.88	5.16	5.33	5.24	4.91	5.18	5.23	4.79	5.52	5.21	5.34	4.88	5.14
GPT-4o	4.74	4.99	5.21	5.07	4.66	5.03	4.93	4.75	5.35	4.92	5.11	4.75	4.96
Gemini-2.5-Flash-Lite	4.97	5.08	5.11	5.05	4.63	4.96	4.72	4.84	5.40	5.16	5.07	4.97	5.00
Grok-4.1	4.99	5.34	5.22	5.22	5.03	5.04	5.17	5.00	5.43	5.25	5.17	5.30	5.18
Open-source LLMs													
LLaMA-3-8B-Instruct-262k	4.40	3.36	3.89	4.19	3.20	3.41	3.29	3.50	4.33	3.81	3.38	3.32	3.67
LLaMA-3.3-70B-Instruct	4.87	4.83	5.00	5.04	4.57	4.44	4.43	4.61	5.14	4.80	4.81	4.80	4.78
Mixtral-8x7B-Instruct-v0.1	4.47	3.00	3.15	4.81	2.90	3.51	3.15	2.96	4.68	4.19	3.12	3.73	3.64
Qwen2.5-7B-Instruct	4.13	3.52	4.66	4.61	4.04	3.82	4.28	4.31	4.97	3.93	3.65	3.30	4.10
Qwen2.5-14B-Instruct	4.46	4.08	5.00	4.89	4.50	4.32	4.66	4.59	5.26	4.53	4.19	3.66	4.51
Qwen2.5-32B-Instruct	4.71	4.40	5.06	4.99	4.64	4.60	4.82	4.69	5.22	4.74	4.35	4.05	4.69
Qwen2.5-72B-Instruct	4.89	4.84	5.24	5.01	4.72	4.67	4.93	4.85	5.37	5.02	4.74	4.70	4.91
Qwen3-4B	4.01	3.76	4.57	4.43	4.10	3.52	4.24	4.24	4.83	3.82	3.89	3.46	4.07
Qwen3-8B	4.47	4.25	4.91	4.50	4.50	4.02	4.65	4.61	5.03	4.38	4.42	4.05	4.48
Qwen3-14B	4.69	4.65	5.07	4.85	4.62	4.24	4.73	4.83	5.31	4.78	4.73	4.27	4.73
Qwen3-32B	4.66	4.47	5.12	4.92	4.64	4.39	4.77	4.83	5.19	4.71	4.86	4.34	4.74
DeepSeek-R1	5.01	5.08	5.24	5.16	4.65	4.97	4.84	4.70	5.28	5.05	5.07	5.22	5.02
DeepSeek-V3.2	5.05	5.21	5.31	5.10	4.82	4.88	5.07	4.86	5.41	5.10	5.15	5.00	5.08
Specialized MT Models													
Seed-X-PPO-7B	4.78	4.69	5.14	4.86	4.60	4.21	4.28	4.69	5.37	5.07	4.45	4.96	4.76
NLLB-200-3.3B	4.04	3.15	3.07	3.96	2.83	2.67	2.49	3.24	3.20	2.93	2.71	2.86	3.10
LLaMAX3-8B-Alpaca	4.07	3.90	4.12	4.45	3.61	3.70	3.94	3.81	4.46	3.89	3.77	3.69	3.95
Production Systems													
Google Translate	4.45	4.64	4.70	4.54	4.06	4.35	4.02	4.54	5.00	4.61	4.33	4.72	4.50
Youdao Translate	3.59	3.72	4.89	4.11	3.46	3.61	4.67	3.28	5.14	3.35	4.35	3.55	3.98

Table 4: Overall translation performance across language directions. For the Qwen3 series and Seed-X models, only the non-reasoning variants are included.

Model	w/o think	with think	
Qwen3-4B	4.07	4.25	↑0.18
Qwen3-8B	4.48	4.63	↑0.15
Qwen3-14B	4.73	4.82	↑0.09
Qwen3-32B	4.74	4.89	↑0.15
Seed-X-PPO-7B	4.76	4.85	↑0.09

Table 5: A comparative analysis of model performance with and without the "think" mode.

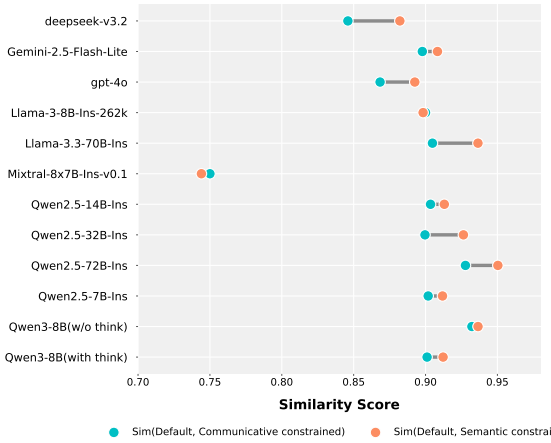

Figure 5: Similarity of default translations to semantic and communicative-constrained translations.

To better understand how different constraints shape translation behavior, we analyze the directional preferences induced by the two constrained translation strategies. As shown in Figure 4 (b), the communicative constraint consistently yields improvements in Naturalness, Functional Equivalence, and Cultural Adaptation across most models,suggesting that the communicative strategy encourages translations that are more fluent, functionally appropriate, and readily comprehensible to target-language readers.

In contrast, the semantic constraint demonstrates more conservative and asymmetric trade-offs across dimensions. Under this strategy, Fidelity and Contextual Accuracy remain consistently more stable than other dimensions, indicating a clear prioritization of semantic faithfulness and contextual precision. These results suggest that the semantic translation strategy is more effective at preserving the original author's intended meaning, favoring accurate transmission of source semantics over target-oriented adaptation.

6 Discussion and Analysis

6.1 Default Translation Preference Across Different Strategies

To characterize default translation behavior of LLMs, we measure cosine similarity between translations under the unconstrained setting and under semantic or communicative constraints.

As shown in Figure 5, default translations consistently exhibit higher similarity to semantic-constrained translations across most models, indicating that in the absence of explicit constraints, models default to a semantic translation mode. This pattern is consistent with Table 6, which shows that communicative constraints induce larger performance fluctuations.

Model	Default Translation					Communicative Translation					Semantic Translation				
	CTX	CAD	FEQ	FID	NAT	CTX	CAD	FEQ	FID	NAT	CTX	CAD	FEQ	FID	NAT
deepseek-v3.2	5.17	5.30	5.07	5.07	4.80	4.95 (-.21)	5.20 (-.10)	4.95 (-.12)	4.72 (-.35)	4.71 (-.09)	5.20 **(+.04)**	5.32 **(+.02)**	5.04 (-.04)	5.14 **(+.07)**	4.58 (-.22)
Gemini-2.5-Flash-Lite	5.04	5.22	4.97	5.08	4.68	5.01 (-.03)	5.24 **(+.02)**	5.05 **(+.08)**	5.00 (-.09)	4.83 **(+.15)**	5.10 **(+.06)**	5.20 (-.02)	4.83 (-.14)	5.07 (-.02)	4.43 (-.24)
gpt-4o	5.03	5.17	4.90	5.10	4.61	5.02 (-.01)	5.19 **(+.02)**	5.07 **(+.17)**	4.94 (-.16)	4.94 **(+.33)**	5.16 **(+.13)**	5.24 **(+.07)**	5.05 **(+.05)**	5.18 **(+.09)**	4.52 (-.09)
Llama-3-8B-Ins-262k	3.61	4.14	3.60	3.51	3.50	3.60 (-.01)	4.09 (-.04)	3.60 (.00)	3.50 (-.02)	3.61 **(+.11)**	3.61 (.00)	4.15 **(+.01)**	3.53 (-.08)	3.47 (-.04)	3.44 (-.07)
Llama-3.3-70B-Ins	4.81	5.03	4.72	4.84	4.50	4.68 (-.13)	4.99 (-.04)	4.80 **(+.08)**	4.65 (-.19)	4.67 **(+.17)**	4.84 **(+.03)**	4.95 (-.07)	4.69 (-.03)	4.81 (-.03)	4.35 (-.15)
Mixtral-8x7B-Ins-v0.1	3.71	4.12	3.50	3.48	3.38	3.79 **(+.08)**	4.21 **(+.09)**	3.47 (-.03)	3.34 (-.14)	3.44 **(+.06)**	3.67 (-.04)	4.08 (-.04)	3.33 (-.17)	3.38 (-.10)	3.19 (-.19)
Qwen2.5-14B-Ins	4.53	4.78	4.49	4.49	4.27	4.53 (.00)	4.86 **(+.07)**	4.70 **(+.21)**	4.52 **(+.03)**	4.49 **(+.23)**	4.56 **(+.02)**	4.85 **(+.07)**	4.53 **(+.04)**	4.58 **(+.10)**	4.15 (-.12)
Qwen2.5-32B-Ins	4.73	5.00	4.64	4.72	4.34	4.63 (-.11)	4.93 (-.08)	4.75 **(+.10)**	4.55 (-.17)	4.58 **(+.23)**	4.74 **(+.01)**	4.93 (-.08)	4.60 (-.04)	4.68 (-.04)	4.24 (-.10)
Qwen2.5-72B-Ins	4.95	5.12	4.89	4.97	4.64	4.91 (-.05)	5.14 **(+.02)**	4.95 **(+.06)**	4.89 (-.07)	4.83 **(+.19)**	4.99 **(+.04)**	5.11 (-.02)	4.87 (-.02)	5.04 **(+.08)**	4.57 (-.07)
Qwen2.5-7B-Ins	4.13	4.49	4.06	4.00	3.84	4.17 **(+.04)**	4.56 **(+.07)**	4.14 **(+.08)**	4.02 **(+.03)**	3.98 **(+.14)**	4.21 **(+.08)**	4.55 **(+.06)**	4.11 **(+.05)**	4.09 **(+.10)**	3.86 **(+.02)**
Qwen3-8B(w/o think)	4.46	4.75	4.50	4.47	4.25	4.41 (-.04)	4.73 (-.02)	4.51 **(+.02)**	4.43 (-.03)	4.29 **(+.04)**	4.42 (-.04)	4.72 (-.03)	4.40 (-.10)	4.35 (-.12)	4.10 (-.15)
Qwen3-8B(with think)	4.61	4.87	4.65	4.70	4.32	4.67 **(+.06)**	4.88 **(+.01)**	4.71 **(+.06)**	4.69 (-.01)	4.41 **(+.10)**	4.51 (-.10)	4.78 (-.09)	4.49 (-.17)	4.67 (-.04)	4.11 (-.21)

Table 6: Effect of translation strategy constraints across three strategies on five evaluation dimensions: Contextual Accuracy (CTX), Cultural Adaptation (CAD), Functional Equivalence (FEQ), Fidelity (FID), and Naturalness (NAT). Default Translation refers to model-generated translations produced without any explicit strategy constraints. Differences relative to the Default Translation are indicated in parentheses.

Cultural Category	Ctx. Acc.	Cul. Adapt.
Geographic & Ecological	**5.03**	**5.30**
Language Symbols	4.39	4.73
Material Culture	4.64	4.85
Organizations & Inst.	4.88	5.05
Social Culture & Customs	4.58	4.83

Table 7: Translation performance across CSI categories. **Ctx. Acc.** and **Cul. Adapt.** denote Contextual Accuracy and Cultural Adaptation scores respectively.

6.2 Performance on Category-wise Culture-Specific Items

To examine how CSI translation performance varies across types, we analyze translations with respect to Contextual Accuracy and Cultural Adaptation, which primarily reflect CSIs translation quality. The categorization framework is adapted from Newmark's taxonomy of CSI (Newmark, 1988). To maintain analytical clarity, the study is restricted to sentences that contain solely one CSI category. For completeness,detailed category definitions and the automatic CSI classification procedure are provided in Appendix F.

As shown in Table 7, Geographic and ecological items achieve the highest scores, whereas Language symbols consistently exhibit the lowest performance,reflecting inherent differences in cultural content: geographic and ecological references typically admit direct and conventionalized correspondences across languages, facilitating accurate and adaptive translation, while language symbols are inherently abstract and non-compositional. Representative case studies illustrating these phenomena are provided in Appendix F.3.

6.3 Cultural Translation Knowledge Analysis

Furthermore, to investigate the relationship between CSIs translation knowledge and their translation quality, we conduct a probing analysis of cultural translation knowledge, focusing on models' ability to identify the appropriate rendering of the CSI. In our experiments, we use GPT-4o and construct single-choice questions for CSI translations based on reference translations, allowing the model to answer these questions to assess whether it has correctly mastered the translations. The experimental details are presented in the Appendix G.1.

As shown in Table 8, models exhibit a consistent gain for which all questions are answered correctly compared to sentences where they are not, highlighting the role of cultural translation knowledge in achieving high-quality CSI translation.

However, possessing knowledge does not inherently guarantee is faithful application. Figure 6 illustrates the distribution of scores specifically within the knowledge-correct subset, which we define as the collection of instances where the model successfully passes all corresponding probing questions for a given CSI. We identify a persistent "knowledge-application gap": a non-negligible proportion of translations still result in low-quality outputs despite the models "knowing" the correct rendering in the probing task. In the Appendix G.2, we provide a case to illustrate this phenomenon.

6.4 The Role of Reference Translation in *LLM-as-a-Judge*

Following (Qian et al., 2024), we investigate the role of reference translations in cultural translation evaluation. Under the setting described in Section 3.2, we perform no-reference evaluation and

Model	Ctx. Acc.		Cul. Adapt.	
	w/ Know.	w/o Know.	w/ Know.	w/o Know.
Llama-3-8B-Ins-262k	4.01	3.23	4.44	3.85
Llama-3.3-70B-Ins	5.02	4.34	5.17	4.64
Mixtral-8x7B-Ins-v0.1	4.21	3.54	4.54	3.99
Qwen2.5-7B-Ins	4.37	3.74	4.64	4.26
Qwen2.5-14B-Ins	4.77	4.13	4.97	4.47
Qwen2.5-32B-Ins	4.89	4.36	5.17	4.59
Qwen2.5-72B-Ins	5.11	4.52	5.26	4.72

Table 8: Performance of translation under Contextual Accuracy (Ctx. Acc.) and Cultural Adaptation (Cul. Adapt.). Scores are shown separately for cases where the model correctly answered all CSI probing questions (*w/ Know*) and where it did not (*w/o Know*).

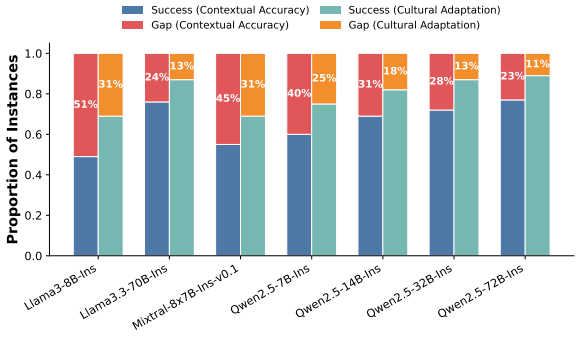

Figure 6: Distribution of CSI translation scores within the knowledge-correct subset. The lower segments indicate scores ≥ 4 in the corresponding dimension.

compute Kendall's τ to measure the rank correlation between automatic scores and human judgments across evaluation dimensions.

As shown in Table 9, incorporating reference translations generally improves the agreement between automatic evaluation metrics and human judgments. Our results show that reference translations play a crucial role in cultural translation evaluation. They provide reference renderings for CSIs and serve as a factual baseline for detecting fine-grained semantic errors, while also helping calibrate whether the translation style aligns with target-language norms. Detailed qualitative analyses of specific cases, presented in Appendix H, further illustrate these effects.

7 Conclusion

In this paper, we presented **CanMT**, a novel-driven benchmark for culture-aware machine translation spanning 12 directions. Under five theoretically grounded evaluation dimensions, we systematically assessed modern LLMs. Beyond establishing that scaling and reasoning improve performance, our systematic evaluation reveals that while communicative strategies significantly enhance target readers' comprehension, semantic strategies are more effective at accurately conveying the author's intent. Notably, we found that LLMs exhibit a default bias toward semantic translation in unconstrained settings. Our analysis across CSIs identifies a difficulty hierarchy, where abstract language symbols remain the most challenging category. Crucially, we discovered a persistent "knowledge-application gap", demonstrating that possessing cultural knowledge does not inherently guarantee its faithful application in translation. Finally, we showed that reference translations are indispensable for calibrating automatic metrics with human judgment. Collectively, **CanMT** provides a rigorous diagnostic framework for the community, highlighting that achieving functional cultural adequacy requires bridging the gap between knowledge possession and its strategic, context-aware deployment.

Dimension	τ **w/ Ref**	τ **w/o Ref**
Contextual Accuracy	0.45	0.42
Cultural Adaptation	0.39	0.36
Fidelity	0.46	0.41
Functional Equivalence	0.45	0.42
Naturalness	0.47	0.45

Table 9: Agreement between LLM scores and human judgments measured by Kendall's τ, with and without reference translations.

Limitations

Despite the systematic approach taken in this study, several limitations remain. First, regarding data source, our benchmark relies primarily on literary fiction. While novels offer high-density cultural content, they may not fully represent the linguistic diversity found in other culturally rich domains, such as social media, historical archives, or oral history records. Additionally, the reliance on public domain or classic texts might introduce a temporal bias, potentially overlooking contemporary cultural neologisms. Second, our analysis focuses on sentence-level translation. Cultural meaning is often constructed discursively across broader contexts (paragraph or document level). Future work should extend this evaluation to discourse-level settings to capture long-range cultural dependencies and consistency.

Ethical Considerations

The **CanMT** dataset is constructed from classic novels. For works that may still be under copy-

right protection in certain jurisdictions, we adhere to the principles of fair use for academic research, extracting only sentence-level parallel segments rather than reproducing full texts. We will release the dataset with strict licenses prohibiting commercial redistribution of the copyrighted segments.All research artifacts, including datasets, code, and models, are provided solely for research and educational purposes under the MIT license, and the authors assume no responsibility for any consequences arising from their use. In this paper, we use Gemini to correct grammatical errors.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Yong Cao, Yova Kementchedjhieva, Ruixiang Cui, Antonia Karamolegkou, Li Zhou, Megan Dare, Lucia Donatelli, and Daniel Hershcovich. 2023. Cultural adaptation of recipes. *Preprint*, arXiv:2310.17353.

Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. 2025. Benchmarking LLMs for translating classical Chinese poetry: Evaluating adequacy, fluency, and elegance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33007–33024, Suzhou, China. Association for Computational Linguistics.

Shanbo Cheng, Yu Bao, Qian Cao, Luyang Huang, Liyan Kang, Zhicheng Liu, Yu Lu, Wenhao Zhu, Jingwen Chen, Zhichao Huang, Tao Li, Yifu Li, Huiying Lin, Sitong Liu, Ningxin Peng, Shuaijie She, Lu Xu, Nuo Xu, Sen Yang, and 7 others. 2025. Seed-x: Building strong multilingual translation llm with 7b parameters. *Preprint*, arXiv:2507.13618.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

DeepSeek-AI. 2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

DeepSeek-AI. 2025b. Deepseek-v3.2: Pushing the frontier of open large language models.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891.

Zhaopeng Feng, Jiayuan Su, Jiamei Zheng, Jiahan Ren, Yan Zhang, Jian Wu, Hongwei Wang, and Zuozhu Liu. 2025. M-MAD: Multidimensional multi-agent debate for advanced machine translation evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7084–7107, Vienna, Austria. Association for Computational Linguistics.

Baban Gain, Dibyanayan Bandyopadhyay, and Asif Ekbal. 2025. Bridging the linguistic divide: A survey on leveraging large language models for machine translation. *CoRR*, abs/2504.01919.

M.A.K. Halliday. 1978. *Language as Social Semiotic: The Social Interpretation of Language and Meaning*. Open University set book. Edward Arnold.

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*, abs/2302.09210.

Tianyi Hu, Maria Maistro, and Daniel Hershcovich. 2024. Bridging cultures in the kitchen: A framework and benchmark for cross-cultural recipe retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1068–1080, Miami, Florida, USA. Association for Computational Linguistics.

Yichong Huang, Baohang Li, Xiaocheng Feng, Wenshuai Huo, Chengpeng Fu, Ting Liu, and Bing Qin. 2024. Aligning translation-specific understanding to general understanding in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5028–5041, Miami, Florida, USA. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. Mixtral of experts. *ArXiv*, abs/2401.04088.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *Preprint*, arXiv:2301.08745.

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

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

Shuang Li, Jiangjie Chen, Siyu Yuan, Xinyi Wu, Hao Yang, Shimin Tao, and Yanghua Xiao. 2023. Translate meanings, not just words: Idiomkb's role in optimizing idiomatic translation with language models. *Preprint*, arXiv:2308.13961.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Yinquan Lu, Wenhao Zhu, Lei Li, Yu Qiao, and Fei Yuan. 2024. LLaMAX: Scaling linguistic horizons of LLM by enhancing translation capabilities beyond 100 languages. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10748–10772, Miami, Florida, USA. Association for Computational Linguistics.

P. Newmark. 1981. *Approaches to Translation*. Language teaching methodology series. Pergamon Press.

Peter Newmark. 1988. *A textbook of translation*, volume 66. Prentice hall New York.

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

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, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Shenbin Qian, Archchana Sindhujan, Minnie Kabra, Diptesh Kanojia, Constantin Orasan, Tharindu Ranasinghe, and Fred Blain. 2024. What do large language models need for machine translation evaluation? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3660–3674.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S Yu. 2024. Multilingual large language model: A survey of resources, taxonomy and frontiers. *arXiv preprint arXiv:2404.04925*.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. Qwen2.5 Technical Report. *arXiv preprint*. ArXiv:2412.15115 [cs].

Pushpdeep Singh, Mayur Patidar, and Lovekesh Vig. 2024. Translating across cultures: LLMs for intralingual cultural adaptation. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 400–418, Miami, FL, USA. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. 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 1342–1348.

L. Venuti. 2008. *The Translator's Invisibility: A History of Translation*. The Translator's Invisibility: A History of Translation. Routledge.

Minghan Wang, Viet-Thanh Pham, Farhad Moghimifar, and Thuy-Trang Vu. 2025. Proverbs run in pairs: Evaluating proverb translation capability of large language model. *Preprint*, arXiv:2501.11953.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

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.

Yangfan Ye, Xiachong Feng, Xiaocheng Feng, Weitao Ma, Libo Qin, Dongliang Xu, Qing Yang, Hongtao Liu, and Bing Qin. 2024a. GlobeSumm: A challenging benchmark towards unifying multi-lingual, cross-lingual and multi-document news summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10803–10821, Miami, Florida, USA. Association for Computational Linguistics.

Yangfan Ye, Xiaocheng Feng, Xiachong Feng, Lei Huang, Weitao Ma, Qichen Hong, Yunfei Lu, Duyu Tang, Dandan Tu, and Bing Qin. 2025a. Langgps: Language separability guided data pre-selection for joint multilingual instruction tuning. *arXiv preprint arXiv:2511.10229*.

Yangfan Ye, Xiaocheng Feng, Xiachong Feng, Libo Qin, Yichong Huang, Lei Huang, Weitao Ma, Qichen Hong, Zhirui Zhang, Yunfei Lu, and 1 others. 2024b. Exploring cross-lingual latent transplantation: Mutual opportunities and open challenges. *arXiv preprint arXiv:2412.12686*.

Yangfan Ye, Xiaocheng Feng, Zekun Yuan, Xiachong Feng, Libo Qin, Lei Huang, Weitao Ma, Yichong Huang, Zhirui Zhang, Yunfei Lu, Xiaohui Yan, Duyu Tang, Dandan Tu, and Bing Qin. 2025b. CC-tuning: A cross-lingual connection mechanism for improving joint multilingual supervised fine-tuning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19036–19051, Vienna, Austria. Association for Computational Linguistics.

Enze Zhang, Jiaying Wang, Mengxi Xiao, Jifei Liu, Ziyan Kuang, Rui Dong, Eric Dong, Sophia Ananiadou, Min Peng, and Qianqian Xie. 2025. Diting: A multi-agent evaluation framework for benchmarking web novel translation. *arXiv preprint arXiv:2510.09116*.

Zhonghe Zhang, Xiaoyu He, Vivek Iyer, and Alexandra Birch. 2024. Cultural adaptation of menus: A fine-grained approach. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1258–1271, Miami, Florida, USA. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. 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.

Appendix

A Scoring Rubrics for Evaluation Dimensions

To ensure consistency and interpretability of the multi-dimensional evaluation, we provide detailed scoring rubrics for each evaluation dimension. All dimensions are rated on a 7-point Likert scale, where higher scores indicate better translation quality.

A.1 Contextual Accuracy

Contextual Accuracy measures whether the translation of the CSI correctly reflects the meaning intended by the source text within its discourse context.

- **1 point (Extremely Poor)**: The translation completely distorts or ignores the source context and culturally specific items, leading to severe misinterpretation of cultural, historical, or narrative elements.
- **2 points (Poor)**: The translation shows major inaccuracies in reflecting the source context, with significant distortions or omissions of culturally specific items.
- **3 points (Relatively Poor)**: The translation partially captures the meaning of culturally specific items but includes notable errors or incomplete incorporation of the background information related.
- **4 points (Average)**: The translation adequately reflects the meaning of culturally specific items on a basic level, with minor inaccuracies in cultural or historical details.
- **5 points (Relatively Good)**: The translation mostly captures the culturally specific item effectively, with good incorporation of background but room for improvement.
- **6 points (Good)**: The translation accurately and thoroughly reflects the culturally specific items, incorporating relevant cultural and historical elements well, although at the minor cost of not adopting the most well-received translation.
- **7 points (Excellent)**: The translation uses the exact commonly accepted translation. Or The translation perfectly embodies the culturally specific items.

A.2 Cultural Adaptation

Cultural Adaptation assesses whether the translation appropriately adapts the CSI for target-language readers, ensuring cultural intelligibility and avoiding cultural conflict or confusion.

- **1 point (Extremely Poor)**: The translation of CSIs causes severe cultural conflicts, with no effective adaptation or retention of core elements.
- **2 points (Poor)**: The translation of CSIs shows poor adaptation, leading to significant

cultural clashes or poor understanding by target readers.

- **3 points (Relatively Poor)**: The translation of CSIs includes some adaptation but still results in notable cultural issues or incomplete retention of elements.
- **4 points (Average)**: The translation of CSIs provides basic cultural adaptation, avoiding major conflicts but with room for better alignment.
- **5 points (Relatively Good)**: The translation of CSIs adapts well to target norms, ensuring good understanding while mostly retaining core cultural elements.
- **6 points (Good)**: The translation of CSIs effectively balances adaptation and retention, minimizing conflicts and enhancing reader comprehension.
- **7 points (Excellent)**: The translation of CSIs masterfully adapts to the target culture without any conflicts, perfectly retaining and integrating core cultural elements. If the translation is widely recognized, it can be awarded with 7 points.

A.3 Functional Equivalence

Functional Equivalence measures whether the translation fulfills the communicative function of the source text, such as informing, persuading, or expressing attitude.

- **1 point (Extremely Poor)**: The translation utterly fails to achieve any of the source's intended functions, resulting in a complete loss of pragmatic effect.
- **2 points (Poor)**: The translation achieves minimal functional equivalence, with major failures in conveying the source's purpose (e.g., criticism or emotion).
- **3 points (Relatively Poor)**: The translation partially realizes the source's functions but with significant shortcomings in evoking the intended response.
- **4 points (Average)**: The translation adequately fulfills the basic functions of the source, though not fully effectively.
- **5 points (Relatively Good)**: The translation mostly achieves the source's functions well, with good conveyance of intended effects like the offering of information or expression emotion.
- **6 points (Good)**: The translation effectively realizes the source's functions, successfully evoking the desired reader's emotional response or informing them more about the story.
- **7 points (Excellent)**: The translation perfectly captures and enhances all intended functions of the source, delivering an optimal pragmatic impact.

A.4 Fidelity

Fidelity evaluates the extent to which the translation preserves the literal meaning and core informational content of the source text.

- **1 point (Extremely Poor)**: The translation severely deviates from the source's literal meaning and structure, with extensive errors, omissions, or complete semantic loss.
- **2 points (Poor)**: The translation shows major deviations from the source, with significant semantic loss due to inaccuracies or unnecessary changes.
- **3 points (Relatively Poor)**: The translation retains some literal elements but includes notable errors or omissions that affect fidelity.
- **4 points (Average)**: The translation basically preserves the source's literal meaning and structure, with only minor deviations or omissions.
- **5 points (Relatively Good)**: The translation mostly maintains high fidelity to the source, with good retention of literal meaning and few minor issues.
- **6 points (Good)**: The translation effectively preserves the source's literal meaning and structure, avoiding semantic loss almost entirely.
- **7 points (Excellent)**: The translation impeccably retains every aspect of the source's literal meaning and structure, with no errors or omissions whatsoever.

A.5 Naturalness

Naturalness assesses the fluency and idiomaticity of the translation in the target language.

- **1 point (Extremely Poor)**: The translation is extremely awkward and unnatural, reading like a forced or incomprehensible transplant.
- **2 points (Poor)**: The translation lacks fluency, with major stiffness or non-native phrasing that hinders readability.
- **3 points (Relatively Poor)**: The translation is somewhat readable but includes noticeable unnatural elements or awkward flow.
- **4 points (Average)**: The translation achieves basic naturalness, reading adequately but without full native fluency.
- **5 points (Relatively Good)**: The translation is mostly natural and fluent, with good readability and minor improvements possible.
- **6 points (Good)**: The translation flows naturally, resembling native expression with high readability.
- **7 points (Excellent)**: The translation is perfectly natural and seamless, indistinguishable from original target-language writing in fluency and rhythm.

B Instruction for human annotators

During our annotation process, all human annotators were compensated appropriately.

B.1 human data filtering

The manual data filtering procedure is illustrated in Figure 7.

B.2 human Evaluation

The manual evaluation procedure is illustrated in Figure 8–12.

C Prompts for LLM Evaluation

For reproducibility, we provide the exact prompts used in our evaluation for each dimension, as shown in Figure 13–17.

==========

The goal of filtering is to ensure that each pair of sentences (A, B) constitutes a genuine parallel corpus—i.e., both express the same core meaning and carry equivalent information.

Examples:

1. Completely Unmatched Content:

Sentence A and B describe entirely different topics or situations.

Example:

A: “The meeting was canceled due to rain.”

B: “天气很好，适合散步。”

(Completely inconsistent in meaning — should be deleted.)

2. Obvious Non-Translation Concatenation:

One side contains multiple sentences while the other includes only part of the content.

Example:

A: “She smiled. Then she walked away.”

B: “她笑了。”

(The latter half is missing — not a complete parallel pair.)

2.2 Mismatched Information Volume (To Be Modified)

1. Contextual Misinterpretation:

Translation omits or generalizes specific details (e.g., person names, times, places) and should restore the original meaning.

Example:

A: “Finn loves cats.”

B: “他喜欢猫。”

→ Modify to: “费恩喜欢猫。”

(Add explicit subject to restore contextual information.)

2. Partial Alignment: The source covers only part of the target or multiple sentences are incorrectly merged.

Example:

A: “Finn loves cats.”

B: “费恩喜欢猫。我喜欢狗。”

→ Modify to: “费恩喜欢猫。”

3. Redundant Information: If the translation adds subjective comments or extra information not in the source, the redundant part should be removed.

Example:

A: “It started to rain.”

B: “天开始下雨，真是糟糕。”

→ Modify to: “天开始下雨。”

==========

Figure 7: Instruction for human data filtering.

D Dimension-level Evaluation Results

To provide a more complete view of model behavior under our evaluation framework, this appendix reports the dimension-level scores for all evaluated models across all language directions. While the main paper focuses on overall scores for clarity and comparability, we include detailed results here to demonstrate the full evaluation coverage and to support further inspection. The complete results are presented in Tables 10–13.

E Language-level Results under Constraints

This appendix presents language-level evaluation results of different models under both semantic and communicative translation constraints across multiple language directions. While the main paper reports only averaged scores over language pairs for conciseness, we provide detailed per-language results here to offer a finer-grained view of model performance under different translation strategies. Results under semantic translation constraints are summarized in Tables 14–17, while results under communicative translation constraints are reported in Tables 18–21.

Model	Contextual Accuracy			Cultural Adaptation			Functional Equivalence			Fidelity			Naturalness		
	En → Es	En → Ja	En → Zh	En → Es	En → Ja	En → Zh	En → Es	En → Ja	En → Zh	En → Es	En → Ja	En → Zh	En → Es	En → Ja	En → Zh
Proprietary LLMs															
GPT-4	4.93	5.32	5.50	5.28	5.49	5.65	4.68	5.13	5.18	4.91	5.20	5.41	4.59	4.66	4.88
GPT-4o	4.69	5.31	5.65	5.11	5.32	5.57	4.57	4.84	5.05	5.00	5.14	5.15	4.35	4.35	4.64
Gemini-2.5-Flash-Lite	5.00	5.26	5.39	5.31	5.34	5.50	4.75	4.91	4.97	5.22	5.25	4.97	4.59	4.64	4.70
Grok-4.1	5.06	5.62	5.47	5.28	5.76	5.44	4.77	5.24	5.16	5.26	5.46	5.36	4.57	4.60	4.70
Open-source LLMs															
LLaMA-3-8B-Instruct-262k	4.65	3.42	4.09	5.01	3.98	4.48	4.15	3.21	3.73	4.25	3.24	3.57	3.95	2.95	3.59
LLaMA-3.3-70B-Instruct	4.89	5.36	5.08	5.30	5.17	5.30	4.61	4.53	4.93	5.11	4.92	5.05	4.44	4.18	4.63
Mixtral-8x7B-Instruct-v0.1	4.47	3.26	3.65	4.95	3.70	3.86	4.25	2.69	2.74	4.45	2.81	2.76	4.23	2.53	2.73
Qwen2.5-7B-Instruct	4.33	3.79	4.78	4.68	4.15	5.23	3.92	3.22	4.44	4.06	3.37	4.54	3.66	3.07	4.30
Qwen2.5-14B-Instruct	4.64	4.28	5.23	5.01	4.50	5.34	4.23	3.95	4.95	4.29	3.98	4.90	4.12	3.67	4.59
Qwen2.5-32B-Instruct	4.77	4.57	5.30	5.18	4.93	5.42	4.52	4.10	4.95	4.78	4.50	5.11	4.28	3.88	4.51
Qwen2.5-72B-Instruct	5.08	5.06	5.45	5.12	5.37	5.54	4.67	4.65	5.03	5.01	4.83	5.30	4.57	4.27	4.87
Qwen3-4B (w/o think)	4.08	3.99	4.78	4.63	4.21	4.96	3.75	3.47	4.43	3.89	3.70	4.43	3.70	3.43	4.26
Qwen3-4B (with think)	4.11	4.46	4.94	4.68	4.52	5.14	3.79	3.89	4.54	3.92	4.00	4.69	3.76	3.44	4.32
Qwen3-8B (w/o think)	4.78	4.54	5.05	5.05	4.73	5.27	4.08	4.17	4.80	4.38	4.11	4.89	4.08	3.69	4.53
Qwen3-8B (with think)	4.78	4.69	5.14	5.03	4.71	5.38	4.47	4.18	5.01	4.66	4.21	5.04	4.34	3.72	4.67
Qwen3-14B (w/o think)	5.02	4.90	5.14	5.15	5.05	5.38	4.46	4.56	4.95	4.60	4.62	5.20	4.23	4.12	4.68
Qwen3-14B (with think)	4.89	4.81	5.06	5.18	4.83	5.29	4.59	4.49	5.06	4.79	4.64	5.04	4.34	3.89	4.64
Qwen3-32B (w/o think)	4.71	4.72	5.47	5.10	4.66	5.52	4.48	4.53	4.95	4.80	4.44	4.93	4.21	4.01	4.71
Qwen3-32B (with think)	4.99	4.91	5.31	5.21	5.01	5.41	4.60	4.67	5.01	4.92	4.82	5.23	4.39	4.12	4.57
DeepSeek-R1	5.09	5.30	5.49	5.33	5.31	5.57	4.87	5.06	5.20	5.08	5.09	5.12	4.66	4.62	4.85
DeepSeek-V3.2	5.23	5.62	5.55	5.29	5.55	5.82	4.91	5.04	5.15	5.11	5.09	5.14	4.74	4.75	4.87
Specialized MT Models															
Seed-X-PPO-7B (w/o CoT)	4.79	5.01	5.15	5.00	4.98	5.47	4.49	4.56	5.05	5.09	4.74	5.03	4.53	4.15	4.98
Seed-X-PPO-7B (with CoT)	4.86	5.04	5.61	5.35	5.15	5.77	4.72	4.51	5.42	4.92	4.80	5.31	4.52	4.30	5.24
NLLB-200-3.3B	4.10	3.48	3.34	4.62	4.09	3.85	3.65	2.84	2.78	4.11	2.75	2.82	3.72	2.61	2.55
LLaMAX3-8B-Alpaca	4.08	4.17	4.28	4.81	4.45	4.57	3.62	3.66	3.97	4.09	3.70	3.99	3.76	3.50	3.77
Production Systems															
Google Translate	4.76	5.08	4.99	5.09	5.22	5.11	4.10	4.20	4.54	4.32	4.62	4.61	3.98	4.09	4.24
Youdao Translate	3.83	3.93	5.10	4.22	4.38	5.26	3.38	3.48	4.72	3.40	3.54	4.84	3.14	3.27	4.52

Table 10: Fine-grained evaluation of translation quality across multiple dimensions for En→Es, En→Ja, and En→Zh directions.

```
================================================================================
Assess whether the culturally specific item in the sentence precisely reflects the source language's unspoken context, including paragraph, cultural, historical, or narrative background. During the evaluation of contextual accuracy, you need to assume the identity of the original author, infer the unprovided context, and consider whether the translation can be accurately applied to the context, and whether it accurately incorporates the relevant historical and cultural background.

- **1 point (Extremely Poor)**: The translation completely distorts or ignores the source context and culturally specific items, leading to severe misinterpretation of cultural, historical, or narrative elements.
- **2 points (Poor)**: The translation shows major inaccuracies in reflecting the source context, with significant distortions or omissions of culturally specific items.
- **3 points (Relatively Poor)**: The translation partially captures the meaning of culturally specific items but includes notable errors or incomplete incorporation of the background information related.
- **4 points (Average)**: The translation adequately reflects the meaning of culturally specific items on a basic level, with minor inaccuracies in cultural or historical details.
- **5 points (Relatively Good)**: The translation mostly captures the culturally specific item effectively, with good incorporation of background but room for improvement.
- **6 points (Good)**: The translation accurately and thoroughly reflects the culturally specific items, incorporating relevant cultural and historical elements well, although at the minor cost of not adopting the most well-received translation.
- **7 points (Excellent)**: The translation perfectly embodies the culturally specific items or uses the exact commonly accepted translation.

**Example**His estate, Longbourn, is entailed to the male line.
--from Pride and Prejudice, the translation should ensure that inheritance to the male line of the Bennet family gets properly translated to ensure contextual accuracy.
================================================================================
```

Figure 8: Instruction for human eval Contextual Accuracy.

F CSI Categorization Details

F.1 CSI Taxonomy

We categorize culture-specific items (CSIs) based on Newmark's taxonomy (Newmark, 1988), which classifies CSIs into five main types: geographic and ecological items, material culture items, social culture and customs, organizations and institutions, and language symbols. Building on this framework, we adapt and refine the definitions to better suit the context of our translation evaluation. Figure 18 provides the resulting definitions and representative examples for each category.

F.2 Automatic CSI Classification

In this work, culture-specific items are automatically identified and classified using GPT-4o. Each CSI instance is assigned to one of the five CSI categories. To illustrate this process, Figure 19 shows the prompt used for the automatic classification of CSIs, specifying the five categories and the expected JSON output format.

F.3 Category-wise CSI Examples

As discussed in Section 6.2 and shown in Table 6, Geographic and Ecological items tend to achieve the highest translation scores, whereas Language Symbols consistently exhibit the lowest performance. To illustrate these patterns qualitatively, we present representative examples of CSI translations across different categories. Examples of Geography and Ecology CSIs are shown in Figures 20, while examples of Language Symbols CSIs are illustrated in Figures 21.

G Cultural Knowledge Probing

G.1 Generation of Questions

To probe models' cultural translation knowledge, we employ GPT-4o to automatically generate single-choice questions for each CSI. For each item, the model is asked to select the most appropriate

Model	Contextual Accuracy			Cultural Adaptation			Functional Equivalence			Fidelity			Naturalness		
	Es → En	Es → Zh	Ja → En	Es → En	Es → Zh	Ja → En	Es → En	Es → Zh	Ja → En	Es → En	Es → Zh	Ja → En	Es → En	Es → Zh	Ja → En
Proprietary LLMs															
GPT-4	5.31	4.92	5.34	5.44	4.89	5.45	5.26	5.11	5.05	5.27	4.89	5.29	4.95	4.75	4.78
GPT-4o	5.16	4.72	5.27	5.36	4.93	5.26	5.01	4.67	4.76	5.16	4.71	5.13	4.66	4.27	4.74
Gemini-2.5-Flash-Lite	5.28	4.48	5.21	5.34	4.88	5.42	5.02	4.77	4.67	4.99	4.51	4.92	4.64	4.50	4.55
Grok-4.1	5.46	4.97	5.35	5.31	5.04	5.53	5.19	5.28	4.79	5.34	5.20	4.91	4.81	4.66	4.59
Open-source LLMs															
LLaMA-3-8B-Instruct-262k	4.25	2.98	3.19	4.71	3.68	3.80	4.06	3.23	3.27	3.87	2.98	3.23	4.08	3.10	3.53
LLaMA-3.3-70B-Instruct	5.12	4.47	4.39	5.36	4.72	4.84	4.92	4.74	4.28	5.07	4.54	4.39	4.72	4.40	4.29
Mixtral-8x7B-Instruct-v0.1	5.01	3.01	3.30	5.18	3.69	4.03	4.61	2.68	3.38	4.83	2.55	3.18	4.40	2.59	3.64
Qwen2.5-7B-Instruct	4.67	3.86	3.78	4.93	4.37	4.11	4.66	4.07	3.84	4.54	3.98	3.59	4.25	3.91	3.78
Qwen2.5-14B-Instruct	4.90	4.28	4.24	5.17	4.62	4.61	4.91	4.68	4.15	4.88	4.60	4.25	4.61	4.32	4.35
Qwen2.5-32B-Instruct	5.02	4.49	4.73	5.31	4.81	4.87	5.05	4.86	4.42	4.99	4.65	4.46	4.57	4.41	4.52
Qwen2.5-72B-Instruct	4.93	4.58	4.84	5.22	4.80	5.07	5.09	4.91	4.47	4.95	4.70	4.49	4.84	4.60	4.49
Qwen3-4B (w/o think)	4.54	4.02	3.39	4.90	4.27	3.93	4.34	4.31	3.51	4.30	3.98	3.35	4.10	3.92	3.44
Qwen3-4B (with think)	4.60	4.17	3.45	4.81	4.63	3.91	4.30	4.28	3.47	4.28	4.19	3.45	4.03	3.98	3.47
Qwen3-8B (w/o think)	4.72	4.26	3.80	4.77	4.70	4.23	4.49	4.70	4.02	4.42	4.37	3.90	4.13	4.45	4.15
Qwen3-8B (with think)	4.63	4.27	3.97	4.86	4.66	4.46	4.59	4.75	4.18	4.72	4.64	4.12	4.28	4.25	4.13
Qwen3-14B (w/o think)	5.04	4.49	3.94	5.13	4.87	4.54	4.80	4.74	4.28	4.81	4.54	4.17	4.48	4.43	4.28
Qwen3-14B (with think)	4.81	4.75	4.41	5.20	4.78	4.72	4.55	5.02	4.26	4.84	4.81	4.43	4.44	4.51	4.12
Qwen3-32B (w/o think)	5.05	4.46	4.23	5.14	4.87	4.78	4.92	4.80	4.39	4.90	4.54	4.25	4.60	4.54	4.29
Qwen3-32B (with think)	4.99	4.63	4.37	5.31	4.87	4.79	5.04	4.88	4.23	4.98	4.60	4.44	4.69	4.55	4.35
DeepSeek-R1	5.27	4.34	5.18	5.49	4.93	5.28	5.13	4.78	4.66	5.11	4.52	5.07	4.80	4.67	4.67
DeepSeek-V3.2	5.24	4.64	5.19	5.50	4.90	5.34	5.01	5.12	4.51	5.05	4.69	4.81	4.72	4.73	4.57
Specialized MT Models															
Seed-X-PPO-7B (w/o CoT)	4.81	4.33	3.97	5.13	4.81	4.43	5.00	4.74	4.22	4.72	4.36	4.05	4.63	4.78	4.41
Seed-X-PPO-7B (with CoT)	5.14	4.59	4.17	5.28	4.98	4.50	4.95	5.11	4.22	4.75	4.55	4.07	4.64	5.01	4.40
NLLB-200-3.3B	4.07	2.51	2.34	4.46	3.38	2.93	3.82	2.85	2.61	3.80	2.80	2.49	3.64	2.63	2.98
LLaMAX3-8B-Alpaca	4.38	3.51	3.46	4.72	3.96	4.03	4.57	3.72	3.65	4.41	3.39	3.48	4.19	3.46	3.89
Production Systems															
Google Translate	4.90	3.99	4.46	5.19	4.34	4.85	4.34	4.14	4.09	4.37	4.13	4.24	3.90	3.70	4.14
Youdao Translate	4.30	3.55	3.81	4.59	3.91	4.13	3.94	3.36	3.40	4.07	3.31	3.38	3.64	3.15	3.34

Table 11: Fine-grained evaluation of translation quality across multiple dimensions for Es→En, Es→Zh, and Ja→En directions.

translation of the CSI given its context from four candidate options. During question construction, a reference translation is used internally to ensure the correctness of the target option. The detailed prompt template is presented in Figure 22.

G.2 Knowledge-Application Gap Analysis

As shown in Figure 23, although the model selected the correct option for the CSI, it still failed to produce the correct translation in the final output.

This example shows that even with correct knowledge of a CSI, the model may fail to produce a contextually accurate or culturally faithful translation, highlighting the gap between knowing and applying cultural translation knowledge.

H Impact of Reference Translations

In this appendix, we present detailed case studies to qualitatively analyze the impact of reference translations across different evaluation dimensions. For each dimension, we illustrate a representative case (See Figure 24–28) by comparing the evaluator's reasoning and scores in reference-free and reference-based settings. These examples demonstrate how reference translations support the evaluation of cultural-specific items, facilitate the detection of fine-grained semantic errors, and help calibrate whether the translation style aligns with target-language norms.

==========

Assess whether the translation achieves adaptation to the target language culture through adjustments to the culturally specific item, thereby ensuring better understanding by target language readers while avoiding potential cultural conflicts. During the evaluation of cultural adaptation, you need to assume the identity of a target language reader, and assess whether the translation in the target language conforms to the target culture's norms and expectations, trigger the resonance from the readers, while also retaining core cultural elements effectively.

- **1 point (Extremely Poor)**: The translation causes severe cultural conflicts or misunderstandings, with no effective adaptation or retention of core elements.
- **2 points (Poor)**: The translation shows poor adaptation, leading to significant cultural clashes or poor understanding by target readers.
- **3 points (Relatively Poor)**: The translation includes some adaptation but still results in notable cultural issues or incomplete retention of elements.
- **4 points (Average)**: The translation provides basic cultural adaptation, avoiding major conflicts but with room for better alignment.
- **5 points (Relatively Good)**: The translation adapts well to target norms, ensuring good understanding while mostly retaining core cultural elements.
- **6 points (Good)**: The translation effectively balances adaptation and retention, minimizing conflicts and enhancing reader comprehension.
- **7 points (Excellent)**: The translation masterfully adapts to the target culture without any conflicts, perfectly retaining and integrating core cultural elements.

Example1 "I am not religious," he said. "But I will say ten Our Fathers and ten Hail Marys that I should catch this fish, and I promise to make a pilgrimage to the Virgin de Cobre if I catch him. That is a promise."
--from The Old Man and the Sea, the translation should sensitively handle the Catholic religious references (e.g., prayers like Our Fathers and Hail Marys, and the pilgrimage to the Virgin de Cobre) that could be unfamiliar or potentially offensive in non-Christian or secular cultures, providing subtle explanations or cultural equivalents while retaining the character's sincere vow to ensure cultural adaptation.
Example2 She did not understand the beauty he found in her, through touch upon her living secret body, almost the ecstasy of beauty. For passion alone is awake to it.
--from Lady Chatterley's Lover, The conflict in this description lies in the direct presentation of the female nude and "no sense of shame." In some cultures, this depiction is seen as a challenge to the patriarchal gaze and a celebration of bodily autonomy; while in other cultural contexts, it may be interpreted as immoral or shameless. If the translator uses euphemistic and vague terms, it completely distorts the core attitude of "no flicker of shame" in the original text. However, if translated literally, it may face publishing censorship or moral criticism from certain reader groups.

==========

Figure 9: Instruction for human eval Cultural Adaptation.

Model	Contextual Accuracy			Cultural Adaptation			Functional Equivalence			Fidelity			Naturalness		
	Ja → Zh	Ru → Zh	Zh → En	Ja → Zh	Ru → Zh	Zh → En	Ja → Zh	Ru → Zh	Zh → En	Ja → Zh	Ru → Zh	Zh → En	Ja → Zh	Ru → Zh	Zh → En
Proprietary LLMs															
GPT-4	5.22	4.92	5.78	5.42	5.10	5.65	5.16	4.70	5.42	5.47	4.83	5.59	4.87	4.38	5.14
GPT-4o	4.96	4.66	5.38	5.09	5.00	5.57	4.91	4.77	5.25	5.11	4.81	5.52	4.58	4.50	5.02
Gemini-2.5-Flash-Lite	4.58	4.75	5.54	4.93	4.99	5.54	4.71	4.98	5.37	4.93	4.87	5.61	4.44	4.62	4.91
Grok-4.1	5.40	5.00	5.82	5.43	5.16	5.52	5.01	5.10	5.34	5.34	5.06	5.61	4.68	4.66	4.87
Open-source LLMs															
LLaMA-3-8B-Instruct-262k	3.22	3.42	4.30	3.98	3.89	4.54	3.12	3.48	4.33	3.16	3.27	4.18	2.96	3.44	4.33
LLaMA-3.3-70B-Instruct	4.32	4.51	5.26	4.77	4.75	5.23	4.31	4.64	5.08	4.48	4.54	5.30	4.27	4.61	4.86
Mixtral-8x7B-Instruct-v0.1	3.12	3.07	4.63	3.53	3.59	4.77	3.07	2.75	4.75	2.96	2.74	4.70	3.08	2.66	4.53
Qwen2.5-7B-Instruct	4.25	4.30	5.19	4.55	4.69	5.19	4.29	4.29	4.94	4.26	4.19	4.97	4.04	4.06	4.58
Qwen2.5-14B-Instruct	4.42	4.58	5.47	4.67	4.86	5.45	4.79	4.42	5.23	4.80	4.54	5.26	4.60	4.54	4.90
Qwen2.5-32B-Instruct	4.63	4.66	5.31	5.08	4.94	5.43	4.92	4.70	5.14	4.94	4.61	5.31	4.53	4.56	4.92
Qwen2.5-72B-Instruct	4.95	4.73	5.49	5.07	5.04	5.58	4.92	4.90	5.34	5.03	4.95	5.34	4.66	4.65	5.10
Qwen3-4B (w/o think)	4.18	4.04	4.87	4.78	4.49	4.95	4.22	4.30	4.81	4.04	4.06	4.88	3.97	4.30	4.62
Qwen3-4B (with think)	4.35	4.33	5.02	4.83	4.42	5.21	4.29	4.42	5.02	4.24	4.41	4.99	4.04	4.17	4.72
Qwen3-8B (w/o think)	4.42	4.48	5.10	5.09	4.82	5.22	4.58	4.68	5.13	4.75	4.51	5.10	4.42	4.58	4.58
Qwen3-8B (with think)	4.78	4.49	5.22	5.16	4.91	5.25	4.55	4.60	5.20	4.68	4.64	5.34	4.36	4.41	4.86
Qwen3-14B (w/o think)	4.74	4.84	5.51	4.92	4.99	5.41	4.67	4.89	5.34	4.73	4.78	5.31	4.58	4.63	4.96
Qwen3-14B (with think)	4.78	4.74	5.50	5.14	5.07	5.45	4.75	4.85	5.26	4.95	5.02	5.40	4.52	4.62	4.91
Qwen3-32B (w/o think)	4.62	4.68	5.30	5.01	5.16	5.32	4.88	4.89	5.24	4.72	4.74	5.23	4.61	4.66	4.88
Qwen3-32B (with think)	4.73	4.93	5.53	4.92	5.11	5.45	4.78	4.88	5.31	4.98	5.02	5.58	4.51	4.63	4.89
DeepSeek-R1	4.78	4.54	5.45	5.13	4.82	5.52	4.99	4.87	5.23	4.77	4.57	5.30	4.54	4.73	4.90
DeepSeek-V3.2	5.08	4.86	5.49	5.21	4.99	5.50	5.11	4.93	5.45	5.08	4.76	5.42	4.85	4.75	5.20
Specialized MT Models															
Seed-X-PPO-7B (w/o CoT)	3.86	4.44	5.59	4.54	4.90	5.49	4.42	4.72	5.25	4.03	4.44	5.42	4.56	4.93	5.09
Seed-X-PPO-7B (with CoT)	4.17	4.53	5.79	4.57	4.82	5.66	4.29	4.82	5.34	4.08	4.46	5.39	4.75	4.91	5.09
NLLB-200-3.3B	2.33	3.06	3.06	3.04	3.55	3.57	2.32	3.32	3.12	2.38	3.31	3.04	2.40	2.98	3.19
LLaMAX3-8B-Alpaca	3.75	3.62	4.38	4.42	4.06	4.67	3.92	3.90	4.48	3.84	3.66	4.39	3.77	3.82	4.38
Production Systems															
Google Translate	3.86	4.32	5.23	4.46	4.58	5.06	4.09	4.70	4.94	3.78	4.71	5.19	3.90	4.38	4.55
Youdao Translate	4.83	3.06	5.30	4.95	3.65	5.31	4.61	3.36	4.98	4.69	3.23	5.34	4.28	3.12	4.77

Table 12: Fine-grained evaluation of translation quality across multiple dimensions for Ja→Zh, Ru→Zh, and Zh→En directions.

```
================================================================================
Examine whether the translation achieves the function of the source language sentence, such as providing more information to the reader, expressing criticism, persuading, or evoking certain emotions in the reader. During the evaluation of functional equivalence, you also need to assume the identity of the translator, infer what function the original author intended to achieve in the source language text, and then assess whether the translation realizes such a function.

- **1 point (Extremely Poor)**: The translation utterly fails to achieve any of the source's intended functions, resulting in a complete loss of pragmatic effect.
- **2 points (Poor)**: The translation achieves minimal functional equivalence, with major failures in conveying the source's purpose (e.g., criticism or emotion).
- **3 points (Relatively Poor)**: The translation partially realizes the source's functions but with significant shortcomings in evoking the intended response.
- **4 points (Average)**: The translation adequately fulfills the basic functions of the source, though not fully effectively.
- **5 points (Relatively Good)**: The translation mostly achieves the source's functions well, with good conveyance of intended effects like persuasion or emotion.
- **6 points (Good)**: The translation effectively realizes the source's functions, successfully evoking the desired reader response.
- **7 points (Excellent)**: The translation perfectly captures and enhances all intended functions of the source, delivering an optimal pragmatic impact.

**Example**I could not help it: the restlessness was in my nature; it agitated me to pain sometimes.
--from Jane Eyre, the translation should preserve the raw emotional turmoil and self-revelation intended by the author to evoke empathy and inner conflict in the reader to ensure functional equivalence.
================================================================================
```

Figure 10: Instruction for human eval Functional Equivalence.

```
================================================================================
Assess whether the translation maximally retains the literal meaning and source language structure of the culturally specific item, avoiding semantic loss caused by paraphrasing or rewriting. Be faithful to the precise expression and literal meaning of the source language text. During the evaluation of fidelity, you need to assume the identity of the translator, and assess whether the translation itself accurately reflects the original meaning, and whether there are errors or omissions.

- **1 point (Extremely Poor)**: The translation severely deviates from the source's literal meaning and structure, with extensive errors, omissions, or complete semantic loss.
- **2 points (Poor)**: The translation shows major deviations from the source, with significant semantic loss due to inaccuracies or unnecessary changes.
- **3 points (Relatively Poor)**: The translation retains some literal elements but includes notable errors or omissions that affect fidelity.
- **4 points (Average)**: The translation basically preserves the source's literal meaning and structure, with only minor deviations or omissions.
- **5 points (Relatively Good)**: The translation mostly maintains high fidelity to the source, with good retention of literal meaning and few minor issues.
- **6 points (Good)**: The translation effectively preserves the source's literal meaning and structure, avoiding semantic loss almost entirely.
- **7 points (Excellent)**: The translation impeccably retains every aspect of the source's literal meaning and structure, with no errors or omissions whatsoever.
================================================================================
```

Figure 11: Instruction for human eval Fidelity.

Model	Contextual Accuracy			Cultural Adaptation			Functional Equivalence			Fidelity			Naturalness		
	Zh → Es	Zh → Ja	Zh → Ru	Zh → Es	Zh → Ja	Zh → Ru	Zh → Es	Zh → Ja	Zh → Ru	Zh → Es	Zh → Ja	Zh → Ru	Zh → Es	Zh → Ja	Zh → Ru
Proprietary LLMs															
GPT-4	5.29	5.31	4.99	5.39	5.35	4.90	5.15	5.36	5.16	5.29	5.59	5.01	4.90	5.09	4.36
GPT-4o	4.84	5.11	4.61	4.98	5.06	4.80	4.93	5.10	4.88	5.11	5.42	4.87	4.73	4.88	4.58
Gemini-2.5-Flash-Lite	5.26	5.02	4.72	5.29	5.03	5.04	5.11	5.23	5.11	5.27	5.26	5.16	4.86	4.79	4.84
Grok-4.1	5.42	5.33	5.27	5.36	5.15	5.31	5.27	5.18	5.49	5.40	5.50	5.49	4.80	4.69	4.93
Open-source LLMs															
LLaMA-3-8B-Instruct-262k	3.46	3.28	3.04	4.03	3.91	3.63	3.90	3.31	3.41	3.71	3.39	3.32	3.94	2.99	3.18
LLaMA-3.3-70B-Instruct	4.75	4.90	4.68	4.89	4.97	4.99	4.79	4.91	4.87	4.93	4.90	4.84	4.63	4.35	4.60
Mixtral-8x7B-Instruct-v0.1	4.02	3.28	3.71	4.43	3.67	4.05	4.34	3.02	3.76	4.10	2.96	3.72	4.08	2.66	3.41
Qwen2.5-7B-Instruct	3.85	3.59	3.15	4.17	3.99	3.77	4.01	3.67	3.35	3.79	3.53	3.14	3.83	3.45	3.09
Qwen2.5-14B-Instruct	4.51	4.42	3.41	4.64	4.51	4.02	4.64	4.11	3.80	4.49	4.21	3.61	4.38	3.68	3.44
Qwen2.5-32B-Instruct	4.85	4.32	4.13	4.99	4.69	4.40	4.68	4.42	3.97	4.86	4.42	4.03	4.32	3.89	3.73
Qwen2.5-72B-Instruct	5.02	4.70	4.60	5.08	4.79	4.79	5.05	4.91	4.74	5.15	4.92	4.94	4.82	4.39	4.44
Qwen3-4B (w/o think)	3.63	3.82	3.27	3.93	4.17	3.65	4.01	3.87	3.59	3.76	4.03	3.40	3.77	3.54	3.38
Qwen3-4B (with think)	3.99	4.01	3.49	4.36	4.34	3.97	4.36	4.20	4.15	4.36	4.39	4.07	4.06	3.68	3.64
Qwen3-8B (w/o think)	4.29	4.29	3.72	4.62	4.56	3.95	4.47	4.47	4.39	4.32	4.67	4.18	4.21	4.12	4.01
Qwen3-8B (with think)	4.66	4.62	4.06	4.90	4.61	4.48	4.89	4.86	4.53	5.05	4.88	4.45	4.47	4.11	4.21
Qwen3-14B (w/o think)	4.67	4.65	4.04	4.77	4.77	4.43	4.89	4.82	4.47	4.90	5.01	4.20	4.66	4.39	4.18
Qwen3-14B (with think)	5.02	4.80	4.56	5.09	4.85	4.82	4.92	5.17	4.91	5.14	5.34	4.84	4.64	4.45	4.57
Qwen3-32B (w/o think)	4.65	4.83	4.06	4.89	4.92	4.51	4.64	5.01	4.60	4.77	5.10	4.26	4.61	4.46	4.24
Qwen3-32B (with think)	5.28	5.00	4.68	5.19	5.06	5.00	4.85	5.15	4.95	5.10	5.23	5.01	4.73	4.59	4.35
DeepSeek-R1	5.08	5.09	5.25	5.14	5.27	5.42	5.11	5.16	5.28	5.09	5.20	5.28	4.84	4.63	4.86
DeepSeek-V3.2	5.03	5.15	4.92	5.16	5.20	5.18	5.21	5.28	5.15	5.24	5.33	5.05	4.88	4.82	4.69
Specialized MT Models															
Seed-X-PPO-7B (w/o CoT)	4.88	4.35	4.91	5.12	4.57	5.06	5.10	4.55	5.16	5.10	4.53	4.86	5.13	4.24	4.82
Seed-X-PPO-7B (with CoT)	4.85	4.18	4.89	5.01	4.50	5.11	5.17	4.55	5.04	5.19	4.65	4.93	5.05	4.45	4.77
NLLB-200-3.3B	2.62	2.33	2.52	3.24	3.22	3.17	3.00	2.73	2.92	2.75	2.67	2.69	3.04	2.61	3.02
LLaMAX3-8B-Alpaca	3.49	3.60	3.31	4.03	4.01	3.71	4.09	3.87	4.08	3.73	3.83	3.67	4.12	3.55	3.68
Production Systems															
Google Translate	4.71	4.31	4.46	5.05	4.55	4.95	4.38	4.37	4.86	4.63	4.28	4.93	4.28	4.12	4.41
Youdao Translate	3.15	4.26	3.28	3.67	4.64	3.87	3.39	4.36	3.72	3.21	4.45	3.39	3.33	4.06	3.46

Table 13: Fine-grained evaluation of translation quality across multiple dimensions for Zh→Es, Zh→Ja, and Zh→Ru directions.

Model	Contextual Accuracy			Cultural Adaptation			Functional Equivalence			Fidelity			Naturalness		
	En→Es	En→Ja	En→Zh	En→Es	En→Ja	En→Zh	En→Es	En→Ja	En→Zh	En→Es	En→Ja	En→Zh	En→Es	En→Ja	En→Zh
Llama-3-8B-Instruct-262k	4.73	2.76	4.07	4.97	3.66	4.53	3.81	2.24	3.80	4.13	2.54	3.77	3.86	2.11	3.63
Llama-3.3-70B-Instruct	5.08	5.26	5.20	5.28	5.29	5.20	4.79	4.51	4.96	4.90	4.83	5.22	4.42	4.17	4.52
Mixtral-8x7B-Instruct-v0.1	4.67	3.14	3.48	4.91	3.67	3.90	3.98	2.54	2.55	4.10	2.76	2.72	3.85	2.46	2.63
Qwen2.5-14B-Instruct	4.68	4.38	5.14	4.87	4.63	5.46	4.29	3.93	4.89	4.49	4.08	4.97	4.02	3.56	4.55
Qwen2.5-32B-Instruct	4.84	4.59	5.45	4.81	4.99	5.37	4.61	4.03	5.10	4.77	4.12	5.08	4.14	3.72	4.65
Qwen2.5-72B-Instruct	4.93	5.11	5.52	5.19	5.12	5.67	4.63	4.76	5.07	4.97	4.94	5.22	4.46	4.29	4.65
Qwen2.5-7B-Instruct	4.63	4.06	4.77	4.88	4.21	5.06	3.85	3.38	4.41	4.16	3.44	4.55	3.88	3.09	4.23
Qwen3-8B (w/o think)	4.45	4.57	5.03	5.06	4.62	5.32	3.96	4.14	4.76	4.25	4.26	4.71	3.96	3.67	4.40
Qwen3-8B (with think)	4.48	4.34	5.17	4.92	4.67	5.37	4.25	3.91	4.85	4.45	4.17	5.01	4.08	3.58	4.54
DeepSeek-V3.2	4.92	5.55	5.52	5.33	5.74	5.55	4.69	5.22	5.31	5.19	5.47	5.28	4.62	4.63	4.81
Gemini-2.5-Flash-Lite	4.94	5.30	5.40	5.16	5.43	5.40	4.53	4.92	5.00	4.79	4.98	5.07	4.37	4.50	4.54
GPT-4o	4.92	5.30	5.39	5.31	5.37	5.40	4.67	4.89	5.05	5.10	5.34	5.10	4.66	4.39	4.65

Table 14: Fine-grained evaluation of semantic translation quality across multiple dimensions for En→Es, En→Ja, and En→Zh directions.

==========

Examine the fluency and native feel of the translation in the target language, assessing whether the sentence reads like a mother-tongue expression rather than a rigid transplant. During the evaluation of naturalness, you need to assume the identity of a target language reader, and assess whether the translation is readable, natural, and fluent.

- **1 point (Extremely Poor)**: The translation is extremely awkward and unnatural, reading like a forced or incomprehensible transplant.
- **2 points (Poor)**: The translation lacks fluency, with major stiffness or non-native phrasing that hinders readability.
- **3 points (Relatively Poor)**: The translation is somewhat readable but includes noticeable unnatural elements or awkward flow.
- **4 points (Average)**: The translation achieves basic naturalness, reading adequately but without full native fluency.
- **5 points (Relatively Good)**: The translation is mostly natural and fluent, with good readability and minor improvements possible.
- **6 points (Good)**: The translation flows naturally, resembling native expression with high readability.
- **7 points (Excellent)**: The translation is perfectly natural and seamless, indistinguishable from original target-language writing in fluency and rhythm.

ExampleIt is a far, far better thing that I do, than I have ever done; it is a far, far better rest that I go to than I have ever known.
--from A Tale of Two Cities, the translation should render the repetitive, idiomatic structure of this long proverbial farewell idiomatically and fluidly in the target language, as if it were a native expression to ensure naturalness.

==========

Figure 12: Instruction for human eval Naturalness.

Model	Contextual Accuracy			Cultural Adaptation			Functional Equivalence			Fidelity			Naturalness		
	Es→En	Es→Zh	Ja→En	Es→En	Es→Zh	Ja→En	Es→En	Es→Zh	Ja→En	Es→En	Es→Zh	Ja→En	Es→En	Es→Zh	Ja→En
Llama-3-8B-Instruct-262k	4.33	3.19	3.37	4.86	3.69	4.06	4.06	3.24	3.27	3.97	3.17	3.25	4.03	3.13	3.52
Llama-3.3-70B-Instruct	5.21	4.42	4.59	5.15	4.63	4.75	4.76	4.81	3.89	4.96	4.64	4.01	4.34	4.44	4.07
Mixtral-8x7B-Instruct-v0.1	4.99	3.13	3.46	5.13	3.67	4.12	4.73	2.52	3.37	4.74	2.51	3.33	4.29	2.48	3.39
Qwen2.5-14B-Instruct	4.96	4.30	4.36	5.27	4.75	4.68	4.72	4.84	4.30	4.96	4.73	4.35	4.40	4.29	4.22
Qwen2.5-32B-Instruct	5.09	4.55	4.59	5.27	4.72	4.87	4.88	4.87	4.06	4.92	4.75	4.35	4.36	4.42	4.29
Qwen2.5-72B-Instruct	5.10	4.70	4.91	5.25	4.98	4.97	4.93	4.90	4.48	5.16	4.99	4.66	4.57	4.58	4.52
Qwen2.5-7B-Instruct	4.69	4.04	3.63	5.00	4.43	4.11	4.56	4.23	3.80	4.41	4.07	3.68	4.07	3.81	3.78
Qwen3-8B (w/o think)	4.79	4.44	3.82	5.11	4.60	4.10	4.37	4.75	3.86	4.48	4.48	3.76	3.94	4.34	3.91
Qwen3-8B (with think)	4.47	4.30	4.15	4.82	4.52	4.43	4.17	4.58	3.84	4.51	4.57	3.99	3.91	4.10	3.78
DeepSeek-V3.2	5.46	4.98	5.24	5.44	5.19	5.28	4.95	5.17	4.55	5.02	4.95	4.84	4.24	4.52	4.51
Gemini-2.5-Flash-Lite	5.31	4.58	4.91	5.41	4.93	5.16	4.81	4.78	4.33	5.14	4.80	4.72	4.26	4.31	4.25
GPT-4o	5.49	4.80	5.09	5.46	4.99	5.28	5.04	5.01	4.66	5.26	4.80	5.03	4.59	4.37	4.49

Table 15: Fine-grained evaluation of semantic translation quality across multiple dimensions for Es→En, Es→Zh, and Ja→En directions.

Model	Contextual Accuracy			Cultural Adaptation			Functional Equivalence			Fidelity			Naturalness		
	Ja→Zh	Ru→Zh	Zh→En	Ja→Zh	Ru→Zh	Zh→En	Ja→Zh	Ru→Zh	Zh→En	Ja→Zh	Ru→Zh	Zh→En	Ja→Zh	Ru→Zh	Zh→En
Llama-3-8B-Instruct-262k	3.35	3.58	4.34	3.92	3.98	4.59	2.98	3.61	4.38	3.08	3.41	4.27	3.12	3.40	4.37
Llama-3.3-70B-Instruct	4.28	4.56	5.26	4.75	4.76	5.27	4.32	4.73	5.09	4.47	4.75	5.26	4.11	4.45	4.58
Mixtral-8x7B-Instruct-v0.1	3.19	2.87	4.58	3.77	3.60	4.60	2.89	2.62	4.65	2.92	2.57	4.74	2.89	2.60	4.18
Qwen2.5-14B-Instruct	4.49	4.50	5.44	4.84	4.91	5.46	4.61	4.62	5.31	4.75	4.50	5.39	4.32	4.30	4.72
Qwen2.5-32B-Instruct	4.51	4.66	5.38	4.86	5.02	5.34	4.63	4.83	4.98	4.77	4.74	5.30	4.35	4.46	4.58
Qwen2.5-72B-Instruct	4.97	4.71	5.61	5.03	4.95	5.64	4.71	4.94	5.32	5.00	4.92	5.62	4.42	4.71	4.95
Qwen2.5-7B-Instruct	4.38	4.39	5.00	4.78	4.48	5.25	4.34	4.29	4.96	4.42	4.30	4.94	4.29	4.16	4.57
Qwen3-8B (w/o think)	4.16	4.56	5.12	4.66	4.93	5.30	4.03	4.76	5.02	4.18	4.43	5.11	3.98	4.53	4.70
Qwen3-8B (with think)	4.46	4.49	5.09	4.82	4.77	5.12	4.31	4.62	5.10	4.61	4.71	5.35	4.09	4.14	4.56
DeepSeek-V3.2	4.95	4.91	5.54	5.20	5.03	5.42	4.92	5.02	5.32	5.00	4.72	5.54	4.47	4.58	4.76
Gemini-2.5-Flash-Lite	4.58	4.92	5.58	4.96	5.10	5.37	4.51	4.90	4.97	4.80	4.98	5.46	4.25	4.46	4.58
GPT-4o	5.06	4.77	5.66	5.27	5.10	5.31	4.97	4.62	5.18	5.23	4.83	5.54	4.49	4.22	4.71

Table 16: Fine-grained evaluation of semantic translation quality across multiple dimensions for Ja→Zh, Ru→Zh, and Zh→En directions.

```
================================================================================
You are an expert evaluator. Please assess the translation below based on the following instructions.
PART 1: INPUT DATA
Source Language: {src_lang}
Target Language: {tgt_lang}
CSIs: {CSIs}
Source Sentence: {src_text}
Translation: {tgt_text}
Reference Translation: {ref_text}
PART 2: SCORING CRITERIA
Contextual Accuracy (Definition)
Concept: CSI (Culture-Specific Items) expressing unique cultural concepts that were initially difficult to find exact counterpart in different culture.
You need to determine whether the translation of CSI accurately conveys the meaning intended by the author. Other parts of the sentence are not within the scope of consideration.
When considering contextual accuracy assessment, if a concept is closely related to a cultural-specific item (CSI) in the sentence, and its mistranslation would affect the target language reader's understanding of the CSI, then the translation of this word/phrase should also be included in the evaluation.
Other parts of the sentence are not within the scope of consideration. The semantic fidelity of the entire sentence is not within the scope of consideration in this dimension.
Scoring Rubric (1-7):
{Scoring Rubrics for Evaluation }
Critical Rules
MANDATORY RE-TRANSLATION: If score < 7, you MUST provide a better translation of CSI, or the translation of content that affects the understanding of CSI in the "reason" part.
Untranslated Content Limit: If CSIs are NOT translated into {tgt_lang}, the score MUST NOT exceed 2.
Output Language: The "reason" must be in English.
Focus only on the CSI provided.
The semantic fidelity of the entire sentence is not within the scope of consideration in this dimension.
PART 3: OUTPUT REQUIREMENT
Please output the result in strictly valid JSON format exactly as shown below:{
"Contextual_Accuracy": {
"score": 1-7,
"reason": "Your explanation here, including the CSIs' translation in the reference translation. IF SCORE < 7, PROVIDE BETTER TRANSLATION."
}}
================================================================================
```

Figure 13: Prompt for eval Contextual Accuracy.

```
================================================================================
You are an expert evaluator. Please assess the translation below based on the following instructions.
PART 1: INPUT DATA
Source Language: {src_lang}
Target Language: {tgt_lang}
CSIs: {CSIs}
Source Sentence: {src_text}
Translation: {tgt_text}
Reference Translation: {ref_text}
PART 2: SCORING CRITERIA
Cultural Adaptation (Definition)
Concept: CSI (Culture-Specific Items) expressing unique cultural concepts.
Assess whether the translation of culture-specific items achieves adaptation to the {tgt_lang} culture through adjustments to the culturally specific item, thereby ensuring better understanding by {tgt_lang} readers while avoiding potential cultural conflicts. During the evaluation of cultural adaptation, assume the identity of a {tgt_lang} reader, first identify the corresponding translation of the culturally specific item and related concepts, then consider whether they correctly convey the meaning, retain the cultural nuance, conform to {tgt_lang} culture norms, and whether the translation could cause cultural conflicts.
Using a standard, accepted translation for a culturally specific term is not just adequate—it can be the optimal solution, worthy of the highest score. Such a translation has undergone cultural negotiation, achieving perfect adaptation, ensuring immediate recognition, zero misunderstanding, and seamless integration for the target reader.
Other parts of the sentence are not within the scope of consideration. The semantic fidelity of the entire sentence is not within the scope of consideration in this dimension.
Scoring Rubric (1-7)
{Scoring Rubrics for Evaluation }
Critical Rules
MANDATORY RE-TRANSLATION: If score < 7, you MUST provide a better translation of CSI, or the translation of content that affects understanding of CSI in the "reason" part.
Untranslated Content Limit: If CSIs are NOT translated into {tgt_lang}, the score MUST NOT exceed 4.
Output Language: The "reason" must be in English.
Focus only on the CSI provided.
The semantic fidelity of the entire sentence is not within the scope of consideration in this dimension.
PART 3: OUTPUT REQUIREMENT
Please output the result in strictly valid JSON format exactly as shown below:
{"Cultural_Adaptation": {
"reason": "Your explanation here, including the CSIs' translation in the reference translation. IF SCORE < 7, PROVIDE BETTER TRANSLATION about CSI.",
"score": 1-7}}
================================================================================
```

Figure 14: Prompt for eval Cultural Adaptation.

Model	Contextual Accuracy			Cultural Adaptation			Functional Equivalence			Fidelity			Naturalness		
	Zh→Es	Zh→Ja	Zh→Ru	Zh→Es	Zh→Ja	Zh→Ru	Zh→Es	Zh→Ja	Zh→Ru	Zh→Es	Zh→Ja	Zh→Ru	Zh→Es	Zh→Ja	Zh→Ru
Llama-3-8B-Instruct-262k	3.35	3.29	2.99	3.92	3.85	3.74	3.91	3.34	3.67	3.56	3.20	3.34	3.72	2.98	3.36
Llama-3.3-70B-Instruct	4.91	4.58	4.72	4.91	4.57	4.85	4.83	4.72	4.82	5.05	4.80	4.86	4.49	4.19	4.38
Mixtral-8x7B-Instruct-v0.1	3.95	3.13	3.46	4.26	3.43	3.89	3.92	2.71	3.52	3.97	2.72	3.49	3.86	2.44	3.25
Qwen2.5-14B-Instruct	4.34	4.41	3.67	4.59	4.45	4.29	4.55	4.42	3.84	4.48	4.34	3.93	4.26	3.75	3.36
Qwen2.5-32B-Instruct	4.69	4.48	4.03	4.83	4.66	4.39	4.66	4.47	4.12	4.83	4.47	4.09	4.27	3.90	3.76
Qwen2.5-72B-Instruct	5.03	4.73	4.60	4.92	4.79	4.79	5.02	4.85	4.83	5.19	4.92	4.93	4.79	4.37	4.52
Qwen2.5-7B-Instruct	3.83	3.78	3.31	4.21	4.12	4.00	3.94	3.96	3.60	3.93	3.92	3.29	3.75	3.57	3.08
Qwen3-8B (w/o think)	4.23	4.17	3.71	4.51	4.52	3.96	4.55	4.36	4.18	4.30	4.37	3.88	4.07	3.83	3.88
Qwen3-8B (with think)	4.68	4.32	4.20	4.92	4.53	4.42	4.68	4.82	4.68	4.83	5.08	4.73	4.29	4.15	4.09
DeepSeek-V3.2	5.21	5.26	4.89	5.27	5.19	5.21	4.93	5.33	5.01	5.13	5.58	4.93	4.76	4.73	4.35
Gemini-2.5-Flash-Lite	5.03	5.41	5.21	5.00	5.18	5.27	4.86	5.18	5.17	5.23	5.52	5.29	4.42	4.68	4.57
GPT-4o	5.18	5.18	5.14	5.16	4.96	5.25	4.97	5.10	5.22	5.18	5.34	5.43	4.61	4.57	4.51

Table 17: Fine-grained evaluation of semantic translation quality across multiple dimensions for Zh→Es, Zh→Ja, and Zh→Ru directions.

Model	Contextual Accuracy			Cultural Adaptation			Functional Equivalence			Fidelity			Naturalness		
	En→Es	En→Ja	En→Zh	En→Es	En→Ja	En→Zh	En→Es	En→Ja	En→Zh	En→Es	En→Ja	En→Zh	En→Es	En→Ja	En→Zh
Llama-3-8B-Instruct-262k	4.59	3.17	4.23	5.03	3.81	4.50	3.84	2.76	3.83	4.16	2.78	3.78	3.97	2.74	3.64
Llama-3.3-70B-Instruct	4.92	5.17	4.99	5.16	5.39	5.20	4.82	4.76	4.92	4.80	4.79	4.93	4.63	4.30	4.83
Mixtral-8x7B-Instruct-v0.1	4.69	3.24	3.69	5.14	3.74	3.94	4.41	2.44	2.58	4.25	2.58	2.65	4.17	2.53	2.57
Qwen2.5-14B-Instruct	4.45	4.72	5.30	4.76	4.78	5.45	4.45	4.32	5.05	4.29	4.25	4.90	4.21	4.00	4.91
Qwen2.5-32B-Instruct	4.50	4.65	5.35	4.95	4.95	5.32	4.51	4.37	5.06	4.42	4.30	5.02	4.28	4.13	4.75
Qwen2.5-72B-Instruct	4.92	5.06	5.59	5.18	5.20	5.64	4.75	4.69	4.98	4.86	4.72	5.19	4.71	4.46	5.03
Qwen2.5-7B-Instruct	4.29	4.00	4.78	4.81	4.31	5.14	3.92	3.29	4.49	4.09	3.35	4.52	3.82	3.14	4.30
Qwen3-8B (w/o think)	4.73	4.53	5.16	5.02	4.73	5.32	4.26	4.27	4.85	4.43	4.44	4.83	4.16	3.95	4.49
Qwen3-8B (with think)	4.75	4.79	5.33	5.08	4.92	5.53	4.50	4.55	4.93	4.59	4.23	4.97	4.28	3.85	4.77
DeepSeek-V3.2	4.99	5.04	5.32	5.34	5.31	5.54	4.92	4.95	5.10	4.71	4.95	4.90	4.69	4.60	4.95
Gemini-2.5-Flash-Lite	4.77	5.43	5.28	5.24	5.54	5.27	4.81	5.07	5.02	4.83	5.00	4.93	4.58	4.84	4.69
GPT-4o	5.09	5.27	5.63	5.15	5.44	5.55	4.59	5.05	5.23	4.71	4.93	5.19	4.66	4.81	5.02

Table 18: Fine-grained evaluation of communicative translation quality across multiple dimensions for En→Es, En→Ja, and En→Zh directions.

```
================================================================================
You are an expert evaluator. Please assess the translation below based on the following instructions.
PART 1: INPUT DATA
Source Language: {src_lang}
Target Language: {tgt_lang}
Source Sentence: {src_text}
Translation: {tgt_text}
Reference Translation: {ref_text}
PART 2: SCORING CRITERIA
Functional Equivalence (Definition)
Examine whether the translation achieves the function of the {src_lang} sentence, such as providing information to the reader, expressing criticism, persuading, or evoking certain emotions.
During the evaluation, assume the identity of the translator, infer what function the original author intended in the {src_lang} text, and assess whether the translation realizes that function.
Scoring Rubric (1-7)
{Scoring Rubrics for Evaluation}
Critical Rules
Language Penalty: If the translation contains non-{tgt_lang} words (language mixing), score it 1 or 2.
Untranslated Content Limit: If there are words that are NOT translated into {tgt_lang}, the score MUST NOT exceed 4.
Output Language: The "reason" must be in English.
PART 3: OUTPUT REQUIREMENT
Please output the result in strictly valid JSON format exactly as shown below:
{"Functional_Equivalence": {
"score": "Score from 1 to 7",
"reason": ""}}
================================================================================
```

Figure 15: Prompt for eval Functional Equivalence.

```
================================================================================
You are an expert evaluator. Please assess the translation below based on the following instructions.

PART 1: INPUT DATA
Source Language: {src_lang}
Target Language: {tgt_lang}
Source Sentence: {src_text}
Translation: {tgt_text}
Reference Translation: {ref_text}

PART 2: SCORING CRITERIA
Fidelity (Definition)
Assess whether the translation maximally retains the literal meaning, core informational content, and source language structure of the original text, specifically avoiding semantic loss caused by paraphrasing or rewriting.
This dimension focuses strictly on the accuracy of informational transfer (what is said), NOT style, tone, or communicative effect (how it is said).
During the evaluation of fidelity, assume the identity of the translator, and assess whether there are errors, omissions, additions of unstated information, or factual distortions in the target text.
The translation must stay faithful to the precise expression and literal meaning of the source language text.
Scoring Rubric (1-7)
{Scoring Rubrics for Evaluation}
Critical Rules
Language Penalty:If the translation contains non-{tgt_lang} words (language mixing), score it 1 or 2.
Untranslated Content Limit:If there are words that are NOT translated into {tgt_lang}, the score MUST NOT exceed 4.
Output Language:The "reason" must be in English.

PART 3: OUTPUT REQUIREMENT
Please output the result in strictly valid JSON format exactly as shown below:
{
"Fidelity": {
"score": "Score from 1 to 7",
"reason": "Your explanation here"}}
================================================================================
```

Figure 16: Prompt for eval Fidelity.

Model	Contextual Accuracy			Cultural Adaptation			Functional Equivalence			Fidelity			Naturalness		
	Es→En	Es→Zh	Ja→En	Es→En	Es→Zh	Ja→En	Es→En	Es→Zh	Ja→En	Es→En	Es→Zh	Ja→En	Es→En	Es→Zh	Ja→En
Llama-3-8B-Instruct-262k	4.38	3.22	3.19	4.78	3.73	3.72	4.16	3.28	3.39	3.91	3.08	3.33	4.10	3.22	3.68
Llama-3.3-70B-Instruct	4.82	4.36	4.44	5.20	4.67	4.88	5.09	4.72	4.43	4.51	4.44	4.40	4.99	4.44	4.70
Mixtral-8x7B-Instruct-v0.1	4.74	3.23	3.35	5.16	3.66	4.02	4.81	2.74	3.57	4.43	2.66	3.17	4.73	2.67	3.97
Qwen2.5-14B-Instruct	4.96	4.30	4.40	5.19	4.78	4.68	5.04	4.73	4.36	4.70	4.45	4.34	4.66	4.55	4.54
Qwen2.5-32B-Instruct	4.89	4.39	4.67	5.20	4.63	4.99	5.00	4.76	4.64	4.60	4.40	4.53	4.92	4.63	4.78
Qwen2.5-72B-Instruct	5.04	4.57	4.80	5.31	5.00	5.00	5.07	5.07	4.72	4.81	4.62	4.80	5.00	4.69	4.86
Qwen2.5-7B-Instruct	4.77	3.99	3.61	5.05	4.33	4.26	4.60	4.23	3.91	4.27	3.98	3.53	4.46	4.05	3.92
Qwen3-8B (w/o think)	4.63	4.25	3.82	5.00	4.67	4.27	4.49	4.69	4.04	4.47	4.35	3.89	4.38	4.33	4.18
Qwen3-8B (with think)	4.88	4.47	4.33	5.01	4.55	4.61	4.65	4.70	4.38	4.67	4.48	4.44	4.47	4.43	4.36
DeepSeek-V3.2	5.24	4.69	4.95	5.29	5.06	5.21	5.08	5.10	4.75	4.62	4.48	4.68	5.05	4.67	4.82
Gemini-2.5-Flash-Lite	4.96	4.40	5.08	5.25	4.79	5.28	5.04	4.93	4.76	4.96	4.63	4.89	5.05	4.75	4.78
GPT-4o	4.91	4.65	4.97	5.17	4.95	5.17	5.15	4.89	4.89	4.66	4.58	5.01	4.98	4.71	5.00

Table 19: Fine-grained evaluation of communicative translation quality across multiple dimensions for Es→En, Es→Zh, and Ja→En directions.

Model	Contextual Accuracy			Cultural Adaptation			Functional Equivalence			Fidelity			Naturalness		
	Ja→Zh	Ru→Zh	Zh→En	Ja→Zh	Ru→Zh	Zh→En	Ja→Zh	Ru→Zh	Zh→En	Ja→Zh	Ru→Zh	Zh→En	Ja→Zh	Ru→Zh	Zh→En
Llama-3-8B-Instruct-262k	3.25	3.26	4.46	3.74	3.91	4.58	3.11	3.55	4.50	3.30	3.25	4.18	3.32	3.64	4.43
Llama-3.3-70B-Instruct	4.40	4.22	5.10	4.73	4.64	5.25	4.50	4.66	5.30	4.40	4.34	4.93	4.52	4.58	5.08
Mixtral-8x7B-Instruct-v0.1	3.11	3.33	4.78	3.54	3.83	4.98	2.93	2.72	4.82	2.78	2.74	4.74	3.00	2.78	4.68
Qwen2.5-14B-Instruct	4.16	4.39	5.30	4.72	4.83	5.41	4.73	4.74	5.38	4.76	4.43	5.08	4.71	4.70	5.04
Qwen2.5-32B-Instruct	4.62	4.51	5.06	5.00	4.78	5.38	4.91	4.76	5.19	4.79	4.47	4.98	4.74	4.79	4.99
Qwen2.5-72B-Instruct	5.08	4.76	5.33	5.19	5.00	5.54	4.95	4.98	5.34	5.01	5.03	5.18	4.86	4.95	5.07
Qwen2.5-7B-Instruct	4.36	4.19	5.15	4.70	4.69	5.16	4.40	4.40	4.92	4.24	4.09	4.97	4.29	4.25	4.74
Qwen3-8B (w/o think)	4.40	4.36	4.98	4.89	4.77	5.08	4.36	4.74	5.08	4.50	4.47	5.01	4.33	4.61	4.79
Qwen3-8B (with think)	4.52	4.44	5.30	4.90	4.77	5.36	4.67	4.66	5.24	4.85	4.60	5.21	4.59	4.41	4.90
DeepSeek-V3.2	5.07	4.64	5.29	5.20	5.06	5.38	5.01	4.99	5.26	4.72	4.60	5.06	4.71	4.69	5.04
Gemini-2.5-Flash-Lite	4.92	4.66	5.63	5.02	5.15	5.54	5.02	4.84	5.32	4.99	4.73	5.35	4.75	4.70	5.16
GPT-4o	4.89	4.49	5.48	5.05	4.97	5.47	4.96	4.77	5.41	4.97	4.52	5.20	5.02	4.79	5.32

Table 20: Fine-grained evaluation of communicative translation quality across multiple dimensions for Ja→Zh, Ru→Zh, and Zh→En directions.

```
================================================================================
PART 1: INPUT DATA
Source Language: {src_lang}
Target Language: {tgt_lang}
Source Sentence: {src_text}
Translation: {tgt_text}
Reference Translation: {ref_text}
PART 2: SCORING CRITERIA
Naturalness (Definition)
Examine the fluency and native feel of the translation in {tgt_lang}, assessing whether the
sentence reads like a mother-tongue expression rather than a rigid transplant.
During the evaluation, assume the identity of a {tgt_lang} reader, and assess whether the
translation is readable, natural, and fluent.
Scoring Rubric (1-7)
{Scoring Rubrics for Evaluation}
Critical Rules
Language Penalty: If the translation contains non-{tgt_lang} words (language mixing), score it 1 or
2.
Untranslated Content Limit: If there are words that are NOT translated into {tgt_lang}, the score
MUST NOT exceed 4.
Output Language: The "reason" must be in English.
PART 3: OUTPUT REQUIREMENT
Please output the result in strictly valid JSON format exactly as shown below:
{"Naturalness": {
"score": "Score from 1 to 7",
"reason": ""}}
================================================================================
```

Figure 17: Prompt for eval Naturalness.

Model	Contextual Accuracy			Cultural Adaptation			Functional Equivalence			Fidelity			Naturalness		
	Zh→Es	Zh→Ja	Zh→Ru	Zh→Es	Zh→Ja	Zh→Ru	Zh→Es	Zh→Ja	Zh→Ru	Zh→Es	Zh→Ja	Zh→Ru	Zh→Es	Zh→Ja	Zh→Ru
Llama-3-8B-Instruct-262k	3.23	3.34	2.85	3.83	3.88	3.63	3.80	3.47	3.47	3.59	3.36	3.23	3.90	3.22	3.44
Llama-3.3-70B-Instruct	4.73	4.72	4.34	4.99	4.88	4.85	4.86	4.80	4.74	4.81	4.81	4.66	4.88	4.37	4.71
Mixtral-8x7B-Instruct-v0.1	4.29	3.23	3.84	4.63	3.71	4.20	4.40	2.80	3.46	4.02	2.82	3.23	4.26	2.61	3.26
Qwen2.5-14B-Instruct	4.33	4.47	3.58	4.71	4.72	4.27	4.83	4.61	4.12	4.60	4.57	3.84	4.52	4.20	3.88
Qwen2.5-32B-Instruct	4.47	4.28	4.12	4.79	4.66	4.46	4.75	4.62	4.39	4.53	4.47	4.11	4.56	4.33	4.02
Qwen2.5-72B-Instruct	4.82	4.57	4.35	5.11	4.82	4.65	5.01	4.88	4.97	5.01	4.71	4.79	4.99	4.63	4.69
Qwen2.5-7B-Instruct	3.91	3.80	3.20	4.32	4.20	3.75	4.13	3.93	3.42	3.97	3.88	3.39	4.01	3.60	3.17
Qwen3-8B (w/o think)	4.16	4.26	3.65	4.57	4.49	4.01	4.38	4.55	4.45	4.29	4.37	4.15	4.24	4.01	4.02
Qwen3-8B (with think)	4.68	4.56	3.98	4.82	4.64	4.35	4.77	4.82	4.60	4.83	5.05	4.38	4.49	4.31	4.05
DeepSeek-V3.2	4.79	5.06	4.37	5.12	5.22	4.68	5.01	5.32	3.93	4.78	5.07	4.02	4.65	4.93	3.69
Gemini-2.5-Flash-Lite	5.01	5.12	4.81	5.18	5.40	5.23	5.08	5.43	5.25	5.09	5.44	5.11	4.89	4.99	4.73
GPT-4o	4.84	4.90	5.06	4.96	5.26	5.16	5.18	5.47	5.25	5.01	5.38	5.11	4.99	5.03	4.92

Table 21: Fine-grained evaluation of communicative translation quality across multiple dimensions for Zh→Es, Zh→Ja, and Zh→Ru directions.

CSI Category	Definition	Examples
Geographic and Ecological Items	Elements representing natural or geographical phenomena specific to a culture.	Everest, Yellowstone
Material Culture Items	Tangible objects and artifacts unique to a culture's lifestyle and craftsmanship.	Baseball, Cowboy Hat
Social Culture and Customs	Practices, rituals distinctive to a culture.	Thanksgiving, Halloween
Organizations and Institutions	Structured entities or systems that organize societal, political, or educational life.	NASA, Harvard University
Language Symbols	Linguistic expressions or symbolic language tied to a specific culture.	No pain, no gain

Figure 18: Definitions and representative examples of CSI categories.

```
================================================================================
Target Word: {word}
Task: Analyze the cultural semantics of the target word and classify it.
Please classify the target word into EXACTLY ONE of the following 5 major categories based or
its cultural semantic meaning.
You must provide both the Category ID and the exact Category Name.
    ID 1: Geographic and ecological items
        - Covers: Ecological concepts, geography, landforms, climatic phenomena.
    ID 2: Material culture items
        - Covers: Food & drink, clothing, architecture, crafts, weights & measures.
    ID 3: Social culture and customs
        - Covers: Festivals, etiquette, taboos, customs, kinship, religion, mythology.
    ID 4: Organizations and institutions
        - Covers: Political concepts, government institutions, legal systems, social welfare
    education.
    ID 5: Language symbols
        - Covers: Proverbs, slang, poetry, allusions, idioms, wordplay, pop culture.
Output Requirement:
    - Return ONLY a JSON object.
    - "category_id": Must be an integer from 1 to 5.
    - "category_name": Must be the EXACT string name corresponding to the ID.
    - Format strictly as:
        {
        "category_id" : ,
        "category_name" : "",
        "reason": "Brief reason for this classification"
        }
================================================================================
```

Figure 19: Prompt used for automatic classification of Culture-Specific Items.

```
================================================================================
Source:
一过长江，我想逃跑的心也死了，离家越远我也就越没有胆量逃跑。
Translation:
As soon as I crossed the Yangtze River, my heart died and I lost the courage to run away from
home.
CSI: 长江 → Yangtze River
Commentary:
The geographic CSI admits a stable and conventionalized English equivalent.The model output
adopt standard renderings, resulting in high contextual accuracy and cultural adaptation.
================================================================================
```

Figure 20: Translation of Geography and Ecology CSI

```
================================================================================
Source:
我从小就不可救药，这是我爹的话。
Translation:
I was raised without medicine, my mother said.
CSIs: 不可救药
Commentary:
The CSI "不可救药" conveys an idiomatic judgment meaning "hopeless beyond remedy," but the
translation renders it literally and alters both the meaning and the speaker, resulting in semantic
distortion and pragmatic mismatch.
================================================================================
```

Figure 21: Translation of Language symbols CSI.

```
You are a professional translator and translation evaluator.
Task:
Given a source sentence in {src_lang} containing a Culture-Specific Item (CSI): "{csi_text}",
design a single-choice question asking:
What is the most appropriate translation of this CSI into {tgt_lang} in this context?

You are also given a reference translation for internal judgment only.
IMPORTANT:
    - The reference translation is for answer calibration ONLY and must not be mentioned.
    - The question must ask about the best translation choice in general.
Input:
    - Source Text: "{src_text}"
    - Reference Translation (internal use only): "{tgt_text}"
    - Target CSI: "{csi_text}"
Requirements:
1. The question must explicitly ask how to translate "{csi_text}" into {tgt_lang} in this context.
The question text must be written in English.
2. All options must be written in {tgt_lang} and be grammatically well-formed.
3. Exactly ONE option must preserve both the semantic meaning and cultural function of the CSI.
4. All distractor options must be lexically or structurally close to the correct translation,
but express a meaning that is semantically incompatible with the CSI in this context,
such that choosing them would lead to serious misunderstanding.
Distractors MUST contain a clear semantic or cultural error, such as:
    - incorrect referent or denotation,
    - wrong cultural function or usage,
    - altered pragmatic meaning,
    - misinterpreted idiom, metaphor, or culture-specific practice.
Differences based only on wording preference, style, register, or literalness are STRICTLY
FORBIDDEN.
5. The correct option must be consistent with how the CSI is treated in the reference translation
(even if translated implicitly or paraphrased).
6. The analysis MUST:
- justify why the correct option preserves the CSI's meaning and cultural function;
- explicitly explain why each distractor fails due to a semantic or cultural mismatch;
- NOT suggest that any incorrect option is acceptable in another context.
Output strictly in JSON (no markdown):{
    "question": "The question stem...",
    "options": { "A": "Option A text",      "B": "Option B text",
                 "C": "Option C text",      "D": "Option D text"},
    "answer": "A",
    "analysis": "Explanation of why this option is the most appropriate translation in context."}
```

Figure 22: Prompt for Generating Culture-Specific Translation SCQs.

Source (ZH):女人都是一个心眼，她认准的事谁也不能让她变。
CSI:心眼
Reference (EN): Women are stubborn—once they get their minds set on something, no one can budge them.

Question: What is the most appropriate translation of "心眼" into English in this context?
Options: A: stubbornness B: eyes in one's heart C: inner wisdom D: mind's eye

---

Model Choice: A – Correct
Model Trans:Women are all heart-eyed, and once they set their minds on something, no one can change their minds.

Figure 23: Case Study: Knowledge–Application Gap

"Source": "没出三个月，那四十**斤**米全吃光了。"
"Translation": "Within three months, all the 40 **jin** of rice were eaten up."
"Reference": "In less than three months the forty **jin** of rice was gone."

Evaluation w/o Reference:
"score": 2.0
"reason": "The CSI '斤' is left untranslated in the target, which hinders English readers from understanding the cultural unit of weight. A better translation would translate the quantity into a familiar unit: 'Within three months, all twenty kilograms of rice were eaten.' 40jin = 20kg."

Evaluation with Reference:
"score": 6.0
"reason": "The CSI '斤' is correctly identified and conveys a quantity. The given translation uses the direct unit 'jin' , which is technically accurate but not natural for English readers..... "

Figure 24: Case Study on Contextual Accuracy: The evaluator penalizes the retention of the specific Chinese unit "jin" (Score: 2), suggesting a domesticated conversion to "kilograms" instead. However, the provision of a reference that also retains "jin" validates the accuracy of the target translation, confirming that preserving the source cultural unit is the correct strategy in this context (Score: 6).

"Source": ""俗话说是**笨鸟先飞**，我还得笨鸟多飞。"
"Translation": "As the saying goes, **the early bird catches the worm**, but I have to be the early bird that flies more."
"Reference": "As the saying goes, "**Slow birds need an early start.**" Well, I was the slow bird who never finished"

Evaluation w/o Reference:
"score": 7.0
"reason":"The CSIs map to standard English equivalents: 笨鸟先飞 -- 'the early bird catches the worm"....."

Evaluation with Reference:
"score": 4.0
"reason": "..... but translates '笨鸟先飞' as 'the early bird' and '笨鸟多飞' as 'the early bird that flies more', which distorts the meaning: it replaces '笨' (slow/not clever) with 'early bird', an unrelated positive stereotype ..... "

Figure 25: Case Study on Cultural Adaptation: The evaluator initially praises the target for replacing the source CSI with the familiar English idiom "the early bird" (Score: 7). However, the reference translation reveals that this adaptation distorts the original cultural image of "slow bird". The reference enables the evaluator to identify the loss of the specific cultural connotation of humility, resulting in a penalized score (Score: 4).

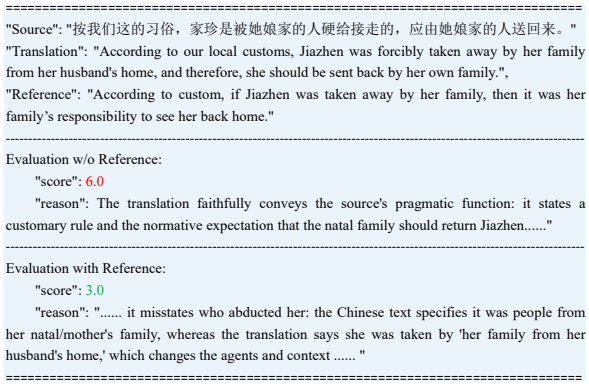

Figure 26: Case Study on Functional Equivalence: Without a reference, the evaluator overlooks kinship logic and assigns a high score for capturing the custom's pragmatic function (Score: 6). With the reference, a critical agent error is identified—misattributing Jiazhen's removal to her husband's family rather than her natal family—revealing a factual distortion that reduces functional equivalence (Score: 3).

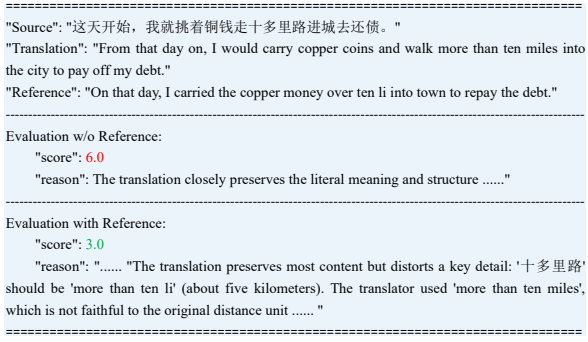

Figure 27: Case Study on Fidelity: The evaluator initially assigns a high score, assuming the translation closely preserves the literal meaning and structure of the source (Score: 6). However, the reference translation provides a precise baseline that reveals a critical fine-grained semantic error: the mistranslation of the Chinese distance unit "li" as "miles". This unit distortion, which significantly alters the factual scale of the narrative, is only identified through direct comparison with the reference, resulting in a more rigorous evaluation (Score: 3).

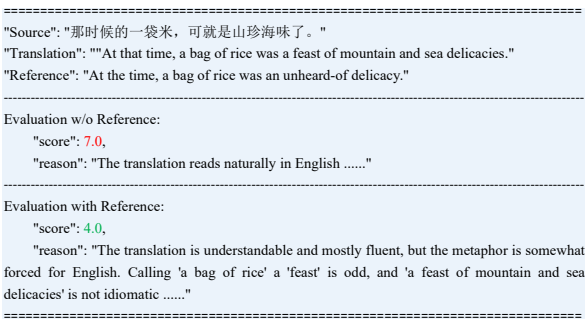

Figure 28: Case Study on Naturalness: The reference translation provides a benchmark, favoring a more natural and idiomatic English expression. This comparison reveals that the target's literal translation is actually forced and unidiomatic in the given context, leading to a corrected lower score (Score: 4).