On the Systematic Challenges of Culturally Loaded Machine Translation: *Dream of the Red Chamber* as the Cultural Lens

Yiming Wang*
School of Computer Science,
Shanghai Jiao Tong University
yiming.wang@sjtu.edu.cn

Jiayuan Di*
School of Foreign Languages,
East China University of Science and Technology
mira.jiayuan@gmail.com

Abstract

Culturally loaded translation poses unique challenges for machine translation (MT), as meanings are deeply embedded in socio-cultural contexts beyond surface linguistic forms. Although large language models (LLMs) have enabled MT systems to achieve human-like quality in many scenarios, their ability to handle culturally loaded expressions remains underexplored. In this study, we systematically investigate the challenges posed by culturally loaded translation in LLM-based MT systems. We construct a Chinese-Japanese bilingual dataset from the culturally representative corpus *Dream of the Red Chamber*[1], containing 500 segments across diverse cultural categories. Using a comprehensive evaluation protocol, we reveal three main challenges: (1) *task challenges*, where frontier LLMs exhibit notable performance gaps and struggle with culturally loaded content; (2) *human evaluation challenges*, where evaluator backgrounds lead to substantial disagreement in translation judgments; and (3) *automatic evaluation challenges*, where widely used metrics fail to reliably assess translation quality for this task. These findings may offer valuable insights for culture-oriented translation research in both computational science and linguistics.

1 Introduction

Translation aims to convey the source message in the target language through the closest natural equivalent while preserving semantic, stylistic, and pragmatic effects. However, language is inseparable from culture, and cultural priorities shape linguistic expression (Nida, 1964). Each culture organizes its vocabulary around its own areas of emphasis, leading certain domains of meaning to become increasingly detailed and complex (Larson, 1998). This culturally shaped development gives rise to a category of lexical items termed *culturally loaded words* by Xu (1980). Such words are deeply embedded in specific socio-cultural contexts and reflect the traditions, beliefs, and value systems of a community (Baker, 1992; Newmark, 1998).

The translation of culturally loaded expressions is **highly contextualized**. Its context extends beyond the immediate text to the broader cultural background behind corresponding languages. Differences in cultural focus across languages may lead to semantic mismatches or lexical gaps (Newmark, 1988; Bader, 1994). Therefore, translators must consider both source and target cultural contexts and balance interpretation with communication. *At the interpretive level*, they should pursue conceptual equivalence rather than rely on surface correspondence, ensuring that target readers grasp the intended meaning, since literal translation is often insufficient (Nida, 1964). *At the communicative level*, these expressions carry substantial cultural weight and should be rendered in ways that preserve, as far as possible, meanings rooted in the source culture (Venuti, 1995). These place high demands on the professional expertise of translators.

On the computer science side, the rapid development of large language models (LLMs) (Brown et al., 2020; Touvron et al., 2023; Bai et al., 2023) has enabled machine translation (MT) systems to achieve human-like quality in many scenarios (Vilar et al., 2023; Jiao et al., 2023; He et al., 2024). However, capturing rich contextual information remains a major challenge for current MT systems (Vieira et al., 2021; Moghe et al., 2025; Zhang et al., 2025; Shen et al., 2025), making the culturally loaded translation task particularly difficult for current models due to its inherently contextualized nature. Evaluating such translations adds another layer of difficulty. Translation quality is

*Equal Contribution. Data are available at https://github.com/mira-jiayuan/RedChamber-ZHJA.

[1]The Chinese novel 《红楼梦》, also known as 《石头记》, was first translated as *Dream of the Red Chamber* by Wang Jizhen (1929). David Hawkes later adopted the alternative title *The Story of the Stone* in his influential translation (1973–1986). Following broader academic and general convention, we use the former title to refer to the original work.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 1292–1314
October 24-29, 2026 ©2026 Association for Computational Linguistics

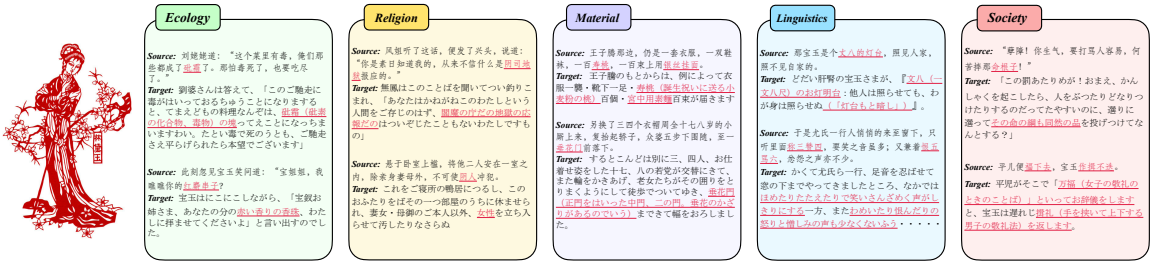


Figure 1: Dataset examples from our *Dream of the Red Chamber* corpus across the five cultural categories. Detailed category definitions are in §2.2, and more examples are provided in §A.1.

already subjective, and cultural elements amplify this subjectivity: evaluators from different cultural backgrounds may judge the same output differently based on their cultural familiarity and preferences (see §2.3 for the theoretical basis), introducing unpredictable variance in human evaluation. Furthermore, it remains unclear whether widely used automatic metrics can reliably assess translation quality in these scenarios. Thus, from the task itself to its evaluation, culturally loaded translation presents unique challenges for modern MT research.

In this study, we systematically investigate the above challenges that it poses for LLM-based MT. We construct a Chinese-Japanese bilingual dataset (§3) based on the highly representative cultural corpus of *Dream of the Red Chamber*. The dataset comprises 500 representative culturally loaded segments, and spans diverse cultural categories, with representative examples shown in Fig.1. We also design a specific evaluation protocol (§4) that incorporates diverse evaluation dimensions and evaluator backgrounds, ensuring the completeness and comprehensiveness of the results. All preliminary model translation qualities are based on human evaluation, which serves as the gold standard for our formal analysis. Based on these, we reveal three comprehensive core challenges:

- **Task Difficulty (§5.1):** Frontier LLMs still underperform on culturally loaded translation, showing substantial gaps both across models and relative to human references. Their performance also varies across cultural categories and is sensitive to category-specific textual characteristics and difficulty. This indicates considerable room for improvement in LLM-based MT systems.
- **Human Evaluation Disagreement (§5.2):** Human evaluation of culturally loaded translation is sensitive to evaluator background, including native cultural environment and expertise. This indicates that evaluating this task requires careful design of evaluator diversity and comprehensiveness; otherwise, systematic bias may arise.
- **Automatic Evaluation Unreliability (§5.3):** Mainstream automatic metrics struggle to reliably capture model rankings and totally fail to distinguish sample-level quality differences in culturally loaded translation. This highlights the need for more task-aware automatic methods.

In addition, we conduct extended error and translation strategy analyses to better understand model behavior patterns and key bottlenecks in this task. We hope these findings provide valuable insights for both linguistic and computational research.

2 Theoretical Foundations

2.1 Culturally Loaded Translation

Xu (1980) first coined the term *culturally loaded words* to describe lexical items that encode society's traditions, beliefs, and value systems. Similar concepts in translation studies, such as *cultural words* (Newmark, 1988) and *culture-specific items* (Aixelá, 1996), generally refer to expressions that are deeply rooted in particular socio-cultural contexts and whose meanings cannot be fully understood without cultural background knowledge (Baker, 1992; Newmark, 1998).

Taking Chinese culture as examples, “布衣” (*bùyī*) in ancient China referred to ordinary people and implied a modest lifestyle associated with coarse cloth garments. If translated literally as “coarse clothes,” English readers may find it confusing, since the phrase only refers to a type of fabric and does not convey the social meaning in Chinese. Another example is “鸿雁” (*hóngyàn*), a bird image in Chinese ecological culture. Although “swan goose” is its literal zoological equivalent, it fails to capture the cultural meaning. In Chinese tradition, “鸿雁” often symbolizes a messenger carrying letters. Translating it as “message-bearing swans”[2] better conveys this cultural imagery, even

[2]Translated by Chinese translator Yuanchong Xu.

though it sacrifices strict lexical equivalence. These examples show that translating culturally loaded expressions requires conveying both meaning and cultural significance.

2.2 Cultural Category

As the foundational scholar of modern translation studies, Eugene A. Nida pioneered the categorization of culture into five types: ecology, religion, material, linguistics, and society (Nida, 1945, 1964):

- **Ecology**: Natural elements such as flora, fauna, climate, geographical landscapes, and ecological phenomena that carry cultural meaning.
- **Religion**: Spiritual beliefs including ancestor worship, folk superstitions, and mythological concepts that shape worldviews.
- **Material**: Tangible artifacts of daily life like clothing, food, tools, and other physical objects that define a civilization's material culture.
- **Linguistics**: Culturally embedded expressions like idioms, proverbs, slang, riddles, and fixed phrases that exhibit unique rhetorical patterns.
- **Society**: Social structures, customs, institutions, kinship systems, etiquette norms, and conventions that govern interpersonal relations.

This classification has shown strong theoretical adaptability and continues to inform contemporary translation research (Oo, 2020; Pian and Chen, 2022; Prompan, 2024; Al Rujaibi et al., 2025). In this study, we adopt this five-fold taxonomy as the framework for dataset construction.

2.3 Cross-culture Translation Theory

Cross-cultural translation has long been a central topic across disciplines. Among the most influential frameworks is Venuti's domestication and foreignization taxonomy (Venuti, 1995). Domestication adapts a text to target-culture norms to enhance readability, whereas foreignization preserves source-culture elements to foreground cultural difference. This framework highlights the translator's agency and offers a critical perspective for evaluating cross-cultural transfer. It has been widely used to analyze how translators handle culture-specific items by balancing accessibility and authenticity.

Other theoretical perspectives further deepen the understanding of translation as a culturally embedded practice. The dynamic equivalence theory (Nida, 1964) prioritizes equivalent reader response rather than literal correspondence. The cultural turn thoery (Bassnett and Lefevere, 1990) conceptualizes translation as a form of cultural rewriting shaped by ideological forces; The norms theory (Toury, 1995) emphasizes the socio-cultural constraints that guide translators' decisions. Together, these frameworks suggest that translation is not a neutral linguistic transfer but a process of cultural negotiation — an important foundation for examining how language models handle culturally loaded translation, which we will also explore in our subsequent experiments.

3 Corpus

3.1 Data Source

We used *Dream of the Red Chamber* as our source corpus. First, widely regarded as the "encyclopedia of Chinese classical culture" (Edwards, 2005; Plaks, 2015), it displays exceptional linguistic richness and cultural depth. Second, it covers all five cultural categories in §2.2, enabling comprehensive evaluation across cultural dimensions within a unified textual framework (Xia, 2010). Third, its long translation history and the extensive scholarship surrounding it, often referred to as "redology" (Saussy, 1995; Moratto et al., 2022), allow findings derived from this corpus to connect with broader research and cross-cultural communication.

3.2 Target Language

Building on Chinese culture as the source, we select Japanese as the target language, as its cultural interpretive distance (Bassnett and Lefevere, 1990; Toury, 1995) from Chinese strikes a balance. Unlike translations between Chinese and Western languages, which often require extensive cultural explanation, Chinese and Japanese belong to the East Asian cultural sphere and share elements such as the kanji writing system and related aesthetic traditions. This shared foundation allows many culturally loaded concepts to be transferred more directly, while still requiring careful strategic choices when meanings diverge. For example, the Chinese term "阴司" (*yīnsī*) is rendered in the Japanese translation as 「閻魔の庁」. This substitution draws on shared religious conceptions between Chinese and Japanese cultures, where both traditions envision judgment by King Yama after death. In contrast, Western readers' imagination of the "nether world" tends to evoke Greek Hades or the Christian purgatory, lacking the distinct East Asian imagery of Yama's judgment. This intermediate cultural dis-

Table 1: Mean lengths and standard deviations for source and target texts per cultural category.

	Ecology	Religion	Material	Language	Society
Source	$23.40_{\pm 12.35}$	$31.70_{\pm 16.21}$	$24.20_{\pm 16.69}$	$15.20_{\pm 9.42}$	$24.90_{\pm 24.19}$
Target	$50.00_{\pm 25.43}$	$64.30_{\pm 36.92}$	$56.50_{\pm 44.72}$	$32.05_{\pm 19.42}$	$51.05_{\pm 43.33}$

tance makes the Chinese–Japanese pair particularly revealing for studying culturally loaded translation.

3.3 Translation Version

We prioritize target-native translations because conveying culturally loaded meanings in the target language requires generative competence in its linguistic and cultural norms (Nida, 1993; Bassnett, 2007). Mastery of the target culture is thus more consequential than source-culture familiarity, leading us to adopt a translation by a native Japanese speaker.

Japanese translations of *Dream of the Red Chamber* date back to Mori Kainan in 1892. Over the following century, they evolved from partial renderings into at least 38 full versions. Among these, the full translations by Matsueda Shigeo, Ito Sohei, and Inami Ryoichi are widely regarded as the three pillars of the Japanese tradition (Song, 2019). We adopt Ito Sohei's version for this study. Unlike Inami Ryoichi's modern Japanese rendering, which prioritizes reader accessibility, Ito's translation achieves a more balanced approach by preserving source fidelity while providing rich annotations that deepen cultural interpretation (Wu, 2019). Furthermore, developed over nearly fifty years with four revisions, it has become the most influential and widely studied version in academic research (Hu, 1993; Maruyama, 1998; Sun, 2006).

3.4 Collection and Annotation

We collected the bilingual Chinese-Japanese edition from the state-sponsored *Library of Chinese Classics* series[3], translated by Ito Sohei. This edition was compiled under the auspices of Chinese government agencies and extensively revised by renowned scholars, ensuring its authority and reliability. Also, its format enables direct comparisons between source and target texts (see §A.2), ensuring rigor throughout the data processing.

Currently, only image versions of this edition are available, so OCR processing is required. Four Chinese graduate students conducted the initial annotation, each selecting 480 culturally loaded segments (4 per chapter across all 120 chapters) and categorizing them according to the taxonomy in §2.2.

[3]https://zh.wikipedia.org/wiki/大中华文库

Table 2: Language models used in our experiments.

Model Name	Abbreviation	Affiliation	Citation
Non-Reasoning Model			
Deepseek-v3	*DS-v3*	Deepseek	DeepSeek-AI (2024)
Qwen3-235B-A22-Non-Thinking	*Qwen3-T*	Alibaba	Yang et al. (2025)
GPT-4.1	*GPT-4.1*	OpenAI	OpenAI (2025a)
Gemini-2.5-Flash	*Gemini-2.5*	Google	Comanici et al. (2025)
Reasoning Model			
Deepseek-r1	*DS-r1*	Deepseek	Guo et al. (2025)
Qwen3-235B-A22-Thinking	*Qwen3-NT*	Alibaba	Yang et al. (2025)
OpenAI-o4-mini	*o4-mini*	OpenAI	OpenAI (2025b)
Claude-Sonnet-4	*Claude-4*	Anthropic	Anthropic (2025)

Table 3: Evaluation dimensions used in our experiments.

Dimension Name	Abbreviation	Target	Scope
Content Accuracy	*Acc.*	Content	General
Language Fluency	*Flu.*	Language	General
Cultural Appropriateness	*Cult.*	Content	Culture-specific
Native Readability	*Read.*	Language	Culture-specific

Four Japanese graduate students then performed a second round of screening and revision. To control the workload of subsequent human evaluation, the dataset was further refined to 500 representative segments, balanced across the five cultural categories (100 per category). The entire annotation process took 20 days. These annotators also participated in human evaluations, so their backgrounds are presented in subsequent §4.1.

Text length statistics are shown in Tab.1 using the Qwen3 tokenizer (Yang et al., 2025). Overall, target texts are about twice as long as source texts. Among categories, *Religion* tends to be longer, *Linguistics* shorter, and the other three are similar. Importantly, there are no marked differences, meaning that category difficulty is not driven by text length.

4 Evaluation Protocol

We evaluated eight language models (see Tab.2), selected for two aspects of diversity: (1) they span five organizations and multiple model families, providing broad source coverage; (2) they include both reasoning and non-reasoning paradigms, enabling us to examine the impact of long chain-of-thought (CoT) ability (Wei et al., 2022; Guo et al., 2025) on culturally loaded translation. Implementation details are provided in §B.1. Based on model outputs, we established the following evaluation protocol.

4.1 Human Evaluation

Given the contextual complexity and inherent subjectivity of this task, our evaluation is human-centered, with human judgments serving as the gold standard. We adopt the standard MQM framework (Lommel et al., 2013), a widely used func-

Table 4: Human evaluation results for eight models and one reference across five cultural categories and the full dataset. For each category, we report overall quality scores (*Overall*) along with four evaluation dimensions. 🏆, 🥉, and 🎗 mark the top three performing models in each dimension.

	Full Dataset					Ecology					Religion				
Model Name	*Overall*	*Acc.*	*Flu.*	*Cult.*	*Read.*	Overall	*Acc.*	*Flu.*	*Cult.*	*Read.*	Overall	*Acc.*	*Flu.*	*Cult.*	*Read.*
Reference	*4.27*	*4.35*	*4.24*	*4.28*	*4.11*	*4.33*	*4.38*	*4.27*	*4.32*	*4.16*	*4.23*	*4.33*	*4.29*	*4.24*	*4.09*
Deepseek-v3	🎗 **3.83**	3.92	🥉 4.05	🏆 **3.83**	3.67	🥉 4.06	🥉 4.11	🥉 4.24	🏆 **4.05**	🎗 **3.89**	🥉 3.88	3.92	🥉 4.15	🏆 **3.83**	3.67
Qwen3-235B-A22B-Non-Thinking	3.30	3.60	3.46	3.37	3.40	3.49	3.79	3.66	3.59	3.56	3.33	3.71	3.54	3.42	3.49
GPT-4.1	3.66	3.75	3.72	3.63	🏆 **3.94**	3.82	3.89	3.85	3.71	🏆 **4.09**	3.64	3.84	3.71	🎗 **3.61**	🏆 **3.91**
Gemini-2.5-Flash	🥉 3.87	🥉 3.94	🎗 **3.91**	🥉 3.82	🥉 3.90	🎗 **4.03**	🎗 **4.10**	🎗 **4.00**	🥉 3.90	🥉 4.05	🎗 **3.84**	🥉 3.99	🎗 **3.89**	🥉 3.74	🥉 3.88
Deepseek-r1	3.37	3.51	3.53	3.33	3.48	3.51	3.60	3.61	3.38	3.60	3.34	3.56	3.49	3.25	3.47
Qwen3-235B-A22-Thinking	3.65	🎗 **3.93**	3.76	3.63	3.31	3.85	🥉 4.11	3.88	3.81	3.50	3.65	🎗 **3.96**	3.81	🎗 **3.61**	3.38
OpenAI-o4-mini	🏆 **3.89**	🏆 **4.00**	🏆 **4.17**	🎗 **3.74**	🎗 **3.72**	🏆 **4.09**	🏆 **4.20**	🏆 **4.33**	🎗 **3.89**	🎗 **3.89**	🏆 **3.90**	🏆 **4.06**	🏆 **4.23**	🥉 3.74	🎗 **3.78**
Claude-Sonnet-4	3.57	3.78	3.80	3.64	3.42	3.73	3.91	3.92	3.72	3.58	3.54	3.85	3.78	🎗 **3.61**	3.39
Average	**3.64**	**3.80**	**3.80**	**3.62**	**3.61**	**3.82**	**3.96**	**3.94**	**3.76**	**3.77**	**3.64**	**3.86**	**3.83**	**3.60**	**3.62**
Range (max-min)	**0.59**	**0.49**	**0.71**	**0.50**	**0.63**	**0.60**	**0.60**	**0.72**	**0.67**	**0.59**	**0.57**	**0.50**	**0.74**	**0.58**	**0.53**

	Material					Linguistics					Society				
Model Name	*Overall*	*Acc.*	*Flu.*	*Cult.*	*Read.*	Overall	*Acc.*	*Flu.*	*Cult.*	*Read.*	Overall	*Acc.*	*Flu.*	*Cult.*	*Read.*
Reference	*4.39*	*4.39*	*4.33*	*4.34*	*4.17*	*4.17*	*4.29*	*4.16*	*4.22*	*4.05*	*4.23*	*4.36*	*4.15*	*4.28*	*4.08*
Deepseek-v3	🥉 4.08	4.12	🥉 4.26	🏆 **4.09**	3.88	🎗 **3.59**	3.73	🥉 3.86	🥉 3.61	3.51	🎗 **3.54**	🎗 **3.68**	🥉 3.75	🥉 3.58	3.42
Qwen3-235B-A22B-Non-Thinking	3.56	3.84	3.72	3.67	3.64	3.10	3.45	3.33	3.20	3.28	3.00	3.35	3.15	3.10	3.14
GPT-4.1	3.85	3.93	3.86	3.72	🏆 **4.13**	3.53	3.66	3.65	3.52	🏆 **3.83**	3.45	3.58	3.50	3.45	🏆 **3.76**
Gemini-2.5-Flash	🥉 4.08	🎗 **4.15**	🎗 **4.04**	🎗 **3.94**	🥉 4.10	🏆 **3.71**	🥉 3.79	🎗 **3.81**	🏆 **3.67**	🥉 3.74	🏆 **3.68**	🏆 **3.76**	🎗 **3.71**	🏆 **3.65**	🥉 3.72
Deepseek-r1	3.59	3.68	3.69	3.46	3.70	3.22	3.38	3.43	3.20	3.35	3.18	3.34	3.33	3.18	3.33
Qwen3-235B-A22-Thinking	🎗 **3.90**	🥉 4.17	3.92	3.85	3.56	3.47	🎗 **3.75**	3.64	3.49	3.21	3.37	3.65	3.47	3.40	3.10
OpenAI-o4-mini	🏆 **4.13**	🏆 **4.19**	🏆 **4.40**	🥉 3.96	🎗 **3.97**	🥉 3.69	🏆 **3.81**	🏆 **4.00**	🎗 **3.58**	🎗 **3.57**	🥉 3.62	🥉 3.74	🏆 **3.88**	🎗 **3.52**	🎗 **3.49**
Claude-Sonnet-4	3.79	3.98	3.97	3.77	3.64	3.43	3.66	3.71	3.51	3.32	3.38	3.61	3.60	3.48	3.27
Average	**3.87**	**4.01**	**3.98**	**3.81**	**3.83**	**3.47**	**3.65**	**3.68**	**3.47**	**3.48**	**3.40**	**3.59**	**3.55**	**3.42**	**3.40**
Range (max-min)	**0.57**	**0.51**	**0.71**	**0.63**	**0.57**	**0.61**	**0.43**	**0.67**	**0.47**	**0.62**	**0.68**	**0.42**	**0.73**	**0.55**	**0.66**

tionalist evaluation approach[4], and adapt it to better capture the characteristics of culturally loaded translation. Specifically, we define four evaluation dimensions, shown in Tab.3 (see §C for detailed explanations and scoring guidelines). The first two assess semantic accuracy and linguistic fluency following MQM. The remaining two focus on cultural aspects, evaluating whether the translation conveys the cultural meaning of the source expression and whether it reads naturally to target-culture readers. In addition, evaluators assign an *overall quality score* to reflect their holistic judgment.

Our evaluators come from diverse cultural and academic backgrounds and consist of sixteen native speakers, including eight Chinese and eight Japanese. Within each language group, four are graduate students specializing in the other language's literature, and four are senior academics serving as university professors in the other country's cultural studies. To ensure evaluation reliability, all student evaluators hold at least JLPT-N1 or HSK-6 certification, or an equivalent qualification. This design yields four groups by language (*zh* / *ja*) and academic level (*stu.* / *prof.*).

Given the data scale, evaluating all translations would be prohibitively labor-intensive: (8 models + 1 reference) × 500 samples = 4,500 translations, resulting in 22,500 scores per evaluator. Therefore, within each group, four evaluators each review 125 samples (25 per category), collectively covering the full dataset. For *Read.*, scores are assigned only by target-language natives. To avoid bias from prior assumptions about model and human abilities, translation sources were anonymized. Evaluators saw only the source text and nine shuffled translations for each sample. Before the evaluation, pilot assessments verified intra-group agreement. Sampling-based comparisons showed that evaluators within the same group had comparable expertise and low judgment variance. After the evaluation, we conducted follow-up interviews with the evaluators to examine their fine-grained decision criteria and interpretations of some anomalous phenomena, which are needed for analysis in §5.2.

4.2 Automatic Evaluation

We adopt three types of metrics: lexical-based BLEU (Papineni et al., 2002), semantic-based xCOMET (Guerreiro et al., 2024), and LLM-as-a-Judge (Zheng et al., 2023). For the latter, we employ two models: GPT-5 (a non-reasoning model, OpenAI (2025c)) and Gemini-3-Pro (a reasoning model, Google (2025)). See §B.2 for the specific prompt.

5 Results

5.1 Task Difficulty (Tab.4)

Model Performance Comparison. We first analyze overall model performances on the full dataset. The top three models in overall score are *OpenAI-o4-mini* (3.89), *Gemini-2.5-Flash* (3.87), and *Deepseek-v3* (3.83). Notably, the first two rank among the top three across all five evaluation dimensions, indicating strong robustness. However,

[4]The standard MQM framework includes five dimensions: Accuracy, Fluency, Verity, Design, and Internationalization.

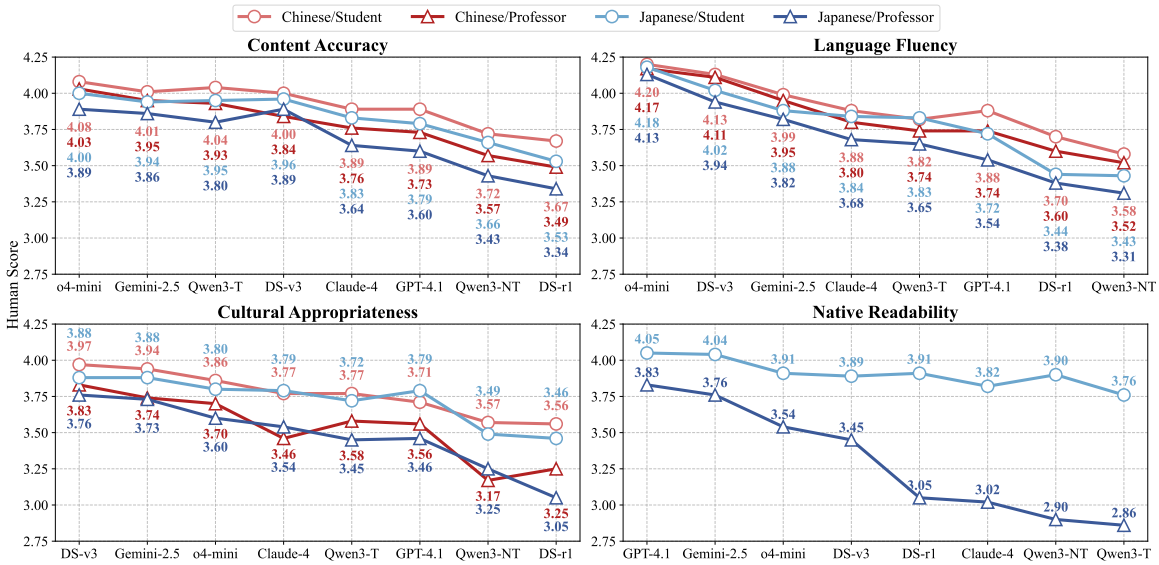


Figure 2: Human evaluation results for eight models across four dimensions, scored by four evaluator groups: Chinese/Student (*zh-stu.*), Chinese/Professor (*zh-prof.*), Japanese/Student (*ja-stu.*), Japanese/Professor (*ja-prof.*). In each subfigure, *models are arranged from left to right in descending order of their overall scores on the full dataset*.

performances vary greatly across models, with overall scores ranging from 3.89 to 3.30. This wide range demonstrates that this task poses a meaningful challenge and effectively differentiates model capabilities. In addition, a considerable gap remains between current models (an average of 3.64) and human references (4.27), whose scores are consistently higher across all dimensions, indicating that culturally loaded translation remains difficult for current LLMs to match human performance.

Another interesting phenomenon is that reasoning and non-reasoning models reveal no systematic advantage for either group. Their average overall scores are nearly identical (3.66 vs. 3.62), and medal counts are similarly balanced (15 vs. 12 medals, with 2 vs. 3 golds). While the top-performing *OpenAI-o4-mini* employs reasoning, the runner-ups *Gemini-2.5-Flash* and *Deepseek-v3* do not, suggesting that longer CoT are not decisive for our culture-oriented translation tasks.

Dimension Difference. Next, we analyze model performance across evaluation dimensions. A clear pattern emerges: the two culturally related dimensions, *Cult.* and *Read.*, consistently receive lower scores than the general two, *Acc.* and *Flu.*. This suggests that while current models perform relatively well on general quality, they struggle more with culturally loaded aspects of translation. Moreover, strong performance on general dimensions does not necessarily imply strong cultural competence. For example, *OpenAI-o4-mini* achieves the highest scores in *Acc.*(4.00) and *Flu.*(4.17), yet its *Cult.*(3.74) and *Read.*(3.72) only rank third. Conversely, *GPT-4.1* attains the highest *Read.* (3.94) but performs less strongly in general dimensions. These phenomena indicate that cultural competence cannot be inferred directly from general translation quality, revealing a gap between linguistic correctness and deeper cultural adaptation.

Cultural Category Difference. Finally, we analyze model performance across cultural categories. The results reveal a clear difficulty hierarchy: *Material* (3.87) and *Ecology* (3.82) score the highest, followed by *Religion* (3.64), while *Linguistics* (3.47) and *Society* (3.40) lag significantly behind.

This gap is largely driven by differences in textual characteristics across cultural categories. *Material* and *Ecology* texts mainly describe concrete objects, entities, and natural phenomena, which are relatively straightforward for models to translate. *Religion* texts, although conceptually abstract, often contain narrative elements and real-world descriptions that models can partially handle using general knowledge. In contrast, *Linguistics* and *Society* texts are considerably more challenging. *Linguistics* frequently includes classical idioms, allusions, and elliptical expressions with dense meanings and flexible syntax, making accurate interpretation and translation difficult. *Social Culture* involves etiquette and honorific expressions that depend on subtle social relationships, requiring pragmatic inference that models often fail to capture, leading to inappropriate tone or meaning.

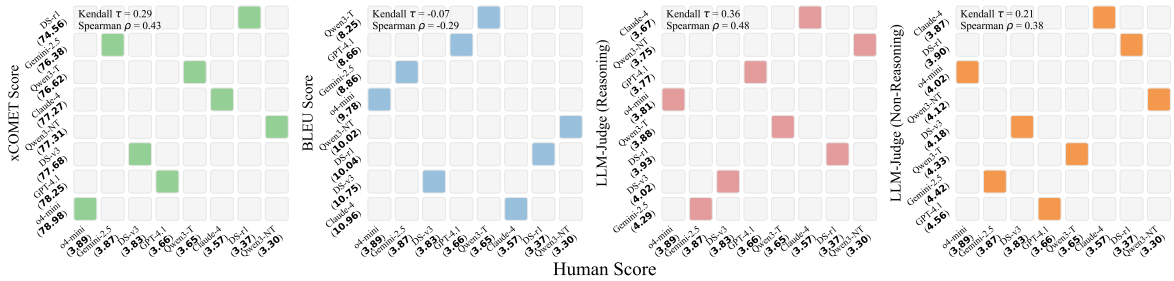


Figure 3: System-level correlation between human overall scores and automatic metrics. In each subplot, models on the x-axis are ordered by human scores from high to low, and models on the y-axis are ordered by the metric from low to high. Colored cells mark matching models, and Kendall's τ and Spearman's ρ are shown in the top-left.

5.2 Human Evaluation Disagreement (Fig.2)

Overall Inter-group Comparison. The inter-group agreement varies across evaluation dimensions. For the first three dimensions, the differences among evaluator groups remain moderate, generally below 0.5 points. In contrast, the *Read.* dimension shows substantially larger divergence. For several lower-ranked models, the gap between evaluator groups approaches 1 point. We analyze this anomaly in detail below.

Despite these variations, the relative ranking of the eight models remains largely consistent across evaluator groups and aligns well with the overall scores. Among the four groups, *ja-prof.* shows the closest agreement with the overall ranking, followed by *zh-prof.*. The two student groups exhibit slightly greater fluctuation, though the differences are not substantial. In addition, agreement tends to decrease as model quality declines, as poorer outputs leave more room for subjective interpretation.

Student vs. Professor. Within the same language group (*zh* or *ja*), students consistently score higher than professors. This may reflect differences in evaluation standards and familiarity with culturally loaded texts. Through the post-evaluation interview, we find that when encountering ambiguous expressions, students tend to accept them more readily, which introduces an upward bias. By contrast, professors generally apply stricter criteria when judging translation adequacy.

This divergence is most evident in the *Read.* dimension, where students and professors show substantial disagreement. Through the post-evaluation interview, we note an important fact: culturally loaded texts are often associated with specific historical or literary registers. Students tend to evaluate readability primarily based on fluency and ease of comprehension. Professors, however, pay closer attention to whether the stylistic register (e.g. classical or modern language) appropriately reflects the cultural and historical context embedded in the source text, even if this reduces immediate readability. In practice, models often produce relatively colloquial outputs that align with students' reading habits, leading them to score such translations more favorably. Professors, however, may consider these stylistically inappropriate, arguing that culturally significant texts should preserve more register-appropriate expressions. This reflects a broader tension between accessibility and authenticity rather than a purely objective notion of correctness.

This is further evidenced by their scoring of the reference on the *Read.* dimension: students and professors gave average scores of 3.70 and 4.52, respectively. Notably, the students' rating is even lower than that of some model outputs, providing additional support for the above conclusions.

Chinese (Source) vs. Japanese (Target). Within the same academic group (*stu.* or *prof.*), Japanese are slightly stricter than Chinese evaluators on the two general dimensions, *Acc.* and *Flu.*. This is expected, as Japanese annotators are native speakers of the target language and can therefore apply more nuanced judgments of translation quality.

In contrast, the relative scoring patterns for the cultural dimension *Cult.* vary across models — sometimes Japanese evaluators score higher, while in other cases Chinese evaluators do. Through the post-evaluation interview, we find that although both groups value the preservation of source culture and the appropriateness of the target expression, they tend to emphasize different aspects. Chinese evaluators often focus more on faithfulness to the source culture, whereas Japanese evaluators place greater weight on appropriateness within the target cultural context. Because current models often struggle to balance these two aspects simultaneously, translations that favor one side may draw stricter scrutiny from evaluators who prioritize the other, causing noticeable fluctuations in scoring.

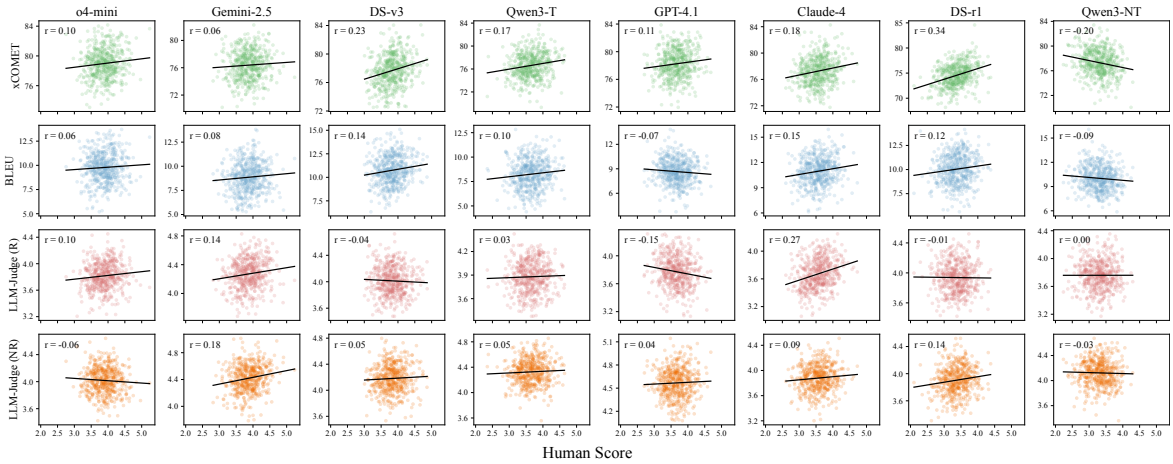


Figure 4: Sample-level correlation between human overall scores and automatic metrics. Each subplot corresponds one model-metric pair with all 500 samples. The black line is the linear fit, and Pearson's r are shown in the top-left.

5.3 Automatic Evaluation Unreliability (Fig.3-4)

System Correlation. Fig.3 compares system-level rankings between human evaluation and automatic metrics. Overall, correlations with human judgments remain limited. xCOMET and two LLM-based judges show modest positive correlations (Kendall's τ between 0.21 and 0.36, Spearman's ρ between 0.38 and 0.48), with several models positioned near the diagonal. However, BLEU exhibits almost no correlation ($\tau < 0.1$), as its reliance on surface-level lexical overlap with reference translations prevents it from capturing semantic and culturally grounded translation quality. These results suggest that while newer metrics offer somewhat more informative signals than BLEU, they still fall short of reliably distinguishing model performance in culturally loaded translation tasks.

Sample Correlation. Fig.4 illustrates the correlation between human evaluation scores and four automatic metrics across all eight models. Overall, correlations with human judgments remain extremely weak at the sample level for all metrics. Although xCOMET and the LLM judges occasionally show slightly stronger positive correlations than BLEU, the overall alignment remains limited, and the scatter plots display substantial dispersion around the fitted regression lines. In many cases, samples with similar automatic scores receive noticeably different human ratings, while samples with comparable human judgments correspond to a wide range of automatic metric values. This pattern suggests that automatic metrics capture only part of the translation quality signal and fail to reliably reflect fine-grained differences among individual outputs. Consequently, the magnitude of an automatic metric score alone is insufficient to determine whether a specific generated sample is actually better or worse according to human evaluation.

6 Extend Analysis

To further explore the model limitations in translating culturally loaded content, we conduct extended analyses of their translation behavior. Due to space limitations, detailed discussions are provided in the appendix, with key conclusions presented here.

Error Type (§D). We classify translation errors into five types: *No Obvious Error*, *Mistranslation*, *Overtranslation*, *Undertranslation*, and *Omission*. *Mistranslation* is the most frequent error across all models, while higher-performing models exhibit a lower proportion of such errors. The remaining three error types occur relatively infrequently. Notably, reasoning models produce substantially more *Overtranslation*, yet this tendency does not translate into better overall translation quality.

Translation Strategy (§E). We focus on the domestication-foreignization framework introduced in §2.3. We find that over 80% of translations adopt a single strategy, with a clear preference for domestication. Under this, models tend to rigidly adapt cultural expressions, often sacrificing source cultural imagery for target-language readability, resulting in notable loss of cultural transmission.

Case Study (§F). To provide a more intuitive understanding of model translation styles, we present several case studies that illustrate how models underperform across various evaluation dimensions.

7 Related Work

Regarding related work on culturally loaded translation, §2 has already discussed foundational theoretical studies. Here, we focus only on MT research that potentially relates to our work.

MT research has long recognized the challenges of culture-oriented translation (Hershcovich et al., 2022). In response, several studies have proposed benchmarks to quantify cultural translation (Riley et al., 2023; Cao et al., 2024; Yao et al., 2024; Zhang et al., 2024), but these often focus on specific domains or concrete contexts (e.g., food) and rely primarily on automatic metrics. Other work has developed systems for identifying culturally loaded content (Pandey et al., 2025). Shen et al. (2025) shares a similar perspective with ours, namely rethinking the challenges of culturally loaded translation in MT, but focuses specifically on the aspect of paratexts. In contrast, our study adopts a more systematic perspective, offering a comprehensive analysis of culturally loaded translation challenges from the task itself to its evaluation.

8 Conclusion

Culturally loaded translation has long been a central concern in translation studies, as language serves not only as a means of communication but also as a carrier of culture. In this work, we connect this linguistic topic with modern MT research. Using a dataset based on Chinese-Japanese translations of *Dream of the Red Chamber*, we systematically uncover the challenges of culturally loaded translation from three perspectives: the task itself, human evaluation, and automatic evaluation. Our analysis may offer insights for future research on culturally loaded translation in computer science. More broadly, the results suggest that, despite rapid progress in LLMs, MT still has a long way to go, as many researchers have long believed (Kocmi et al., 2024; Pang et al., 2025; Ataman et al., 2025).

Limitations

The primary limitation of this study is language diversity. Although we have justified in §3.2 the representativeness of the current language pair for studying culturally loaded translation, including additional languages, particularly those with broader speaker populations such as English, would undoubtedly enhance the generalizability of our findings. However, due to various constraints, including limited funding, the lengthy cycle of human evaluation, and the need for culturally proficient native speakers (crowdsourcing is not a viable option for this study), we are currently unable to extend our study to more languages in the near term. As a mitigation, we will open-source our data sources and provide a raw multilingual corpus to facilitate further research. We believe that future work can build upon this foundation to yield more valuable insights.

Ethical Considerations

Potential Risk. Our study involves engineering experiments and analyses, and does not propose any new architectures or algorithms. Therefore, it does not introduce uncontrollable consequences in practical use. Regarding the research data, we have carefully verified the compliance of both the constructed dataset and the translations generated by LLMs, ensuring no potential risks are involved.

Data Compliance. We have confirmed that our research data, including the constructed dataset and the translations generated by LLMs, does not contain any sensitive elements related to politics, violence, or similar topics. We ensure that no uncontrollable impact will be caused to the human evaluators involved in the experiment or to any individuals who may come into contact with the data.

Human Evaluation. All evaluators were fully informed about the research use of the evaluation data collected from them, and their consent was obtained. Throughout the evaluation process, we consistently respected their autonomy and paid them with mutually agreed-upon remuneration.

AI Assistants Declaration

AI assistants (ChatGPT and Deepseek) were used exclusively for writing polishing and played no role in any other aspect of this work.

Acknowledgments

This work was conducted entirely by student researchers, without external funding or financial support from senior researchers. We thank Natsuki Oe, Kanon Yamaguchi, and Mouye Weng for coordinating the human study with Japanese volunteers, and Yuya Goto of Mejiro University for his helpful suggestions on parts of the background discussion and case studies presented in this manuscript.

References

Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, and 1 others. 2025. Phi-4-Reasoning technical report. *arXiv preprint arXiv:2504.21318*.

Javier Franco Aixelá. 1996. Culture-specific items in translation. In Román Álvarez and M. Carmen-África Vidal, editors, *Translation, Power, Subversion*, pages 52–78. Multilingual Matters, Clevedon.

Raya Nasser Al Rujaibi, Adlina Ariffin, and Rafik Jamoussi. 2025. Cultural representation in children's cartoon programmes: Insights from the Nida/Newmark typology. *International Journal of Language, Literacy and Translation*, 8(2):28–44.

Anthropic. 2025. Introducing Claude 4.

Duygu Ataman, Alexandra Birch, Nizar Habash, Marcello Federico, Philipp Koehn, and Kyunghyun Cho. 2025. Machine translation in the era of large language models: A survey of historical and emerging problems. *Information*, 16(9):723.

Yousef Bader. 1994. Mona baker, in other words: A coursebook on translation (book review). *IRAL: International Review of Applied Linguistics in Language Teaching*, 32(2):169–170.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Mona Baker. 1992. *In Other Words: A Coursebook on Translation*. Routledge, London and New York.

Susan Bassnett. 2007. Culture and translation. In Piotr Kuhiwczak and Karin Littau, editors, *A Companion to Translation Studies*, pages 13–23. Multilingual Matters, Clevedon.

Susan Bassnett and André Lefevere, editors. 1990. *Translation, History and Culture*. Pinter Publishers, London and New York.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Yong Cao, Yova Kementchedjhieva, Ruixiang Cui, Antonia Karamolegkou, Li Zhou, Megan Dare, Lucia Donatelli, and Daniel Hershcovich. 2024. Cultural adaptation of recipes. *Transactions of the Association for Computational Linguistics*, 12:80–99.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, and 1 others. 2025. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next-generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

DeepSeek-AI. 2024. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*.

Louise Edwards. 2005. Fictions of enlightenment: Journey to the west, tower of myriad mirrors and dream of the red chamber, and: Androgyny in late ming and early qing literature (review). *China Review International*, 12(1):154–156.

Google. 2025. Gemini 3 Pro: The frontier of vision AI.

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

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*, 12:229–246.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam De Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, and 1 others. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013.

Wenbin Hu. 1993. 《红楼梦》在国外. 中华书局, 北京.

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. *arXiv preprint arXiv:2301.08745*.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, and 1 others. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46.

Mildred L. Larson. 1998. *Meaning-Based Translation: A Guide to Cross-Language Equivalence*, 2 edition. University Press of America, Lanham, MD.

Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2013. Multidimensional quality metrics: A flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London. Aslib.

Hiroaki Maruyama. 1998. 伊藤漱平訳,『紅楼夢』, 平凡社ラーブラリー, 全十二冊, 平成九年十一月完結. 二松学舎大学人文論叢, 61:102.

Nikita Moghe, Arnisa Fazla, Chantal Amrhein, Tom Kocmi, Mark Steedman, Alexandra Birch, Rico Sennrich, and Liane Guillou. 2025. Machine translation meta-evaluation through translation accuracy challenge sets. *Computational Linguistics*, 51(1):73–137.

Riccardo Moratto, Kanglong Liu, and Di-kai Chao, editors. 2022. *Dream of the Red Chamber: Literary and Translation Perspectives*. Routledge, London and New York.

Peter Newmark. 1988. *A Textbook of Translation*. Prentice-Hall International, New York.

Peter Newmark. 1998. *More Paragraphs on Translation*. Multilingual Matters, Clevedon.

Eugene A. Nida. 1945. Linguistics and ethnology in translation-problems. *Word*, 1(2):194–208.

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

Eugene A. Nida. 1993. *Language, Culture, and Translating*. Shanghai Foreign Language Education Press, Shanghai.

Su Khine Oo. 2020. Culture through keywords in kenneth wong's translation of "spirit food" by nu nu yi (inwa). *Journal of English Language and Linguistics*, 1(2):115–134.

OpenAI. 2025a. Introducing GPT-4.1 in the API.

OpenAI. 2025b. Introducing OpenAI o3 and o4-mini.

OpenAI. 2025c. OpenAI GPT-5 system card. *arXiv preprint arXiv:2601.03267*.

Saurabh Kumar Pandey, Harshit Budhiraja, Sougata Saha, and Monojit Choudhury. 2025. Culturally yours: A reading assistant for cross-cultural content. In *Proceedings of the 31st International Conference on Computational Linguistics: System Demonstrations*, pages 208–216.

Jianhui Pang, Fanghua Ye, Derek Fai Wong, Dian Yu, Shuming Shi, Zhaopeng Tu, and Longyue Wang. 2025. Salute the classic: Revisiting challenges of machine translation in the age of large language models. *Transactions of the Association for Computational Linguistics*, 13:73–95.

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

Yu-wei Pian and Wei Chen. 2022. English translation of culture-loaded words—a corpus-based study. *Journal of Literature and Art Studies*, 12(6):667–673.

Andrew H. Plaks. 2015. *Archetype and Allegory in the "Dream of the Red Chamber"*. Princeton University Press, Princeton, NJ.

Jantra Prompan. 2024. A study of culture-specific items (CSIs) and translation strategies in the blind earthworm in the labyrinth. *Rajapark Journal*, 18(58):212–235.

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

Haun Saussy. 1995. *The Problem of a Chinese Aesthetic*. Stanford University Press, Stanford, CA.

Sherrie Shen, Weixuan Wang, and Alexandra Birch. 2025. Liaozhai through the looking-glass: On paratextual explicitation of culture-bound terms in machine translation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34400–34416.

Dan Song. 2019. 《红楼梦》在日本的翻译与影响研究. 外语教学与研究, 51(1):121–132, 161.

Yuming Sun. 2006. 日本《红楼梦》研究略史. 红楼梦学刊, *No. 5*, pages 224–240.

Gideon Toury. 1995. *Descriptive Translation Studies and Beyond*. John Benjamins, Amsterdam and Philadelphia.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Lawrence Venuti. 1995. *The Translator's Invisibility: A History of Translation*. Routledge, London and New York.

Lucas Nunes Vieira, Minako O'Hagan, and Carol O'Sullivan. 2021. Understanding the societal impacts of machine translation: A critical review of the literature on medical and legal use cases. *Information, Communication & Society*, 24(11):1515–1532.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for translation: Assessing strategies and

performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427.

Yiming Wang, Pei Zhang, Jialong Tang, Haoran Wei, Baosong Yang, Rui Wang, Chenshu Sun, Feitong Sun, Jiran Zhang, Junxuan Wu, and 1 others. 2025. Polymath: Evaluating mathematical reasoning in multilingual contexts. *arXiv preprint arXiv:2504.18428*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Jun Wu. 2019. 阐释的演化：伊藤漱平《红楼梦》日译研究. 知识产权出版社, 北京.

Qing Xia. 2010. On english translation of culture-specific items in the ancient chinese official system: A descriptive and comparative study on hawkes's and yang's english translated cases of hong lou meng. *Cross-Cultural Communication*, 6(4):58–73.

Guozhang Xu. 1980. Culturally loaded words and english language teaching. 现代外语, *No. 4*, pages 21–27.

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.

Ran Zhang, Wei Zhao, and Steffen Eger. 2025. How good are LLMs for literary translation, really? literary translation evaluation with humans and LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10961–10988.

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.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623.

A Dataset

A.1 More Examples

Tab.5 - 9 presents dataset examples from five cultural categories.

A.2 Original PDF Format

Fig.5 and 6 present two cases of the original PDF format of our corpus with direct bilingual comparisons.

B Model Usage

B.1 Inference Implementation

All model implementations follow default settings. For closed-source models, we do not have access to hyperparameters. For open-source models, specific sampling hyperparameters are listed in Tab.10.

Given the randomness of sampling, theoretically, we should generate multiple responses and report the average. However, due to the high cost of human evaluation, both in time and expense, it is impractical to score every sampled output. To address this, we randomly selected 20 samples, had each model generate eight responses, and scored them separately by our human evaluators. We observed that score variance across samples was minimal, unlike in deterministic reasoning tasks, where outputs can vary significantly (Abdin et al., 2025; Chen et al., 2025; Wang et al., 2025). Under this premise, we use only the first sampled response for evaluation.

B.2 LLM-as-a-Judge Prompt

Tab.11 presents the prompt used for our llm-as-a-judge evaluations.

C Human Guideline

Tab.12 presents the explanations and scoring guidelines of four specific evaluation dimensions and one overall dimension used in our experiments.

Note: During human evaluation, all evaluators were presented with guidelines in their respective native languages. The English version is provided in this section for ease of presentation.

D Error Type Analysis

To better understand error patterns of models, we analyze error distributions across eight models in Tab.7. Errors are categorized into five types: *No Obvious Error*, *Mistranslation*, *Overtranslation*, *Undertranslation*, and *Omission* (see Tab.13

Table 5: Dataset examples of the **Ecology** category.

Source: 每日早起，拿上等燕窝一两，冰糖五钱，用银铫子熬出粥来。
Target: 朝、上等の燕（海つばめの。燕）一分と砂糖五分とを、の子で煮いてお粥にこしらえなさい。

Source: 我不吃六安茶
Target: 「わしは六安茶（安徽省霍山—昔、六安郡に属した—の茶）はまないのだよ」

Source: 此刻忽见宝玉笑问道："宝姐姐，我瞧瞧你的红麝串子?
Target: 宝玉はにこにこしながら、「宝おさま、あなたの分の赤い香りの香珠、わたしにませてくださいよ」と言い出すのでした。

Source: 只见晴雯如得了世露一般，一气都灌下去了。
Target: すると晴雯はさながら甘露でも得たかのように、一にこれを咽喉に流しこむのでした。

Source: 刘姥姥道："这个菜里有毒，俺们那些都成了砒霜了。那怕毒死了，也要吃尽了。"
Target: 婆さんは答えて、「このご走に毒がはいっておるちゅうことになりますと、てまえどもの料理なんぞは、砒霜（砒素の化合物、毒物）のってえことになっちまいますわい。たとい毒で死のうとも、ご走さえ平らげられたら本望でございます」

Table 6: Dataset examples of the **Religion** category.

Source: 贾珍便命贾琼，责琛、贾、贾蔷四个人去陪客，一面吩咐去请钦天监阴阳司来择日。
Target: に言いつけて客のに当たらせることとし、一方では使いをやって天（天文法をつかさどる役所）の司の（生、後出。易占をつかさどる）に日柄をたてさせます。

Source: 风姐听了这话，便发了兴头，说道："你是素日知道我的，从来不信什么是阴司地狱报应的。
Target: はこのことばをいてついりこまれ、「あなたはかねがねこのわたしという人をご存じのはず、魔のだの地のだのはついぞじたこともないわたしですもの」

Source: "我昨日叫赖升媳妇出去叫人给宝玉算算命。这先生算得好灵，说要娶了金命的人帮扶他，必要冲冲喜才好。不然只怕保不住。"
Target: わしは昨日、の家内を街へつかわして宝玉のことを占わせてみたのだが、その占い者の立てた卦というのがぴったりなのだよ—金の性の者を嫁にとってれ添わせなさい。どうしてもここでひとつ起直しせぬことにはいかん。でないと恐らくは生命も保つまい、とこういったそうだ。

Source: 你舅舅今日斋戒去了
Target: 伯父さんは今日はお役で戒にお出かけになってお留守なの。

Table 7: Dataset examples of the **Material** category.

Source: 王子腾那边，仍是一套衣服，一双鞋袜，一百寿桃，一百束上用银丝挂面；
Target: 王子のもとからは、例によって衣服一靴下一足寿桃（生祝いに送る小麦粉の桃）百中用素百束が届きます

Source: 这红玉也不梳洗，向镜中胡乱挽了一挽头发，洗了洗手，腰内束了一条汗巾子便来打扫房屋。
Target: 玉とておちおち身づくろいなどしてはおられず、そうそうにに向かってをわがね、手を洗い、腰をしめるなり、部屋の除にやってきました。

Source: 宝琴披着凫靥裘站在那里笑
Target: の毛の装をはおった宝琴がそこに立って笑っています。

Source: 另换了三四个衣帽周全十七八岁的小厮上来，复抬起轿子，众婆五步下围随，至一垂花门前落下。
Target: するとこんどはに三、四人、お仕着せ姿をした十七、八の若党が交替にきて、またをかきあげ、老女たちがそのりをとりまくようにして徒でついてゆき、垂花（正をはいった中、二の。垂花のかざりがあるのでいう）まできて幅をおろしました。

Source: 散押岁钱荷包金银锞；摆上合欢宴来，男东女西归坐，献屠苏酒、合欢汤、吉祥果、如意糕毕。
Target: そこで押（大晦日に上から子供に取らせるお年玉の金。穴あき百枚を赤で通してある）中着金の小粒などを分けて取らせます。また合宴の宴席をけて、男はの、女は西の席につき、屠酒合（スプ）吉祥果（果物）如意糕（蒸し子）を献じえました。

Table 8: Dataset examples of the **Linguistics** category.

Source: 谋事在人，成事在天
Target: 人さまがお膳立て、天道さまがお取りあげ

Source: 那宝玉是个丈八的灯台，照见人家，照不见自家的。
Target: どだい肝の宝玉さまが、『文八（一文八尺）のお灯明台：他人は照らせても、わが身は照らせぬ（「灯台もと暗し」）』。

Source: 巧媳妇做不出没米的粥来
Target: 遣り上手の嫁さんでも米なしでは粥はできぬ

Source: 偏偏凤姐想出一条偷梁换柱之计
Target: あいにくちゃんが替え玉の略を考え出してくれた

Source: 于是尤氏一行人悄悄的来至窗下，只听里面称三赞四，耍笑之音虽多；又兼着恨五骂六，忿怨之声亦不少。
Target: かくて尤氏ら一行、足音を忍ばせての下までやってきましたところ、なかではほめたりたたえたりで笑いさんざめく声がしきりにする一方、またわめいたり恨んだりの怒りと憎しみの声も少なくないふう

Table 9: Dataset examples of the **Society** category.

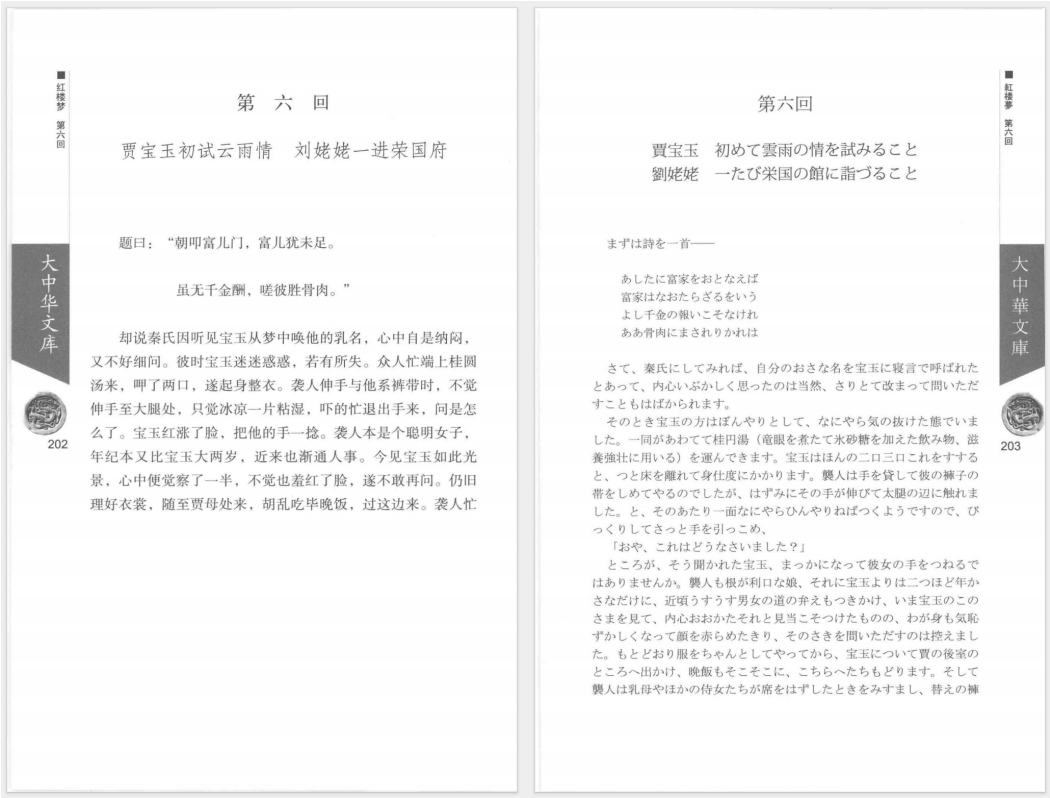
Source: 我们合家大小登门去磕头。
Target: わたくしども家中ってお宅に伺い叩させていただきます

Source: 他是我们这里有名的一个泼皮破落户儿，南省俗谓作辣子。
Target: これはうちではこえたお婆の破落、江南の方なら俗に『辣子』というやつよ

Source: 他又成了香饽饽了，都抢不到手。
Target: あのひとはほかほか（人者の意）なってしまい、い合いでめったに手にははいらないね

Source: 平儿便福下去，宝玉作揖不迭。
Target: 平がそこで「万福（女子の敬礼のときのことば）」といってお辞をしますと、宝玉はれじ揖礼（手を挟いて上下する男子の敬礼法）を返します。

Source: “孽障！你生气，要打骂人容易，何苦摔那命根子！”
Target: 「このあたりめが！おまえ、かんしゃくを起こしたら、人をぶつたりどなりつけたりするのだってたやすいのに、りにってその命のも同然の品を投げつけてなんとする？」



■红楼梦 第六回

第 六 回

贾宝玉初试云雨情　刘姥姥一进荣国府

题曰："朝叩富儿门，富儿犹未足。

虽无千金酬，嗟彼胜骨肉。"

却说秦氏因听见宝玉从梦中唤他的乳名，心中自是纳闷，又不好细问。彼时宝玉迷迷惑惑，若有所失。众人忙端上桂圆汤来，呷了两口，遂起身整衣。袭人伸手与他系裤带时，不觉伸手至大腿处，只觉冰凉一片粘湿，吓的忙退出手来，问是怎么了。宝玉红涨了脸，把他的手一捻。袭人本是个聪明女子，年纪本又比宝玉大两岁，近来也渐通人事。今见宝玉如此光景，心中便觉察了一半，不觉也羞红了脸，遂不敢再问。仍旧理好衣裳，随至贾母处来，胡乱吃毕晚饭，过这边来。袭人忙

大中华文庫

202

■紅楼夢 第六回

第六回

賈宝玉　初めて雲雨の情を試みること
劉姥姥　一たび栄国の館に詣づること

まずは詩を一首——

あしたに富家をおとなえば
富家はなおたらざるをいう
よし千金の報いこそなけれ
ああ骨肉にまされりかれは

さて、秦氏にしてみれば、自分のおさな名を宝玉に寝音で呼ばれたとあって、内心いぶかしく思ったのは当然、さりとて改まって問いただすこともはばかられます。

そのとき宝玉の方はぼんやりとして、なにやら気の抜けた態でいました。一同があわてて桂円湯（竜眼を煮たて氷砂糖を加えた飲み物、滋養強壮に用いる）を運んできます。宝玉はほんの二口三口これをすすると、つと床を離れて身仕度にかかります。襲人は手を貸して彼の褲子の帯をしめてやるのでしたが、はずみにその手が伸びて太腿の辺に触れました。と、そのあたり一面なにやらひんやりねばつくようですので、びっくりしてさっと手を引っこめ、

「おや、これはどうなさいました？」

ところが、そう聞かれた宝玉、まっかになって彼女の手をつねるではありませんか。襲人も根が利口な娘、それに宝玉よりは二つほど年かさなだけに、近頃うすうす男女の道の弁えもつきかけ、いま宝玉のこのさまを見て、内心おおかたそれと見当こそつけたものの、わが身も気恥ずかしくなって顔を赤らめたきり、そのさきを問いただすのは控えました。もとどおり服をちゃんとしてやってから、宝玉について賈の後室のところへ出かけ、晩飯もそこそこに、こちらへたちもどります。そして襲人は乳母やほかの侍女たちが席をはずしたときをみすまし、替えの褲

大中華文庫

203

Figure 5: Original PDF format of our corpus with direct bilingual comparisons (Case 1).

■红楼梦 第十四回

盒、脚踏之类。一面交发，一面提笔登记，某人管某处，某人领某物，开得十分清楚。众人领了去，也都有了投奔，不似先时只拣便宜的做，剩下的苦差没个招揽。各房中也不能趁乱失迷东西。便是人来客往，也都安静了，不比先前一个正摆茶，又去端饭，正陪举哀，又顾接客。如这些无头绪，荒乱推托偷闲窃取等弊，次日一概都蠲了。凤姐儿见自己威重令行，心中十分得意。因见尤氏犯病，贾珍又过于悲哀，不大进饮食，自己每日从那府中煎了各样细粥，精致小菜，命人送来劝食。贾珍也另外吩咐每日送上等菜到抱厦内，单与凤姐。那凤姐不畏勤劳，天天于卯正二刻，就过来点卯理事。独在抱厦内起坐，不与众妯娌合群，便有堂客来往也不迎会。

这日乃五七正五日上，那应〔赴〕僧正开方破狱，传灯照亡，参阎君，拘都鬼，延请地藏王，开金桥，引幢幡；那道士们正伏章申表，朝三清，叩玉帝；禅僧们行香，放焰口，拜水忏；又有十三众尼僧搭绣衣，皂红鞋，在灵前默诵接引诸

大中华文庫

478

■紅楼夢 第十四回

て帳面に控えをとり、だれそれはどこそこを受け持つ、誰それはこれこれを管理するといった式に一目でわかるように付けておきます。一同はそれぞれ受領して立ってゆきましたが、なんといっても自分の持ち場が決まったとなると、これまでみたいにだれも自分の得になる仕事なら飛びついてでもしたがる反面、余されもののありがたくない役目は買って出る者もないというような不都合も解消しました。またこれでは、どの部屋にしても、どさくさまぎれにものが紛失といったことも起こりようがありません。たとえ人の出入りが激しくとも、みな落ちついて構えておられ、これまでのように同じ人間がお茶を出したかと思うとまたご飯を運びに立ったり、客につき合って哀哭しているさいちゅうに、こんどは別の客の取り次ぎにも出なければならぬという状態とはおおちがい。そこでさしもの乱脈を極め、仕事を逃げる、ずるけて休む、ものをちょろまかすたぐいの手のつけようもなかった弊風も、翌日からふっつり地をはらうにいたりました。

熙鳳は自分の威令の並び行われるのを見て、内心満更でもありません。尤氏は病床にあり、賈珍はまた嘆きが過ぎてかどうも食が進まないようだとあって、熙鳳は自身、毎日あちらの屋敷から、手のこんだ粥、しゃれた味の漬物などをこしらえさせては届けさせ、賈珍に召せとすすめます。賈珍も毎日上等の料理を別あつらえでこしらえさせて抱廈まで運ばせ、熙鳳にだけ食べさせるのでした。熙鳳の方では骨身惜しまず、毎日正卯の二刻には出かけてきて点呼をとり、万事を取りしきります。そして自分ひとりは抱廈を根城にしてほかの相嫁たちとは合流せず、女客の訪問があろうとどこ吹く風で、応対に出ようともしません。

さて、その日は五七の正五日の当日に当たっていましたので、応付僧はいましも方便門を開けて地獄に押し入り、灯の力で亡者を照らしてやる一方、閻魔大王に目通りして鬼の頭をつかまえていてもらい、地蔵菩薩を頼んできて、三途の渡しの金橋を開放、幢幡を引いてもらおうという段取り。折から道士たちは、上帝に上奏文を奉り、三清（道教でいう元始天尊・太上道君・太上老君の三神）に拝謁、玉帝に叩首しようというところ、禅僧たちは香を焚いて礼拝し、施餓鬼の経文を誦し、「水懺」（厄ばらいの経文）をあげ故人に代わって頂礼懺悔します。それからまたぬいとり衣をまとい赤い靴を突っかけた十三人の若い尼僧たちが、死者を導くためのもろもろの呪文を霊前にあって黙誦しているというふうで、いかさま大変な賑やかさです。

大中華文庫

479

Figure 6: Original PDF format of our corpus with direct bilingual comparisons (Case 2).

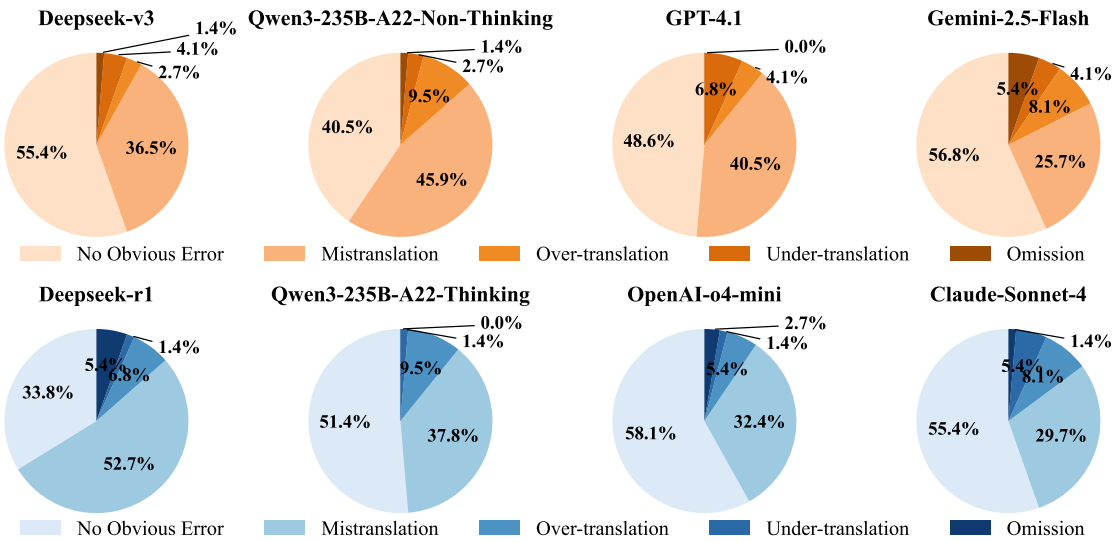


Figure 7: Distribution of translation error types across eight models.

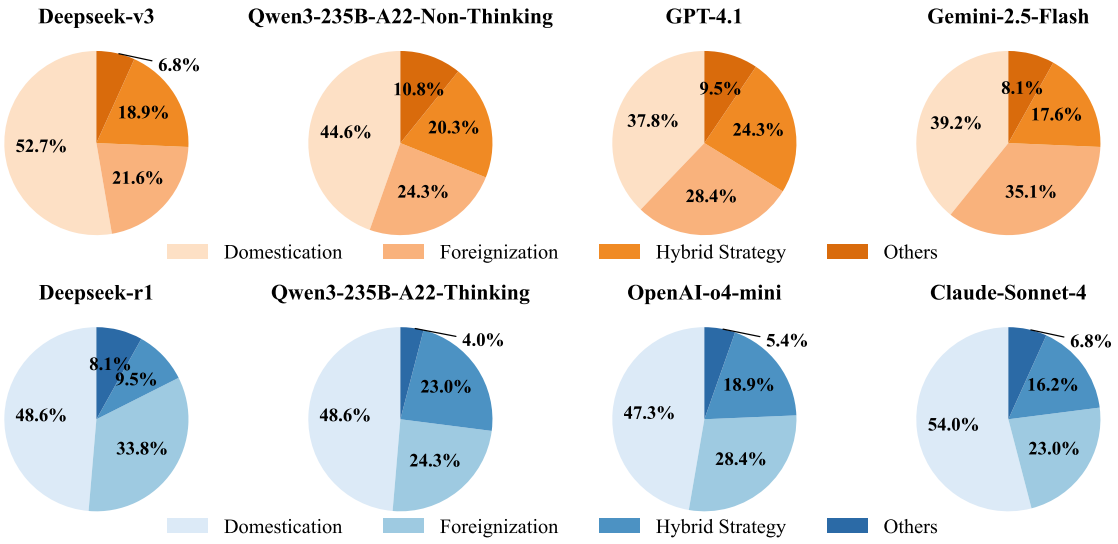


Figure 8: Distribution of culture-oriented translation strategy types across eight models.

Table 10: Inference hyperparameters used in our experiments.

Model	Top-k	Top-p	Temperature T
Deepseek-v3	20	0.95	0.6
Deepseek-r1	20	0.95	0.6
Qwen3-235B-A22B-Non-Thinking	20	0.8	0.7
Qwen3-235B-A22B-Thinking	20	0.95	0.6

for details). Top-performing models, including OpenAI-o4-mini, Gemini-2.5-Flash, and Deepseek-v3, show the highest *No Obvious Error* rates (55–58%) and lowest *Mistranslation* (25–36%). In contrast, Deepseek-r1 achieves only 33.8% error-free and 52.7% mistranslation, confirming that misunderstanding source content is the primary accuracy bottleneck.

The remaining three error types are relatively rare, together accounting for under 15% of cases for most models. *Undertranslation* and *Omission* remain consistently low across all systems (<7% and <5.4% respectively), indicating models rarely refuse to translate or drop content entirely. In addition, *Overtranslation* shows an interesting pattern: reasoning models exhibit notably higher rates (Qwen3-235B-A22-Thinking 9.5%, Claude-Sonnet-4 8.1%, Gemini-2.5-Flash 8.1%) compared to non-reasoning Deepseek-v3 (2.7%). While longer CoT encourages elaboration, such additions often over-explain cultural elements or introduce irrelevant content, failing to improve outcomes. This confirms that "thinking more" does not guarantee better cultural adaptation — effective translation requires appropriately calibrated elaboration.

Table 11: Prompt used for LLM-as-a-judge evaluation.

Instruction:
You are an expert in translation evaluation with native-level proficiency in both source and target languages and deep knowledge of both cultures. You will evaluate a translation based on the following four dimensions. After considering each dimension, provide an **Overall Score** from 0 to 5 that holistically reflects the translation's quality.

—

Evaluation Dimensions:

1. Content Accuracy
Measures whether the translation faithfully and accurately conveys the meaning of the source text.
- High accuracy means no noticeable deviation from the original meaning.
- Low accuracy indicates mistranslations, distortions, or loss of core content.

2. Language Fluency
Measures whether the translation conforms to the linguistic norms of the target language, including grammar, vocabulary, spelling, punctuation, and naturalness of expression.
- Fluent translations read smoothly and naturally.
- Low fluency involves awkward wording, grammatical errors, or obvious translationese.

3. Cultural Appropriateness
Evaluates whether appropriate translation strategies are used when handling cultural elements in the source text, ensuring that cultural connotations are properly conveyed and understandable to target readers.
- High cultural appropriateness preserves cultural nuances without causing confusion.
- Low appropriateness results in cultural loss, distortion, or misunderstanding.

4. Native Readability
Evaluates whether the translation sounds natural and acceptable to native speakers of the target language and can be easily understood without barriers.
- High readability means the text is clear and fully understandable at first reading.
- Low readability requires inference, repeated reading, or leaves only a general impression.

—

Overall Scoring Criteria:
- **5 - Excellent**: Highly accurate, fluent, culturally appropriate, and reads naturally for native speakers. No noticeable issues.
- **4 - Good**: Generally accurate and fluent with minor deviations; cultural meaning largely preserved; occasional awkwardness but easily understood.
- **3 - Adequate**: Main content is understandable, but noticeable issues exist in accuracy, fluency, or cultural handling; may require some effort to read.
- **2 - Poor**: Significant deviations or errors; cultural meaning poorly conveyed; awkward or hard to follow; only general ideas graspable.
- **1 - Very Poor**: Severe distortions or omissions; cultural content missing; highly unnatural or unintelligible.
- **0 - Completely Wrong**: Translation is unrelated to the source or empty.

—

Source Text:
{source_text}

Translation:
{target_text}

Evaluation:
Provide a **brief justification** (one or two sentences) and then the **Overall Score** in the following format:

Justification: [your reasoning]
Overall Score: [0-5]

Table 12: Human evaluation criteria used in our experiments.

Dimension	Definition	Scoring Criteria
Content Accuracy	Measures whether the translation faithfully and accurately conveys the meaning of the source text.	**5**: Highly accurate with no noticeable deviation. **4**: Generally accurate with minor distortion. **3**: Some deviations exist, but the main content is understandable. **2**: Significant deviation; core content difficult to understand. **1**: Severely deviates from the source text; core meaning completely lost.
Language Fluency	Measures whether the translation conforms to the linguistic norms of the target language, including grammar, vocabulary, spelling, punctuation, and naturalness of expression.	**5**: Smooth and natural; linguistically well-formed. **4**: Generally fluent with slight unnaturalness. **3**: Basically understandable but with noticeable issues. **2**: Awkward wording; difficult to understand or obvious translationese. **1**: Frequent errors; difficult to understand.
Cultural Appropriateness	Evaluates whether appropriate translation strategies are used when handling cultural elements in the source text, ensuring that cultural connotations are properly conveyed and understandable to target readers.	**5**: Cultural meanings are accurately conveyed and easily understood. **4**: Generally fluent and appropriate with minor deviations. **3**: Cultural information largely preserved but with some mistranslation. **2**: Cultural meaning poorly conveyed, causing misunderstanding. **1**: Cultural content missing or seriously distorted.
Native Readability	Evaluates whether the translation sounds natural and acceptable to native speakers of the target language and can be easily understood without barriers.	**5**: Meaning is clear and fully understandable. **4**: Overall natural with minor awkwardness. **3**: Comprehension requires inference or repeated reading. **2**: Most content obscure; only general idea graspable. **1**: Highly unnatural; difficult for native readers.
Overall Score	Provides a comprehensive evaluation of the translation's overall quality by considering all aspects above.	**5**: Excellent overall performance across all aspects. **4**: Generally good with minor shortcomings. **3**: Moderate quality with noticeable issues. **2**: Poor overall quality with significant problems. **1**: Extremely poor overall performance.

E Translation Strategy Analysis

We further analyze the translation strategies adopted by the models during the translation process, as shown in Fig.8. Specifically, we draw on the "domestication–foreignization" theory introduced in §2.3 and categorize translation strategies into four types (see Tab.14 for details). Overall, each model employs either domestication or foreignization strategies in over 80% of cases, with mixed or other strategies being relatively infrequent. A clear preference for domestication emerges across models, prioritizing adaptation to target language norms. Notably, reasoning models adopt domestication at higher rates than their non-reasoning counterparts.

Two primary domestication patterns are observed. The first involves Japanese idiomatic expressions, which preserve core meaning and ensure fluency but replace source cultural imagery with target-culture equivalents, resulting in cultural loss. The second relies on generic paraphrasing, which captures only coarse semantics while omitting nuanced components, leading to partial meaning degradation (see §F.1 and F.4 for a typical case). These findings suggest that while domestication enhances readability and target-language acceptability, it often comes at the cost of source culture preservation — underscoring a key limitation of current models in conveying culturally embedded meanings.

Table 13: Error types and corresponding descriptions.

Category	Description
A. No obvious error	The translation accurately conveys the source content without any issues.
B. Mistranslation	The content is incorrectly understood, resulting in a translation that deviates from the original meaning.
C. Overtranslation	Information not present in the source text is added to the translation.
D. Undertranslation	Portions of the source text are left untranslated and directly copied into the target text.
E. Omission	Source information is omitted, resulting in an incomplete translation.

Table 14: Translation strategy types and corresponding descriptions.

Category	Description
A. Domestication	Oriented toward the norms of the target culture, adapting the translation to fit the cognitive and cultural expectations of target readers. Strategies include localization, generalization, cultural substitution, functional equivalence, etc.
B. Foreignization	Preserves the cultural characteristics and heterogeneity of the source text, allowing the translation to reflect the original cultural context. Strategies include literal translation, transliteration, semantic borrowing, formal equivalence, etc.
C. Hybrid Strategies	Combine domestication and foreignization to strike a balance between cultural fidelity and reader comprehension. Strategies include literal translation with annotations, retention with explanation, partial literal translation combined with partial free translation, etc.
D. Others	Translation approaches that do not fall directly under domestication or foreignization, often involving omission or entirely creative renditions. Strategies include omission, innovative translation, etc.

F Case Study

F.1 Case 1

Source:
他又成了香饽饽了，都抢不到手。

Reference:
あのひとはほかほか饅頭（人気者の意）なってしまい、奪い合いでめったに手にははいらないね

Deepseek-v3:
彼はまた人気者になって、みんな手に入れようにも手に入れられないんだ。彼はまた人気者になって、みんなが奪い合っている。

Qwen3-235B-A22B-Non-Thinking:
彼はまたもてはやされる存在になり、手に入れようとしても手に入らないほどだ。

GPT-4.1:
また引っ張りだこになって、誰も手に入れられないよ。

Gemini-2.5-Flash:
彼はまた引っ張りだこになって、もう手に入らないよ。

Deepseek-r1:
彼はまた引く手あまたになってしまって、もらえないんです。

Qwen-235B-A22B-Thinking:
彼はまたもてはやされる存在となり、誰もが手に入れようとするものの、なかなか手に入らない状態です。

OpenAI-o4-mini:
彼はまた人気者になって、みんな手に入れようにも手に入れられないんだ。

Claude-sonnet-4:
彼はまた人気者になって、取り合いになっているよ。

Analysis. The expression “香饽饽” in the original text functions as a key culturally loaded word. Literally, it refers to a “steaming hot bun”, while metaphorically it conveys the cultural meaning of “a highly sought-after person”. Ito renders it as 「ほかほか饅頭（人気者の意）」. This

approach represents a classical paradigm in cultural translation: the source-language imagery is preserved through literal translation, while the metaphorical meaning is made explicit through explanatory notes in parentheses. In this way, the translation both maintains the cultural imagery of the source text and ensures comprehensibility for target-language readers.

LLMs tend to prioritize alignment with the linguistic conventions and expression patterns of the target language at the expense of source-language cultural imagery. This preference is realized primarily through two approaches:

- The first approach involves replacing the source expression with established Japanese idioms. For example, GPT-4.1 and Gemini-2.5-Flash translate the term as 「引っ張りだこ」, while Deepseek-r1 uses 「引く手あまた」. Although these expressions accurately convey the core meaning of "being highly sought after" and produce fluent, idiomatic output, they essentially constitute cultural substitution. As a result, the culturally specific imagery associated with "饽饽", a symbol rooted in Chinese culinary culture, is lost.

- The second approach generalizes the expression into more neutral vocabulary. Examples include 「人気者」 in the translations by Deepseek-v3 and Claude-4-sonnet, and 「もてはやされる存在」 produced by Qwen series. While such renderings convey the attribute of "being popular," they substantially weaken or even eliminate the dynamic semantic component of "being competed for" (i.e., the sense of rivalry or competition implied in the original expression). Consequently, part of the original semantic richness is lost.

F.2 Case 2

Source:
那宝玉是个丈八的灯台，照见人家，照不见自家的。

Reference:
どだい肝腎の宝玉さまが、『丈八（一丈八尺）のお灯明、他人は照らせても、わが身は照らせぬ（「灯台もと暗し」）』。

Deepseek-v3:
宝玉は八尺の灯台で、他人を照らすが、自分の足元は照らさない。

Qwen3-235B-A22B-Non-Thinking:
その宝玉というやつは、丈八（じょうはち）の灯台（とうだい）のようなもので、他人（ひと）のことは照らし出すが、自分自身は照らし出せないのだ。

GPT-4.1:
あの宝玉は、まるで丈八（じょうはち）の灯台のようなもので、人のことはよく見えるが、自分のことは見えないのです。

Gemini-2.5-Flash:
この宝玉（ほうぎょく）は、まさに「灯台下暗し（とうだいもとくらし）」という男だった。

Deepseek-r1:
宝玉は一丈八尺の灯籠台、他人を照らすが、自分の足元は見えぬ

Qwen-235B-A22B-Thinking:
宝玉は丈八（じょうはち）の灯台のようなものだ。他人は照らすが、自分自身は照らせぬ。

OpenAI-o4-mini:
あの宝玉は丈八の燈台みたいなもので、人のことはよく照らせても、自分の家（身内）のことはまったく見えない。

Claude-sonnet-4:
その宝玉は、他人のことはよく分かるが、自分のことは分からない人なのである。

Analysis. This case belongs to a citation-type two-part allegorical saying (歇后语) structure composed of "tenor + vehicle + explanation". The first half establishes the metaphor, while the second half reveals the intended meaning of the pun. From a translation perspective, the major difficulty of this sentence lies in two aspects: the treatment of the cultural image and the explicit rendering of the metaphorical implication. On the one hand, the translation needs to reproduce the exaggerated metaphorical image of "丈八的灯台"; on the other hand, it must convey the satirical meaning embedded in the original expression, namely "seeing others' faults while failing to examine one's own".

Ito's translation adopts a strategy of preserving the source metaphor while providing explanatory support. First, the expression 「丈八（一丈八尺）のお灯明」 retains the original imagery, and the term "丈八" is annotated to ensure comprehension for target-language readers. Second, the latter half is translated literally as 「他人は照らせて

も、わが身は照らせぬ」, which accurately conveys the meaning of the original while maintaining a formal correspondence with the source structure. Finally, the translator supplements the Japanese proverb 「灯台もと暗し」, further facilitating reader understanding. As a result, the translation both preserves the cultural imagery and makes the metaphorical meaning explicit, achieving a desirable balance between cultural fidelity and reader comprehensibility.

For LLMs, their translations can generally be categorized into the following types:

- The first type combines structural preservation with explanatory adaptation. For example, GPT-4.1 adopts a canonical Japanese comparative construction for expressing metaphor and contrast, 「まるで...のようなもので、...よく見えるが、...見えないのです」. The resulting translation is fluent and rhythmic, effectively restoring the syntactic structure of the original sentence.

- The second type adopts a literal translation strategy for the cultural image. Examples include deepseek-v3-250324, o4-mini-2025-04-16, deepseek-r1-250528, and the qwen3 series. A common characteristic of these translations is that "丈八" is directly rendered as 「八尺」, 「丈八」, or 「一丈八尺」. Although such literal translation preserves the cultural features of the source language to the greatest extent, the absence of further explanation or adaptation may reduce readability for Japanese readers unfamiliar with the cultural reference.

- The third type adopts a domestication or paraphrasing strategy, as seen in the translations produced by gemini-2.5-flash and claude-sonnet-4-20250514. While these translations achieve high fluency, the cultural imagery of the source text is entirely replaced or erased. From the perspective of cultural transmission, this strategy may fulfill communicative purposes but inevitably results in a loss of cultural information.

- Finally, some translations generated by the DeepSeek series contain clear errors. In these cases, "丈八的灯台" is interpreted as a literal physical object rather than a metaphor, and the metaphorical relationship is therefore not reflected. Moreover, "照不见自家" is rendered as 「自分の足元は見えない」, which concretizes the expression into a spatial description. Consequently, the original satirical implication—referring to a lack of self-reflection—degenerates into a neutral depiction of physical visibility.

F.3 Case 3

Source:
.....只听里面称三赞四，耍笑之音虽多；又兼着恨五骂六，忿怨之声亦不少。

Reference:
.....なかでは、ほめたりたたえたりで、笑いさんざめく声がしきりにする一方、またわめいたり恨んだりの怒りと憎しみの声も少なくないふう.....

Deepseek-v3:
中では三人を褒め四人を誉め、笑い戯れる声が多いかと思えば、また五人を恨み六人を罵る、怒り怨む声も少なくない。

Qwen3-235B-A22B-Non-Thinking:
中からは三人称えて四人ほめる声、遊び笑う音は多いにあり、また五人を恨み六人を罵る声、怒りと不満の言葉も少なくない。

GPT-4.1:
中からは、「三を褒め四を称える」声や、ふざけ合う音が多く聞こえてきたが、一方では「五を恨み六を罵る」怒りや恨みの声も少なからず混じっていた。

Gemini-2.5-Flash:
(30)中からはお世辞を言ったり、褒めそやしたり、ふざけ合ったりする声が多く聞こえたが、その一方で、罵声や恨み言、不平不満の声も少なくなかった。

Deepseek-r1:
中からは賑やかな笑い声とともに、褒めそやす話し声が多く聞こえる一方で、恨み罵る憤りの声も少なからず聞こえてきた。

Qwen-235B-A22B-Thinking:
中をうかがうと、そこには三人称賛し、四人で笑い声を立てている者が多く、また五人を恨み、六人を罵るような怒りや不満の声も少なくなかった。

OpenAI-o4-mini:
中からは絶え間なくおだてるような笑い声が聞こえてきたが、同時に罵倒や恨み言を口にする怒りの声も少なくなかった。

Claude-sonnet-4:
中からは三を褒め四を讃えるような、戯れ笑

う声が多く聞こえてくるものの、同時に五を恨み六を罵るような、憤懣怨嗟の声も少なくなかった。

Analysis. Taking the expression "称三赞四""恨五骂六", a typical culturally loaded word under the linguistic category, as an example, the key challenge in its translation lies not only in correctly interpreting the rhetorical device of numerical vagueness characteristic of Chinese, but also in reproducing the rhythmic vitality and emotional intensity created by the parallel structure of the original expression.

Ito renders "称三赞四" as 「ほめたりたたえたり」 and "恨五骂六" as 「わめいたり恨んだりの怒り」. This translation employs the Japanese parallel construction 「～たり～たり」, which not only accurately conveys the semantic content of the original but also reproduces its rhythmic pattern and emotional tension. In contrast, LLMs generally fall into two categories.

- On the one hand, the translations by Deepseek-v3, GPT-4.1, and Claude-sonnet-4 mistakenly interpret the numerals "三、四、五、六"("three, four, five, six"), which function rhetorically as vague intensifiers, as literal quantity markers. Consequently, they are rendered as expressions such as 「三人を褒め四人を讃える」 ("praising three people and commending four") or 「五人を恨み六人を罵る」 ("hating five people and insulting six"), resulting in a clear distortion of the original meaning.

- On the other hand, OpenAI-o4-mini, Deepseek-r1, and Gemini-2.5-Flash, render "称三赞四" and "恨五骂六" as expressions such as 「お世辞を言う」, 「褒めそやす」, 「罵倒する」, or 「恨み言」. Although these translations successfully convey the core semantic content of praise and denunciation expressed in the original, they lack the linguistic tension and rhetorical dynamism of the source text, and therefore fail to fully reproduce its tone and emotional force.

F.4 Case 4

Source:
巧媳妇做不出没米的粥来

Reference:
遣繰り上手の嫁さんでも米なしでは粥はできぬ

Deepseek-v3:
巧婦、米なしに粥を作ることはできない。

Qwen3-235B-A22B-Non-Thinking:
巧も米がなければ粥は作れない。

GPT-4.1:
どんなに腕のいい主婦でも、米がなければお粥は作れない。

Gemini-2.5-Flash:
元手がなければ商売はできない

Deepseek-r1:
ない袖は振れぬ

Qwen-235B-A22B-Thinking:
巧婦も米がなければ粥は炊けぬ。

OpenAI-o4-mini:
どんなに器用な嫁でも、米のないお粥は作れない。

Claude-sonnet-4:
米なしでは、どんなに腕の良い嫁でも粥は作れない

Analysis. Overall, the translations produced by LLMs can be broadly divided into two categories.

- The first category adopts a domestication strategy aimed at achieving functional equivalence. A representative example is the DeepSeek-r1 translation 「ない袖は振れぬ」. This type of translation completely abandons the concrete imagery in the original text—such as "米", "粥", and "媳妇"—and instead aligns directly with a semantically equivalent idiomatic expression in Japanese culture. Although this approach diverges significantly from the original in terms of linguistic form, it succeeds in achieving pragmatic equivalence at the functional level.

- The second category adopts a foreignization strategy, preserving the cultural imagery of the source language. For instance, the Qwen3-thinking translation 「巧婦も米がなければ粥は炊けぬ」 retains the original imagery of "巧妇", "米", and "粥". Moreover, by employing expressions such as 「炊けぬ」, which carry a strong classical Sino-Japanese stylistic flavor, the translation closely mirrors the structural and rhetorical style of the original Chinese proverb, thereby achieving a dual form of foreignization in both imagery and syntax.

It is worth noting, however, that when foreignization lacks sufficient cultural contextual support,

it may lead to semantic deviation. Models such as Gemini-2.5-Flash and GPT-4.1 exhibit varying degrees of strategic fluctuation in their sentence-level translations. For example, Gemini renders the proverb as 「元手がなければ商売はできぬ」 ("without capital, business cannot be conducted"). Although the translation maintains the formal structure of a proverb, it introduces a substantial shift in meaning by abstracting the original expression into a commercial context.

F.5 Case 5

Source:
那畸角上我有一个花冢。

Reference:
あの隅の方に、わたくし、花塚をこしらえてあります

Deepseek-v3:
あの隅っこに私の花塚があります。

Qwen3-235B-A22B-Non-Thinking:
その変な角のところに、私はお墓参りをしています。

GPT-4.1:
その歪んだ角に、私は一つの花塚（はなづか）を持っている。

Gemini-2.5-Flash:
その奇妙な角の上に、私は一つの花塚を建てた。

Deepseek-r1:
あの隅っこに私の花の塚があるんだ。

Qwen-235B-A22B-Thinking:
あの片隅に、私の花塚があります。

OpenAI-o4-mini:
あの歪んだ角に、私はひとつの花の塚を持っている。

Claude-sonnet-4:
その隅っこに私は花塚を持っています。

Analysis. "畸角" originally refers to a spatial notion, meaning a "remote, neglected, or unnoticed corner". The translation strategies adopted by LLMs models can be broadly divided into two categories.

- The first category consists of contextually grounded literary translations. Representative examples include the Reference translation, Qwen3-235B-A22B-Thinking, and the Deepseek series. These models correctly identify the spatial attribute embedded in the character "畸" within Chinese rhetoric. The Reference translation uses 「隅の方」, while DeepSeek renders it as 「隅っこ」, both of which accurately anchor the expression in physical spatial orientation. Among them, the translation 「片隅」 produced by Qwen3-235B-A22B-Thinking stands out as particularly effective. In Japanese lexical intuition, 「片隅」 not only denotes a marginal space but also carries a subtle sense of loneliness and isolation, which aesthetically resonates with the imagery of "花冢". This alignment suggests that models equipped with stronger reasoning abilities can move beyond literal lexical correspondence and achieve cross-lingual equivalence at the level of scene or atmosphere, rather than merely word meaning.

- The second category can be described as moderate translations based on literal semantic equivalence. Models such as GPT-4.1, OpenAI-o4-mini, and Gemini-2.5-Flash exhibit a clear tendency toward dictionary-like literal translation. They interpret the character "畸" as indicating abnormal shape, rendering the phrase as expressions such as 「歪んだ角」 or 「奇妙な角」. Similarly, Qwen3-non-thinking translates it as 「変な角」. Such treatments establish only a superficial lexical correspondence and fail to capture the intended meaning of the original expression.

In summary, the quality of translating "畸角" largely depends on whether the model recognizes its metaphorical function as an environmental modifier rather than a description of shape. Models capable of detecting contextual cues and employing expressions such as 「片隅」 or 「隅」, which align with natural usage in the target language, demonstrate clear advantages over literalist approaches in both translational accuracy and literary aesthetic quality.