From Words to Worlds: Benchmarking Culture-Loaded Expression Translation in Large Language Models

Anonymous ACL submission

Abstract

Culture-loaded expressions, such as idioms, slang, and culture-specific items (CSIs), are pervasive in natural language and encode meanings beyond literal linguistic form. Their mistranslation can obscure cultural references and reduce target-language interpretability, especially when culturally grounded meanings must be realized across different target languages. However, existing benchmarks remain fragmented and lack a systematic framework for evaluating culture-loaded translation in multi-target settings. To address this gap, we introduce CulT-Eval, a benchmark centered on Chinese culture-loaded expressions, which encode rich historical, literary, social, and institutional knowledge. CulT-Eval contains 7,959 curated instances with cultural span annotations, Cultural Explications, and diagnostic error labels. Through extensive evaluation of neural machine translation systems and large language models, we identify recurring failure modes that are not adequately captured by existing automatic metrics. Accordingly, we propose CURE, a Cultural Explication-grounded metric that separates cultural validity from target-language realization quality. Results show that current models struggle to preserve culturally grounded meaning in both English and Japanese, while dominant error patterns differ between the two target settings. Our benchmark and code are available at `https://anonymous.4open.science/r/CulT-Eval-413B`.

1 Introduction

Rooted in shared cultural knowledge and social conventions, many natural-language expressions convey meanings that rely on implicit contextual knowledge rather than explicit linguistic form. We refer to them as culture-loaded expressions, including idioms, slang, literary allusions, and culture-specific items (CSIs); representative instances are shown in Figure 1. Their mistranslation can obscure cultural references, distort pragmatic force, and reduce target-language interpretability. Prior studies on idioms, chengyu, interpretative slang, and CSIs further show that such expressions often require recovering intended meaning from cultural and historical context rather than direct lexical mapping (De Luca Fornaciari et al., 2024; Fu et al., 2025; Liang et al., 2025; Yao et al., 2024).

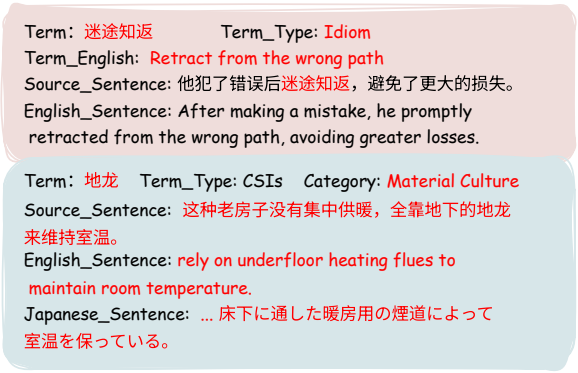

Figure 1: Representative CulT-Eval instances.

Despite their importance, existing benchmarks do not yet provide a unified basis for evaluating culture-loaded translation. Prior resources have made valuable progress on specific expression types and domains, but they often focus on isolated phenomena or use incompatible annotation and evaluation schemes. As summarized in Table 1, few benchmarks jointly support broad culture-loaded expression coverage, span-level annotation, Cultural Explications, explicit error taxonomy, and evaluation beyond surface metrics.

Beyond benchmark coverage, evaluation metrics themselves remain limited for culture-loaded translation. Standard metrics such as BLEU and ChrF mainly reward surface lexical similarity, while learned metrics such as COMET may favor conventional phrasing and penalize stylistic variation (Freitag et al., 2022; Mukherjee et al., 2025). As a result, translations may preserve surface form while

Benchmark	Multiple Expression Types	Culture-Specific Meaning	Figurative or Non-Literal Meaning	Context-Dependent Interpretation	Explicit Error Taxonomy	Error-Level Analysis	Evaluation Beyond Surface Metrics
CHENGYU-BENCH	✗	✓	✓	✗	✗	✗	✗
IdioTS	✗	✗	✓	✓	✗	✗	✗
SlangDIT	✗	✓	✓	✓	✗	✗	✓
CAMT	✓	✓	✗	✓	✗	✗	✓
IdiomEval	✗	✓	✓	△	✓	✓	✓
MENT/RATE	✓	△	✓	✓	✗	△	✓
CulT-Eval	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparison of CulT-Eval with existing culture-related and non-literal translation benchmarks. ✓: supported, ✗: not supported, △: partially supported.

failing to convey idiomatic, implicit, or culturally grounded meaning (Shafayat et al., 2025; Wicks and Post, 2023; Baziotis et al., 2023). Although recent work has introduced context-sensitive challenge sets and phenomenon-specific metrics for idiomaticity, metaphor, and cultural references (Wang et al., 2024; Yao et al., 2024), these efforts still lack a unified structure for taxonomy-driven analysis of culturally induced meaning deviations.

To address these gaps, we introduce CulT-Eval, a benchmark centered on Chinese culture-loaded expressions, which encode rich historical, literary, social, and institutional knowledge. CulT-Eval contains 7,959 curated instances with cultural span annotations, Cultural Explications, and diagnostic error labels, and evaluates how such meanings are realized in English and Japanese target outputs. This design supports fine-grained analysis of whether culturally salient meaning is preserved, distorted, omitted, or pragmatically shifted in translation.

Using CulT-Eval, we evaluate neural machine translation systems and large language models, revealing systematic failure patterns that are not adequately captured by existing automatic metrics. Building on this finding, we propose CURE, a Cultural Explication-grounded metric that separates cultural validity from target-language realization quality. Together, CulT-Eval and CURE provide a framework for diagnosing culturally grounded translation failures beyond sentence-level similarity.

- We present CulT-Eval, a benchmark centered on Chinese culture-loaded expressions, comprising 7,959 instances with English/Japanese target realizations, structured annotations, and diagnostic error labels.
- We propose CURE, a Cultural Explication-grounded metric that separates cultural validity from target-language realization quality and aligns more reliably with human judgments than standard MT metrics.
- We evaluate neural machine translation systems and large language models on CulT-Eval, revealing systematic failure patterns in preserving culturally grounded meaning across English and Japanese targets.

2 Related Work

Culture-loaded Expressions in Machine Translation. Recent work has increasingly studied culture-loaded and non-literal translation, leading to specialized benchmarks for idioms (Li et al., 2023), proverbs (Wang et al., 2025), slang (Liang et al., 2025), social-media expressions (Guo et al., 2025), classical poetry (Chen et al., 2025), culture-specific items (Yao et al., 2024), and domain-specific cultural terminology such as recipes (Cao et al., 2024). More recently, Tian et al. (2026) introduce MENT and RATE to evaluate non-literal translation beyond surface mapping. Despite this progress, existing resources remain fragmented across cultural domains, registers, phenomena, and evaluation criteria, limiting cross-benchmark comparability and cumulative analysis. These limitations motivate a unified taxonomy for characterizing culture-loaded expressions and aligning evaluation across domains and phenomena.

Evaluation Metrics in Machine Translation. Cultural evaluation is still often conducted with general-purpose MT metrics that capture surface overlap or holistic semantic similarity, including BLEU (Papineni et al., 2002), ChrF++ (Popović, 2017), BERTScore (Zhang et al., 2019), COMET (Rei et al., 2020), and QE-based variants (Rei et al., 2022; Juraska et al., 2023; Guerreiro et al., 2024). However, cultural meaning shifts often involve dimension-specific failures, such as referent mismatch, loss of allusive meaning, or register and socio-pragmatic mismatch, which are not reliably reflected by aggregate scores (Yang et al., 2025b; Cheng et al., 2025). Overlap-based

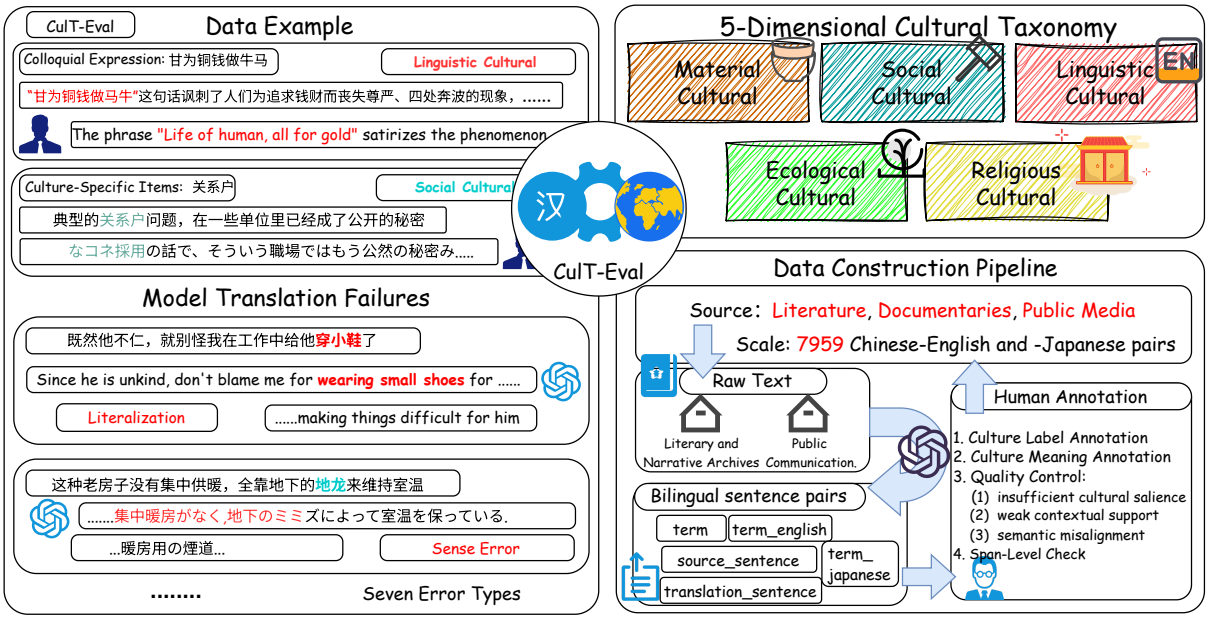

Figure 2: Overview of the CulT-Eval benchmark.

metrics may under-reward legitimate paraphrases or localization choices, while learned metrics may still fail to identify which cultural dimension is violated (Proietti et al., 2025; Tian et al., 2026). These limitations motivate taxonomy-grounded and dimension-aware evaluation for fine-grained assessment of cultural meaning preservation.

3 CulT-Eval Benchmark

3.1 Data Source

Literary and Narrative Archives. We collect source-target excerpts from regional literature, folklore chronicles, and movie subtitles. This domain primarily contributes idioms, slang, colloquial expressions, and Ecological/Material CSIs, many of which require interpretation beyond literal translation and may lack fixed target-language equivalents.

Public and Institutional Communication. We also aggregate official publicity materials, news reports, and documentaries. This domain provides Social/Political CSIs, statutory terms, and formalized culture-specific expressions, where translations are expected to preserve communicative precision and conform to established target-language conventions.

3.2 Cultural Taxonomy

To support fine-grained diagnostic evaluation, each instance in CulT-Eval is assigned a primary cultural category. We adopt a five-way taxonomy adapted from translation studies and intercultural communication (Newmark, 1988; Aixela, 1999): **Material Culture**, covering artifacts and architectural forms (e.g., *Majiazi* [马架子]); **Social Culture**, covering sociopolitical systems and institutional roles (e.g., *Red Tourism* [红色旅游]); **Linguistic Culture**, covering idioms, proverbs, and metaphors (e.g., *Three cobblers* [三个臭皮匠]); **Religious Culture**, covering beliefs, rituals, and philosophical frameworks (e.g., Confucianism); and **Ecological Culture**, covering seasonal, calendrical, and geographical concepts (e.g., *Grain Rain* [谷雨]). For diagnostic clarity, each instance receives a single primary label according to its predominant contextual function.

Metric	Pearson r	Spearman ρ
BLEU	0.30	0.28
ChrF++	0.22	0.20
BERTScore	0.28	0.25
COMET	0.45	0.39
MetricX-QE	0.25	0.23
CURE (Ours)	**0.68**	**0.65**

Table 2: Pearson (r) and Spearman (ρ) correlation coefficients between metrics and human annotations on a 1000 instances subset.

3.3 Benchmark Construction Pipeline

As shown in Fig. 2, CulT-Eval is constructed through a semi-automated pipeline combining LLM-assisted candidate mining with human annotation. We use GPT-5 with domain-adapted

Error Type	Lexical		Semantic	
	BLEU	ChrF	BERT	COMET
Omission	✓	✓	✓	✓
Literalization	✗	✗	✗	✗
Neutralization	✗	✗	△	△
Over-interpretation	✗	✗	✗	✗

Table 3: Metric sensitivity analysis. (✓: Sensitive; ✗: Insensitive; △: Partial)

prompts to mine Chinese sentences likely to contain culture-loaded expressions, prioritizing recall over precision. Final instance acceptance and all benchmark annotations are human verified. Additional robustness checks are provided in App B.1

Trained annotators review the candidates, remove culturally irrelevant or ambiguous sentences, identify the minimal source-side cultural span and its corresponding target-language realization span, and assign a taxonomy label. For each retained instance, they further provide a verified target-language reference, a one-sentence Cultural Explication, and, when applicable, a standardized target-language equivalent or realization strategy.

3.4 Dataset Statistics and Quality Control

After annotation, we applied post hoc filtering to ensure consistency and interpretability. Annotators refined span boundaries, normalized terminology, and removed instances with insufficient cultural salience, weak contextual support, or semantic misalignment between the Chinese source and the target-language reference.

Target-language references were drawn from the original source-target materials, including subtitles, literary translations, and official publications, and were manually checked for pragmatic adequacy and cultural faithfulness. Instances were excluded if the target side relied mainly on unglossed transliteration or failed to convey the intended meaning of the culture-loaded expression. For Japanese references, annotators additionally checked whether retained Chinese-character forms and Sino-Japanese cognates preserved the intended meaning in context.

We also annotate target-language realization spans for each culture-loaded expression, yielding explicit mappings between Chinese source spans and their English or Japanese realizations. After filtering, CulT-Eval contains 7,959 instances across idioms, slang, colloquialisms, literary expressions, and entity-based culture-specific items, as summarized in Table 8.

4 Diagnosing the Limits of Standard MT Metrics

4.1 Evaluation Setup with Standard Metrics

We evaluate representative translation systems on CulT-Eval, covering both multilingual NMT models and instruction-tuned LLMs. The MT baselines include NLLB-200-3.3B (Team et al., 2022), Hunyuan-7B (Zheng et al., 2025), and MADLAD-400-10B (Kudugunta et al., 2023). The LLM baselines include GPT-5.1 (Singh et al., 2025), Llama-3.1-Instruct (Grattafiori et al., 2024), Seed-X-PPO-7B (Cheng et al., 2025), Qwen3-Instruct (Yang et al., 2025a), Gemini-3.0-Pro [1]and DeepSeek-V3 (DeepSeek-AI et al., 2025). All systems translate directly from Chinese into the specified target language, English or Japanese, without access to Cultural Explications or reference translations.

For LLMs, we evaluate both zero-shot and one-shot prompting. The one-shot example is selected from a held-out development pool, excluded from all evaluation instances, and used only to demonstrate the desired output format. NMT systems are evaluated under their standard inference settings without prompt variations.

We report BLEU, ChrF++, BERTScore, COMET, and MetricX-QE at the sentence level, computed separately for the Chinese–English and Chinese–Japanese subsets.

4.2 Standard Metrics under Cultural Evaluation

Table 4 shows that sentence-level metrics distinguish systems and reveal clear performance differences across target languages. However, they do not directly assess whether annotated culture-loaded spans are correctly realized. We therefore measure metric-human alignment using cultural correctness, defined as whether the annotated span is faithfully realized in the target-language output rather than whether the full sentence matches the reference surface form.

As shown in Table 2, BLEU, ChrF++, BERTScore, and COMET exhibit weak and unstable correlations with human cultural-correctness judgments. Thus, higher sentence-level scores do not reliably indicate correct translation of culture-loaded spans.

[1] https://deepmind.google/models/model-cards/

Model	Target	BLEU (↑)		ChrF++ (↑)		BERTScore (↑)		COMET (↑)		MetricX-QE (↓)	
		0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot
Machine Translation Models											
Hunyuan-MT-7B	Zh–En	20.81	–	47.92	–	0.82	–	0.72	–	4.15	–
	Zh–Ja	23.50	–	50.12	–	0.84	–	0.75	–	3.90	–
Madlad400-10B-MT	Zh–En	16.47	–	36.59	–	0.79	–	0.61	–	5.50	–
	Zh–Ja	18.25	–	39.40	–	0.81	–	0.63	–	5.20	–
NLLB200-3-3B	Zh–En	15.69	–	35.81	–	0.78	–	0.59	–	5.85	–
	Zh–Ja	17.80	–	38.95	–	0.80	–	0.62	–	5.45	–
Seed-X-PPO-7B	Zh–En	23.15	–	48.60	–	0.84	–	0.75	–	3.85	–
	Zh–Ja	25.80	–	51.20	–	0.86	–	0.78	–	3.60	–
Open-Sourced and Proprietary LLMs											
Llama-3.1-8B-Instruct	Zh–En	19.45	19.68	42.16	42.55	0.83	0.84	0.74	0.75	3.95	3.88
	Zh–Ja	21.60	21.95	44.80	45.10	0.85	0.86	0.77	0.78	3.65	3.60
DS-R1-D-Qwen-7B	Zh–En	18.37	9.12	37.62	28.15	0.65	0.68	0.55	0.48	6.10	7.25
	Zh–Ja	20.15	10.45	40.50	31.20	0.71	0.67	0.58	0.50	5.90	6.80
Qwen3-8B-Instruct	Zh–En	25.44	25.85	47.80	48.12	0.86	0.87	0.79	0.80	2.85	2.80
	Zh–Ja	28.15	28.60	51.30	51.75	0.88	0.89	0.82	0.83	2.65	2.58
Qwen3-32B-Instruct	Zh–En	28.74	29.05	50.61	50.95	0.88	0.89	0.85	0.86	2.15	2.10
	Zh–Ja	31.40	31.85	54.10	54.55	0.90	0.91	0.87	0.88	2.05	1.98
DeepSeek-V3	Zh–En	***31.32***	***31.55***	***53.43***	***53.85***	0.91	0.92	***0.89***	***0.90***	***1.55***	***1.52***
	Zh–Ja	***34.60***	***34.95***	***57.20***	***57.50***	***0.93***	***0.94***	***0.91***	***0.92***	***1.45***	***1.42***
GPT-5.1	Zh–En	29.34	25.68	51.39	50.82	**0.92**	**0.93**	0.87	0.88	1.62	1.68
	Zh–Ja	32.40	28.50	55.80	54.15	**0.94**	**0.95**	0.90	0.89	1.48	1.55
Gemini-3.0	Zh–En	**31.85**	**32.10**	**54.10**	**54.60**	**0.92**	**0.93**	**0.90**	**0.91**	**1.50**	**1.48**
	Zh–Ja	**35.15**	**35.40**	**58.00**	**58.40**	**0.94**	**0.95**	**0.92**	**0.93**	**1.40**	**1.38**

Table 4: Comprehensive evaluation results on CulT-Eval across both target-language settings. Bold green and underlined blue denote the best and second-best results, respectively.

Type	Content Analysis	Score	Valid
High *Error*	**SRC** ... 强调自己吃软不吃硬... **REF** ... *amenable to coaxing but not* ... **SYS** ... better to persuasion than ...	32.1 0.83	✗
Low *Good*	**SRC** ... 典型的关系户问题，...。 **REF** ... 典型的な縁故採用の問題で... **SYS** ... コネで採用された人をめぐる典型的な問題で...	18.7 0.61	✓

Table 5: Examples of metric–validity misalignment. High Error is high-score invalid; Low Good is low-score valid.

Representative examples in Table 5 further illustrate this misalignment. Some translations receive high sentence-level scores despite literalization, over-generalization, or incorrect sense selection, while culturally valid paraphrases or target-language-appropriate reformulations may receive low scores due to surface divergence from the reference. These results show that sentence-level similarity is an unreliable proxy for cultural correctness and motivate a finer-grained analysis of culture-loaded translation errors.

4.3 Error Taxonomy of Cultural Meaning Distortions

Building on the metric–human misalignment observed in Section 4.2, we conduct a structured error analysis to characterize how culturally salient meanings fail in translation. We introduce a taxonomy of cultural meaning distortions that distinguishes recurrent failure modes and provides a diagnostic framework beyond conventional sentence-level metrics. Since multiple cultural translation errors may co-occur in a single output, each invalid translation is assigned one primary error label according to a fixed priority order from omission to over-interpretation. This single-label policy supports consistent aggregate analysis by associating each instance with its most fundamental source of cultural failure.

4.3.1 Error Categories.

We identify seven error categories for culture-loaded span translation. **Omission** refers to cases where the culture-loaded span is deleted or not meaningfully realized in the target language. **Literalization** occurs when the translation preserves surface form but fails to convey the conventional, idiomatic, or culturally grounded meaning. **Sense Error** refers to selecting an incorrect sense or referent for the culture-loaded expression. **Neutralization** occurs when the basic meaning is conveyed, but culture-specific force, evaluative meaning, or institutional implication is flattened into a generic expression. **Mis-substitution** refers to replacing

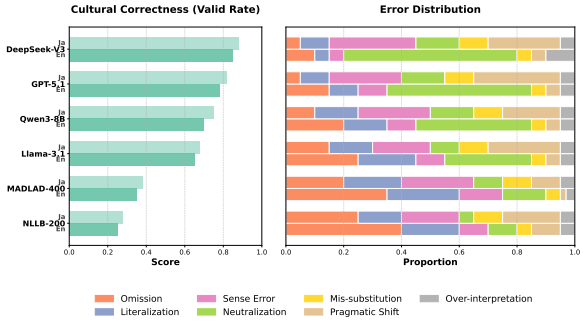

Figure 3: Cross-target cultural correctness and error distribution for selected systems. The left chart reports cultural-validity rates for English and Japanese targets. The right chart shows the language-specific distribution of primary error types among invalid outputs. When multiple errors co-occur, the primary label is assigned according to the priority rule described in Section G.

the source-cultural expression with a misleading or non-equivalent target-culture analogue. **Pragmatic Shift** captures changes in register, politeness, stance, honorific level, or social relation. **Over-interpretation** occurs when the translation adds unsupported cultural explanation, background information, or value judgment. All categories target errors in annotated culture-loaded spans rather than general sentence-level translation errors. Details are provided in Appendix G.

4.3.2 Error Distributions across Systems

We examine cultural correctness and primary error distributions for representative systems under the source-only setting described in Section 4.1. Figure 3 shows that cultural failures differ between the English and Japanese target settings and vary across systems, even when sentence-level translation quality appears competitive. These differences reflect how culture-loaded spans are realized through omission, literalization, neutralization, pragmatic shift, or over-interpretation, rather than surface fluency alone.

4.4 Why Sentence-level Metrics Fall Short

Taken together, the analyses reveal a structural mismatch between standard metrics and cultural correctness. Sentence-level metrics assume that similarity to a reference reflects meaning preservation, but culture-loaded expressions often depend on specific target-language span realizations that may have little effect on overall surface similarity.

Our error taxonomy makes this mismatch explicit. Omission errors are often penalized because content is removed, whereas literalization, neutralization, and over-interpretation may preserve lexical or semantic overlap while distorting culturally salient meaning. As shown in Table 3, this asymmetry is systematic across error categories: standard metrics conflate different cultural failure mechanisms into a single sentence-level score. This motivates an evaluation approach that directly targets the target-language realization of culture-loaded spans.

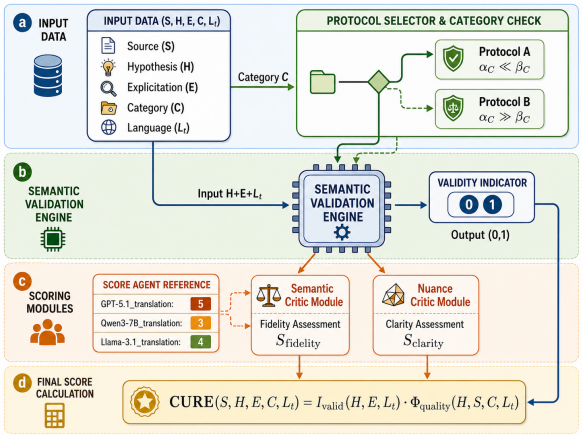

Figure 4: Evaluation pipeline

5 CURE: Cultural Realization Evaluation

We propose CURE (**Cult**ural **R**ealization **E**valuation), an automatic evaluation metric for assessing whether culturally salient meaning is correctly realized in the target language. Unlike standard metrics that rely on surface overlap, CURE is anchored by Cultural Explication annotations, which define the intended meaning of culture-loaded spans and frame cultural evaluation as a structured verification problem.

CURE models cultural realization through two complementary components: **Validity** and **Quality**. Validity determines whether the intended cultural referent is correctly instantiated in the target language, while Quality assesses how appropriately that meaning is expressed for target-language readers. Let S denote the source sentence, H the hypothesis translation, E the associated Cultural Explication, C the category of the culture-loaded span, and L_t the target language. CURE is defined as:

$$\text{CURE}(S, H, E, C, L_t) = \mathbb{I}_{\text{valid}}(H, E, L_t) \cdot \Phi_{\text{quality}}(H, S, C, L_t). \quad (1)$$

Here, $\mathbb{I}_{\text{valid}}(H, E, L_t) \in \{0, 1\}$ is the Semantic Validity Indicator, which verifies whether H realizes the cultural referent defined in E in a way that is appropriate and interpretable for the target

Model	Target	Validity (%)		Fidelity (1–5)		Clarity (1–5)		Final CURE	
		0-shot	1-shot	0-shot	1-shot	0-shot	1-shot	0-shot	1-shot
Machine Translation Models									
Hunyuan-MT-7B	Zh–En	30.5	–	3.85	–	4.10	–	1.21	–
	Zh–Ja	34.2	–	3.92	–	4.15	–	1.35	–
Madlad400-10B-MT	Zh–En	23.2	–	3.66	–	4.00	–	0.89	–
	Zh–Ja	26.5	–	3.75	–	4.08	–	1.05	–
NLLB200-3-3B	Zh–En	12.8	–	3.41	–	3.66	–	0.44	–
	Zh–Ja	15.5	–	3.52	–	3.75	–	0.55	–
Seed-X-PPO-7B	Zh–En	33.0	–	4.05	–	4.15	–	1.45	–
	Zh–Ja	36.5	–	4.12	–	4.22	–	1.62	–
Open-Sourced and Proprietary LLMs									
Llama-3.1-8B-Instruct	Zh–En	28.4	37.8	4.15	4.28	4.25	4.35	1.19	1.63
	Zh–Ja	31.2	40.5	4.22	4.35	4.32	4.42	1.35	1.80
DS-R1-D-Qwen-7B	Zh–En	30.5	22.1	4.50	4.65	4.10	4.55	1.40	0.91
	Zh–Ja	33.8	25.4	4.55	4.70	4.18	4.62	1.58	1.05
Qwen3-8B-Instruct	Zh–En	42.8	45.2	4.30	4.42	4.34	4.40	1.85	1.99
	Zh–Ja	46.5	49.0	4.40	4.52	4.45	4.50	2.10	2.25
Qwen3-32B-Instruct	Zh–En	46.8	49.5	4.44	4.55	4.25	4.45	2.03	2.23
	Zh–Ja	50.5	53.2	4.52	4.65	4.35	4.55	2.25	2.45
DeepSeek-V3	Zh–En	52.4	55.1	4.73	4.82	4.52	4.60	2.42	2.59
	Zh–Ja	**57.5**	**60.5**	4.85	4.90	***4.66***	**4.75**	**2.72**	**2.92**
GPT-5.1	Zh–En	49.6	48.8	4.76	4.80	4.54	4.65	2.31	2.31
	Zh–Ja	54.2	53.5	***4.88***	***4.92***	4.62	4.68	2.55	2.58
Gemini-3.0	Zh–En	51.8	56.0	4.80	4.85	4.55	4.68	2.40	2.65
	Zh–Ja	***56.2***	***59.8***	**4.90**	**4.94**	**4.70**	***4.72***	***2.68***	***2.88***

Table 6: Detailed breakdown of CURE evaluation results.

language L_t. Invalid translations receive a score of zero, preventing fluent but culturally incorrect realizations from being rewarded.

For valid instances, the Quality component is computed as:

$$\Phi_{\text{quality}}(H, S, C, L_t) = \alpha_C \mathcal{S}_{\text{fidelity}}(H, S, L_t) + \beta_C \mathcal{S}_{\text{clarity}}(H, L_t). \quad (2)$$

where $\mathcal{S}_{\text{fidelity}}$ measures preservation of intended meaning and pragmatic force, and $\mathcal{S}_{\text{clarity}}$ measures communicative intelligibility for target-language readers. The target-language variable L_t conditions both validity and quality, since appropriate realization may differ between English and Japanese.

In implementation, CURE uses a fixed LLM-based reference evaluator with deterministic decoding and fixed prompts. Fidelity and Clarity are scored on a 1–5 scale, and the reported CURE score is the average gated quality score over instances. Full prompts and scoring rubrics are provided in Appendix L.

5.1 Category-conditioned Evaluation and Reference Implementation

Based on the taxonomy in Section 3.2, CURE uses category-conditioned protocols without changing the metric formulation. A Category Check routes each instance to one of two protocols according to its category C, while the target language L_t conditions what counts as an appropriate realization.

Protocol A is fact-centric and applies to categories such as cultural concepts, material artifacts, and social institutions. It emphasizes referential validity: the output must preserve the source-cultural referent specified by the Cultural Explication, rather than replacing it with a misleading or non-equivalent target-culture analogue.

Protocol B is style-centric and applies to idioms, literary expressions, and slang. It emphasizes functional equivalence: paraphrase or re-expression is permitted if the pragmatic force and register of the original expression are preserved in the target language.

We instantiate these protocols in CulT-Agent, a reference implementation with three role-based components: a Semantic Validator for Validity, and a Fidelity Critic and Clarity Critic for Quality. Each instance is represented as (S, H, E, C, L_t). Reference translations are used only as stylistic anchors, whereas Cultural Explications serve as the authoritative basis for cultural correctness. Full component descriptions and coordination protocols are provided in Appendix G.

Full CURE	**0.88**	–
w/o Validity Gate ($\mathbb{I}_{\text{valid}}$)	0.76	-0.12
w/o Adaptive Routing	0.81	-0.07
w/o Explication Anchor (E)	0.62	**-0.26**
w/o Reference Anchor (R)	0.85	-0.03

Table 7: Ablation study of CURE.

6 Experiments and Analyses

6.1 Cultural Correctness under CURE

We evaluate CURE using the same translation systems and experimental setup as in Section 4.1. Table 6 reports CURE scores and component breakdowns for representative MT systems and LLMs on CulT-Eval.

The results show that culture-loaded expressions remain difficult for current translation systems. Across both English and Japanese targets, validity failures are frequent, indicating that systems often misinterpret, generalize, or omit culturally salient meanings even when outputs appear fluent. These failures directly lower final CURE scores and expose deficiencies obscured by standard sentence-level metrics.

Stronger LLMs generally achieve higher validity rates than dedicated MT systems, but their improvements are still far from definitive. Substantial variation remains in how systems preserve and express cultural meaning after referential correctness is satisfied. By decomposing cultural correctness into semantic validity and realization quality, CURE offers a more informative view of translation behavior than aggregate metrics that conflate distinct sources of cultural failure.

6.2 Cross-target Analysis

Figure 3 shows that both cultural correctness and error distributions differ between English and Japanese targets. Chinese–English outputs more often exhibit omission and neutralization, producing fluent but culturally flattened translations. Chinese–Japanese outputs more often involve sense errors, pragmatic shifts, and form-preserving confusions, reflecting the risk of shared scripts and culturally proximate but non-identical conventions. These patterns show that culturally proximate translation is not necessarily trivial and motivate target-language-aware evaluation.

6.3 Ablation Study

To assess each CURE component, we conduct an ablation study on translations generated by Qwen3-8B. As shown in Table 7, removing the Validity gate lowers human alignment, confirming its role in filtering fluent but culturally invalid outputs. The largest drop occurs without Cultural Explications, showing that explicit semantic grounding is central to cultural evaluation. By contrast, removing reference translations causes only a marginal decline, indicating that CURE relies more on semantic definitions than surface-form similarity.

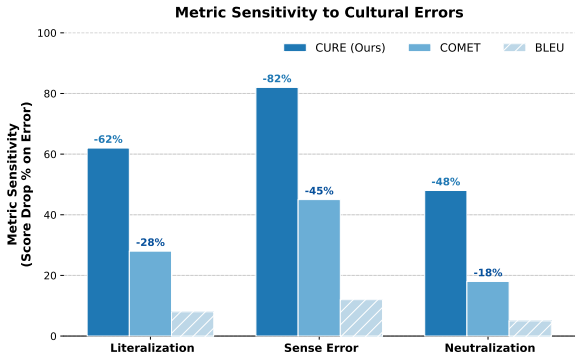

Figure 5: Sensitivity of evaluation metrics.

6.4 Diagnostic Sensitivity to Cultural Errors

Figure 5 compares CURE and standard metrics on representative cultural error types. CURE shows substantial score drops for sense errors, literalization, and neutralization, whereas COMET is only moderately sensitive and BLEU remains largely insensitive. These results suggest that standard metrics may still assign plausible scores to fluent translations that preserve surface similarity but distort the intended cultural meaning. By grounding evaluation in Cultural Explications and target-language realization, CURE captures diagnostic signals missed by sentence-level similarity metrics, explaining the discrepancies discussed above.

7 Conclusion

We introduced CulT-Eval, a benchmark centered on Chinese culture-loaded expressions with English and Japanese target realizations, and CURE, a Cultural Explication-grounded metric for evaluating cultural validity and realization quality. Experiments show that standard sentence-level metrics often miss culturally induced meaning distortions, while CURE better aligns with human judgments and more sensitively captures culture-related errors. Our findings show that cultural meaning preservation remains difficult across both target settings, with distinct failure patterns in English and Japanese.

Limitations

CulT-Eval focuses on Chinese culture-loaded expressions and evaluates their realization in English and Japanese. Although this dual-target design supports comparison between culturally distant and culturally proximate target-language settings, it does not cover all possible cultural transfer scenarios.

Although we include native-speaker target-side validation, we do not claim it as a professional human translation ceiling. A full professional translation baseline remains important for calibrating the absolute CURE score scale and for estimating the remaining gap between current systems and human-level cultural translation.

8 Ethics and Data Statement

CulT-Eval involves culture-loaded expressions whose interpretation may depend on regional background, social norms, political terminology, religious concepts, and target-language conventions. We therefore treat Cultural Explications as context-specific semantic anchors rather than universal cultural definitions. Annotators were instructed to avoid stereotyping, unsupported generalization, and overly prescriptive interpretations, while allowing paraphrase or target-language-appropriate reformulation when the intended meaning was preserved.

Target-side validation was conducted with native English and Japanese participants recruited through Prolific. Participants were compensated according to the estimated completion time, with an expected hourly rate of approximately USD $12 or the equivalent amount in local currency. This validation is used as a target-side acceptability check and should not be interpreted as a professional human translation baseline.

Because the benchmark draws on literary translations, subtitles, documentaries, news reports, and public materials, release is subject to licensing constraints. We will release annotations, cultural spans, Cultural Explications, evaluation scripts, prompts, and system outputs where permitted; for materials that cannot be redistributed, we will provide provenance or reconstruction information when legally permissible. CulT-Eval and CURE should be used as diagnostic tools rather than definitive authorities on cultural correctness.

References

Javier Franco Aixela. 1999. *4 Culture-specific Items in Translation*, pages 52–78. Multilingual Matters, Bristol, Blue Ridge Summit.

Christos Baziotis, Prashant Mathur, and Eva Hasler. 2023. Automatic evaluation and analysis of idioms in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3682–3700, Dubrovnik, Croatia. Association for Computational Linguistics.

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.

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 33019–33036, 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.

Francesca De Luca Fornaciari, Begoña Altuna, Itziar Gonzalez-Dios, and Maite Melero. 2024. A hard nut to crack: Idiom detection with conversational large language models. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 35–44, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

DeepSeek-AI and 1 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yicheng Fu, Zhemin Huang, Liuxin Yang, Yumeng Lu, and Zhongdongming Dai. 2025. CHENGYU-BENCH: Benchmarking large language models for Chinese idiom understanding and use. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2355–2366, Suzhou, China. Association for Computational Linguistics.

Aaron Grattafiori and 1 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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.

Hongcheng Guo, Fei Zhao, Shaosheng Cao, Xinze Lyu, Ziyan Liu, Yue Wang, Boyang Wang, Zhoujun Li, Chonggang Lu, Zhe Xu, and Yao Hu. 2025. Redefining machine translation on social network services with large language models. *Preprint*, arXiv:2504.07901.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. Metricx-23: The google submission to the wmt 2023 metrics shared task. In *Conference on Machine Translation*.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset. *Preprint*, arXiv:2309.04662.

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.

Yunlong Liang, Fandong Meng, Jiaan Wang, and Jie Zhou. 2025. Slangdit: Benchmarking llms in interpretative slang translation. *Preprint*, arXiv:2505.14181.

Ananya Mukherjee, Saumitra Yadav, and Manish Shrivastava. 2025. Why should only high-resource-languages have all the fun? pivot based evaluation in low resource setting. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4779–4788, Abu Dhabi, UAE. Association for Computational Linguistics.

P. Newmark. 1988. *Approaches to Translation*. English language teaching. Prentice Hall.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Lorenzo Proietti, Stefano Perrella, and Roberto Navigli. 2025. Has machine translation evaluation achieved human parity? the human reference and the limits of progress. *Preprint*, arXiv:2506.19571.

Ricardo Rei, José G. C. de Souza, Duarte M. Alves, Chrysoula Zerva, Ana C. Farinha, T. Glushkova, Alon Lavie, Luísa Coheur, and André F. T. Martins. 2022. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Conference on Machine Translation*.

Ricardo Rei, Craig Alan Stewart, Catarina Farinha, and Alon Lavie. 2020. Unbabel's participation in the wmt20 metrics shared task. In *Conference on Machine Translation*.

Sheikh Shafayat, Dongkeun Yoon, Woori Jang, Jiwoo Choi, Alice Oh, and Seohyon Jung. 2025. A 2-step framework for automated literary translation evaluation: Its promises and pitfalls. *Preprint*, arXiv:2412.01340.

Aaditya Singh and 1 others. 2025. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Yanzhi Tian, Cunxiang Wang, Zeming Liu, Heyan Huang, Wenbo Yu, Dawei Song, Jie Tang, and Yuhang Guo. 2026. Beyond literal mapping: Benchmarking and improving non-literal translation evaluation. *Preprint*, arXiv:2601.07338.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, AiTi Aw, and Nancy Chen. 2024. SeaEval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 370–390, Mexico City, Mexico. Association for Computational Linguistics.

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.

Rachel Wicks and Matt Post. 2023. Identifying context-dependent translations for evaluation set production. *Preprint*, arXiv:2311.02321.

An Yang and 1 others. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Cai Yang, Yao Dou, David Heineman, Xiaofeng Wu, and Wei Xu. 2025b. Evaluating llms on chinese idiom translation. *Preprint*, arXiv:2508.10421.

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.

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

Mao Zheng, Zheng Li, Bingxin Qu, Mingyang Song, Yang Du, Mingrui Sun, and Di Wang. 2025. Hunyuan-mt technical report. *Preprint*, arXiv:2509.05209.

A Additional Discussion: Multi-round and Routing-based Translation

Our findings also have implications for multi-round translation, re-translation, and language-routing pipelines. Prior work on language-specific knowledge and multilingual routing suggests that model performance may vary depending on the language or intermediate representation through which information is processed. For culture-loaded expressions, however, the key question is not only whether an intermediate step improves fluency or surface similarity, but whether it helps recover and preserve the intended cultural referent.

In this sense, CulT-Eval provides a diagnostic setting for evaluating such pipelines. Many culture-loaded expressions require contextual reinterpretation rather than direct lexical mapping, and some expressions lack a fully equivalent target-language realization. A multi-round or explanation-first pipeline may therefore help when it reduces literal translation and clarifies the intended cultural meaning, but it may also introduce neutralization or over-explication when the target language lacks a compact equivalent.

CURE is useful for analyzing these cases because it separates semantic validity from realization quality. The Validity component asks whether the intended cultural referent or pragmatic function has been correctly recovered, while the Quality component evaluates whether that meaning is expressed faithfully and clearly in the target language. This decomposition can help evaluate whether routing-based or re-translation pipelines genuinely improve cultural translation, rather than merely increasing sentence-level similarity or fluency.

Category	Count	Ratio (%)	Avg. Term Len. (Char)			Avg. Context
			Source (Zh)	Target (En)	Target (Ja)	Source (Zh)
Linguistic	2,512	31.6	5.13	29.43	6.50	22.29
Social	2,399	30.1	3.21	24.22	5.20	23.72
Material	1,594	20.0	3.06	22.97	4.80	24.39
Ecological	833	10.5	2.82	20.13	4.50	22.66
Religious	621	7.8	3.13	23.01	5.10	24.22
Total / Avg.	**7,959**	**100.0**	**3.74**	**25.09**	**5.45**	**23.33**

Table 8: Statistics of the CulT-Eval benchmark .

B Additional Experiments

B.1 Benchmark Independence and Robustness Checks

This appendix provides additional robustness checks addressing the independence of CulT-Eval and CURE. There are two potential sources of dependence in our pipeline. First, GPT-5 is used during LLM-assisted candidate mining, while GPT-family models are also included among evaluated systems. Second, CURE is implemented with an LLM-based evaluator, which may raise concerns about judge-specific stylistic preferences. We therefore conduct two targeted checks: an alternative-judge evaluation and a human-mined subset evaluation.

Importantly, GPT-5 is used only for high-recall candidate mining. Final instance acceptance, source-side span selection, target-language realization span selection, Cultural Explication formulation, and post hoc filtering are all conducted through human annotation and verification. Thus, the mining model does not determine benchmark labels, reference targets, or evaluation decisions.

Alternative-judge robustness. To examine whether CURE results depend on a single LLM evaluator, we re-evaluate a stratified subset of 500 instances using an alternative judge from an independent model family. The subset is stratified by cultural category, target language, and system output. Table 9 compares system-level CURE scores under the original judge and the alternative judge. The resulting system rankings remain highly correlated, suggesting that the main conclusions are not an artifact of one evaluator's stylistic preferences.

Human-mined subset robustness. To examine whether LLM-assisted candidate mining gives GPT-family systems an implicit advantage, we additionally construct a human-mined subset whose candidate instances are selected without GPT-based mining. This subset follows the same annotation, filtering, and evaluation protocol as the main benchmark.

System	Orig.	Alt.	Δ
DeepSeek-V3	2.59	2.55	-0.04
GPT-5.1	2.31	2.26	-0.05
Qwen3-32B-Instruct	2.23	2.28	+0.05
Hunyuan-MT-7B	1.21	1.19	-0.02
Rank corr.		0.80	

Table 9: Alternative-judge robustness on a stratified subset. Orig. and Alt. denote average CURE scores under the original Grok-based evaluator and the alternative Claude-4 judge, respectively.

System	LLM	Human	Δ
DeepSeek-V3	2.62	2.58	-0.04
GPT-5.1	2.34	2.30	-0.04
Hunyuan-MT-7B	1.23	1.18	-0.05
Rank corr.		1.00	

Table 10: Robustness on a human-mined subset. LLM and Human denote the LLM-mined and human-mined subsets, respectively; the latter is constructed without GPT-based candidate extraction.

Table 10 compares rankings on the LLM-mined and human-mined subsets. The ranking trend remains stable, indicating that the observed system differences are not driven by construction-time exposure to GPT-generated candidates.

Stratified robustness. We further examine whether ranking stability holds across cultural categories and difficult cases. The hard subset contains instances with low baseline cultural correctness, including idiomatic, literary, and highly context-dependent expressions. As shown in Table 11, the main ranking trends remain stable across categories and hard cases, although smaller categories naturally yield slightly lower correlations due to limited sample size.

B.2 Annotation Reliability

To assess the reliability of our annotation scheme, we conduct an inter-annotator agreement analysis on a double-annotated subset of 500 instances. The analysis covers two annotation tasks that are central to CulT-Eval: (i) binary cultural validity judgments, which determine whether the culture-loaded expression is correctly realized in the target-language output, and (ii) the seven-way primary error taxonomy, which characterizes the main type of cultural meaning distortion when the translation is invalid.

Table 13 reports both raw agreement and chance-corrected agreement. For validity judgments, annotators achieve 86.2% raw agreement, with Cohen's $\kappa = 0.72$ and Krippendorff's $\alpha = 0.71$, indicating substantial agreement. For the seven-way error taxonomy, annotators achieve 78.4% raw agreement, with Cohen's $\kappa = 0.68$ and Krippendorff's $\alpha = 0.67$. These results show that the binary validity labels and fine-grained error labels are consistently reproducible beyond chance.

Subset	#Inst.	Alt.	Human
Linguistic	158	0.95	0.93
Social	151	0.94	0.92
Material	100	0.90	0.94
Ecological	52	0.88	0.85
Religious	39	0.85	0.82
Hard	150	0.92	0.91
Overall	500	0.94	0.92

Table 11: Stratified robustness checks across cultural categories and hard cases. Alt. and Human denote Spearman rank correlations under the alternative-judge and human-mined settings, respectively. The top-2 systems remain stable across all subsets.

We further analyze agreement for each category in the seven-way error taxonomy. As shown in Table 14, errors such as Omission, Literalization, and Sense Error show relatively high agreement, reflecting their clearer decision boundaries. Lower agreement is observed for Neutralization and Pragmatic Shift, where the distinction depends on whether the translation primarily loses culture-specific content or changes the social, pragmatic, or register-level force of the expression. These disagreements are therefore concentrated in semantically adjacent categories rather than distributed arbitrarily across the taxonomy.

This pattern suggests that the proposed taxonomy is broadly reliable, while also identifying the most challenging boundary cases. In subsequent annotation rounds, disagreements involving Neutralization and Pragmatic Shift were reviewed through adjudication, with priority given to the most fundamental source of cultural failure.

B.3 Sensitivity to Fidelity–Clarity Weights

CURE computes the Quality component as a weighted combination of Fidelity and Clarity under category-conditioned evaluation protocols. Although the default weights are chosen to reflect the different evaluation priorities of fact-centric and style-centric cultural expressions, the main conclusions should not depend on a single numerical configuration. We therefore conduct a sensitivity analysis over alternative fidelity–clarity weight settings. Each pair denotes the weights assigned to Fidelity and Clarity, respectively.

Model	Target	Final CURE (95% CI)	
		0-shot	1-shot
Machine Translation Models			
NLLB200-3-3B	Zh–En	0.44 ± 0.08	–
	Zh–Ja	0.55 ± 0.09	–
Madlad400-10B-MT	Zh–En	0.89 ± 0.08	–
	Zh–Ja	1.05 ± 0.08	–
Hunyuan-MT-7B	Zh–En	1.21 ± 0.07	–
	Zh–Ja	1.35 ± 0.09	–
Seed-X-PPO-7B	Zh–En	1.45 ± 0.08	–
	Zh–Ja	1.62 ± 0.08	–
Open-Sourced and Proprietary LLMs			
Llama-3.1-8B-Ins	Zh–En	1.19 ± 0.08	1.63 ± 0.09
	Zh–Ja	1.35 ± 0.07	1.80 ± 0.09
DS-R1-D-Qwen-7B	Zh–En	1.40 ± 0.09	0.91 ± 0.08
	Zh–Ja	1.58 ± 0.08	1.05 ± 0.09
Qwen3-8B-Ins	Zh–En	1.85 ± 0.07	1.99 ± 0.08
	Zh–Ja	2.10 ± 0.09	2.25 ± 0.08
Qwen3-32B-Ins	Zh–En	2.03 ± 0.08	2.23 ± 0.07
	Zh–Ja	2.25 ± 0.08	2.45 ± 0.08
GPT-5.1 *(Base)*	Zh–En	2.31 ± 0.07	2.31 ± 0.07
	Zh–Ja	2.55 ± 0.06	2.58 ± 0.07
Gemini-3.0	Zh–En	$2.40^{*} \pm 0.07$	$\mathbf{2.65}^{\dagger} \pm 0.07$
	Zh–Ja	$2.68^{*} \pm 0.07$	$2.88^{\dagger} \pm 0.07$
DeepSeek-V3	Zh–En	$\mathbf{2.42}^{\dagger} \pm 0.07$	$2.59^{\dagger} \pm 0.07$
	Zh–Ja	$\mathbf{2.72}^{\dagger} \pm 0.07$	$\mathbf{2.92}^{\dagger} \pm 0.06$

Table 12: Statistical validation of Final CURE scores. Values following $\pm$ denote the 95% bootstrap confidence intervals (resampled over 1,000 iterations). Pairwise significance tests for the top-tier LLMs are performed against the GPT-5.1 baseline: * denotes $p < 0.05$ and † denotes $p < 0.01$. The highest score in each specific target and prompting setting is highlighted in bold.

Specifically, we compare the default setting with three variants: a fidelity-heavy configuration, a clarity-heavy configuration, and a uniform configuration. As shown in Table 15, the resulting CURE scores vary only slightly across configurations. More importantly, the system ranking remains stable, with Spearman rank correlations ranging from 0.98 to 1.00 against the default ranking. The same top systems are preserved across all tested settings, indicating that the main findings are not an artifact of a particular choice of fidelity–clarity weights.

Task	Agr.	κ	α
Validity Judgment	86.2%	0.72	0.71
Error Taxonomy	78.4%	0.68	0.67

Table 13: Inter-annotator agreement on 500 double-annotated instances. Both tasks show substantial agreement. Agr., κ, and α denote raw agreement, Cohen's κ, and Krippendorff's α, respectively.

Error	Sup.	κ	Main Disagr.
Omission	42	0.88	Literalization
Literalization	86	0.75	Sense Error
Sense Error	95	0.71	Literalization
Mis-subst.	58	0.69	Sense Error
Over-interp.	31	0.74	Prag. Shift
Neutralization	115	0.62	Prag. Shift
Prag. Shift	73	0.58	Neutralization

Table 14: Inter-annotator agreement and primary disagreement patterns for the seven-way error taxonomy. Sup. denotes support; Main Disagr. denotes the most frequent disagreement label. Mis-subst., Over-interp., and Prag. Shift denote Mis-substitution, Over-interpretation, and Pragmatic Shift, respectively.

C Category-wise Analysis

To further examine whether the overall findings are driven by particular cultural categories, we report category-wise CURE results under the five-way taxonomy introduced in Section 3.2. For each category, we compute Validity and final CURE scores separately for English and Japanese outputs. This analysis is intended as a diagnostic breakdown rather than a separate benchmark setting.

Table C reports the category-wise results. The breakdown shows whether the main trends hold across different cultural phenomena and whether category-level difficulty varies between the English and Japanese settings. Smaller categories, especially Religious and Ecological Culture, should be interpreted with caution due to their lower support.

C.1 Human-Metric Alignment Subset and Resampling Stability

To evaluate the alignment between automatic metrics and human cultural correctness judgments, we construct a human–metric alignment subset of 1,000 instances. The subset is sampled using a stratified protocol over the five cultural categories while preserving the original length distribution of CulT-Eval. This sampling strategy is designed to avoid concentrating the analysis on short, obvious, or clear-cut examples, and to ensure that the correlation results reflect a broad range of culture-loaded translation phenomena.

Weights	DS-V3	GPT-5.1	Qwen3	Llama
(.7,.3) / (.4,.6)	2.59	2.31	2.23	1.19
(.8,.2) / (.5,.5)	2.61	2.32	2.25	1.18
(.6,.4) / (.3,.7)	2.57	2.29	2.21	1.20
(.5,.5) / (.5,.5)	2.63	2.35	2.28	1.22

Table 15: Sensitivity of CURE scores to alternative fidelity–clarity weights. Weights are shown as A/B, where A and B denote the fact-centric and style-centric protocols, respectively; each pair gives (fidelity, clarity). The top system ranking remains stable across settings.

Category	#Inst.	En		Ja	
		Val.	CURE	Val.	CURE
Linguistic	2,512	42.5	1.78	46.8	2.05
Social	2,399	51.2	2.25	55.4	2.52
Material	1,594	65.8	3.02	71.5	3.38
Ecological	833	56.4	2.45	61.2	2.78
Religious	621	53.7	2.31	58.6	2.64

Table 16: Category-wise CURE results across English and Japanese targets. Val. denotes validity rate (%), and CURE denotes the average gated quality score. Scores are averaged over top-tier LLMs.

To further examine whether the reported CURE–human correlation depends on a particular sample, we conduct three additional rounds of independent stratified resampling. Each resample contains 500 instances drawn using the same stratification principle. As shown in Table 17, CURE maintains stable correlations with human judgments across resamples, with mean Pearson correlation 0.680 ± 0.003 and mean Spearman correlation 0.650 ± 0.004. This indicates that the reported alignment is not driven by a favorable or unusually easy subset.

D Human Annotation Guidelines and Quality Control

The construction of the CulT-Eval benchmark involved a dedicated team of target-language bilingual annotators with expertise in Chinese-English and Chinese-Japanese translation. All source-side cultural span and category annotations were conducted by native Chinese speakers trained in the proposed taxonomy. Target-side realization spans and reference adequacy checks were conducted by annotators with advanced proficiency in the corresponding target language. Japanese instances were additionally reviewed for the conventionality of retained Chinese-character forms, Sino-Japanese cognate meaning, and register or pragmatic adequacy.

Subset	Pearson r	Spearman ρ
Original Subset	0.684	0.651
Resample 1	0.679	0.648
Resample 2	0.682	0.655
Resample 3	0.676	0.645
Mean $\pm$ Std.	0.680 ± 0.003	0.650 ± 0.004

Table 17: Stability of CURE–human correlation under independent stratified resampling.

D.1 Annotation Workflow

The annotation process followed a multi-stage pipeline designed to minimize subjectivity through collective verification. Initially, each sentence pair identified during the LLM-assisted candidate mining phase was independently screened to verify the presence of genuine culture-loaded expressions; instances lacking sufficient contextual grounding were discarded. For all validated instances, annotators identified the minimal source-side cultural span and the corresponding target-language realization span in English or Japanese. Category labels were assigned according to the source-side cultural function of the Chinese expression. Cultural Explications were then synthesized as concise, factual descriptions of the culturally grounded meaning required for faithful target-language realization.

D.2 Cultural Correctness and Error Annotation

To facilitate the error analysis presented in 3.2, annotators assessed whether system translations successfully instantiated the cultural referent described in the human-authored explications. Notably, cultural correctness was determined by semantic faithfulness to these explications rather than surface-level lexical similarity to the reference translations. When a translation was deemed incorrect, it was assigned a primary label from a seven-category typology. To handle cases where multiple error phenomena co-occurred, we implemented a fixed hierarchical priority order—moving from Omission and Literalization to Sense Error, Neutralization, Mis-substitution, Pragmatic Shift, and finally Over-interpretation. Annotators were instructed to assign the earliest applicable category in this sequence, thereby ensuring a deterministic and interpretable mapping of cultural failures.

E Target-side Human Validation

To provide an independent target-side check, we conduct human validation studies for both En-

Model	Target	Zero-Shot Setting		One-Shot Setting	
		Final CURE	95% Bootstrap CI	Final CURE	95% Bootstrap CI
Machine Translation Models					
NLLB200-3-3B	Zh–En	0.44	[0.36, 0.51]	–	–
	Zh–Ja	0.55	[0.47, 0.64]	–	–
Madlad400-10B-MT	Zh–En	0.89	[0.82, 0.97]	–	–
	Zh–Ja	1.05	[0.96, 1.12]	–	–
Hunyuan-MT-7B	Zh–En	1.21	[1.13, 1.28]	–	–
	Zh–Ja	1.35	[1.26, 1.45]	–	–
Seed-X-PPO-7B	Zh–En	1.45	[1.38, 1.53]	–	–
	Zh–Ja	1.62	[1.54, 1.70]	–	–
Open-Sourced and Proprietary LLMs					
Llama-3.1-8B-Instruct	Zh–En	1.19	[1.12, 1.27]	1.63	[1.55, 1.72]
	Zh–Ja	1.35	[1.28, 1.42]	1.80	[1.71, 1.88]
DS-R1-D-Qwen-7B	Zh–En	1.40	[1.31, 1.48]	0.91	[0.83, 0.99]
	Zh–Ja	1.58	[1.50, 1.66]	1.05	[0.96, 1.13]
Qwen3-8B-Instruct	Zh–En	1.85	[1.79, 1.92]	1.99	[1.91, 2.06]
	Zh–Ja	2.10	[2.02, 2.19]	2.25	[2.18, 2.33]
Qwen3-32B-Instruct	Zh–En	2.03	[1.95, 2.10]	2.23	[2.16, 2.31]
	Zh–Ja	2.25	[2.18, 2.33]	2.45	[2.37, 2.52]
GPT-5.1 *(Baseline)*	Zh–En	2.31	[2.24, 2.37]	2.31	[2.25, 2.38]
	Zh–Ja	2.55	[2.49, 2.61]	2.58	[2.51, 2.65]
Gemini-3.0	Zh–En	2.40^{*}	[2.33, 2.46]	$\mathbf{2.65}^{\dagger}$	[2.58, 2.71]
	Zh–Ja	2.68^{*}	[2.62, 2.75]	$2.88^{\dagger}$	[2.81, 2.95]
DeepSeek-V3	Zh–En	$\mathbf{2.42}^{\dagger}$	[2.35, 2.48]	$2.59^{\dagger}$	[2.52, 2.65]
	Zh–Ja	$\mathbf{2.72}^{\dagger}$	[2.64, 2.79]	$\mathbf{2.92}^{\dagger}$	[2.86, 2.99]

Table 18: Statistical validation of Final CURE scores across all evaluated systems. The 95% confidence intervals (CIs) are computed via bootstrap resampling over 1,000 iterations. For the top-tier LLMs, pairwise significance tests are performed against the GPT-5.1 baseline: * denotes $p < 0.05$ and † denotes $p < 0.01$. The highest score in each specific target and prompting setting is highlighted in bold.

Stage	Removed	Remaining
Raw LLM candidates	–	12,000
Low cultural salience	2,041	9,959
Weak context	1,186	8,773
Semantic mismatch	814	7,959
Final benchmark	**4,041**	**7,959**

Table 19: Filtering statistics during CulT-Eval construction.

glish and Japanese target-language outputs. Participants were recruited through Prolific. We recruited native or near-native speakers of the corresponding target language: English validators for Chinese–English outputs and Japanese validators for Chinese–Japanese outputs. Participants were compensated at an estimated hourly rate of USD $12, or the equivalent amount in local currency, based on the expected completion time of the validation task.

Validators were shown the Cultural Explication and the target-language output, but not the model identity. They were asked to judge whether the output naturally and faithfully conveyed the intended cultural meaning in the target language. This validation is designed as an independent target-side acceptability check. It does not replace the original source-culture annotation, since CulT-Eval is anchored in Chinese culture-loaded expressions and their intended source-side meanings. It should also not be interpreted as a professional human translation baseline or a certified human translation ceiling.

Overall, the target-side validation provides an additional check that the original cultural-validity labels are broadly consistent with judgments from target-language readers. Lower agreement in Lin-

Subset	#Ex.	En Agr.	Ja Agr.
Overall	2000	82.4%	83.6%
Linguistic	734	78.2%	81.2%
Social	586	83.3%	82.1%
Material	427	87.6%	88.1%
Eco./Rel.	253	83.8%	86.6%

Table 20: Target-side human validation on stratified English and Japanese examples. En Agr. and Ja Agr. denote agreement with the original cultural-validity labels by native English and Japanese speakers, respectively. Validators were recruited through Prolific and compensated at an estimated hourly rate of USD $12 or its local-currency equivalent.

guistic Culture is expected, as idioms, slang, and non-literal expressions often admit multiple acceptable target-language realizations. For Japanese, we additionally inspect cases involving retained Chinese-character forms, Sino-Japanese cognates, and register-sensitive expressions, since these are common sources of subtle target-language disagreement.

F Cultural Taxonomy Definitions

CulT-Eval adopts a five-way cultural taxonomy to characterize the primary cultural dimension involved in each culture-loaded expression. The taxonomy is adapted from prior work in translation studies and intercultural communication (Newmark, 1988; Aixela, 1999). Each instance is assigned one primary category according to the predominant cultural function of the annotated span in context. When an expression could plausibly belong to multiple categories, annotators select the category that is most necessary for interpreting the intended meaning of the source sentence.

Material Culture. Material Culture refers to tangible cultural artifacts, physical objects, traditional technologies, food, clothing, architecture, tools, housing forms, and other concrete items that are culturally specific. These expressions often require more than literal object translation because their meaning depends on how the object is used, recognized, or socially situated within the source culture. Examples include traditional dwellings, local food items, regional clothing, agricultural tools, and architectural structures. For instance, *Majiazi* [马架子] refers to a log shelter or temporary wooden structure with a specific regional and material-cultural function, while *dilong* [地龙] in traditional architecture can refer to an underfloor heating flue system rather than its literal surface meaning.

Social Culture. Social Culture covers expressions related to social organization, political institutions, historical movements, interpersonal relations, collective practices, administrative systems, and culturally specific social roles. These expressions encode how social relations or institutions are understood within the source culture, and therefore may not have direct equivalents in the target language. Examples include institutional titles, policy terms, social-status expressions, work-unit culture, kinship-like social relations, and historically situated political concepts. For example, *Red Tourism* [红色旅游] refers not merely to tourism with a color adjective, but to tourism associated with revolutionary history and political memory. Similarly, *guanxihu* [关系户] denotes a person who benefits from personal connections or nepotistic relations in an institutional setting.

Linguistic Culture. Linguistic Culture includes idioms, proverbs, slang, colloquial expressions, metaphors, literary allusions, and other non-compositional linguistic forms whose meaning cannot be reliably inferred from their literal components. These expressions often encode shared cultural knowledge, historical references, social attitudes, or conventional pragmatic force. Correct translation requires recovering the intended figurative or pragmatic meaning rather than preserving surface form. Examples include chengyu, folk sayings, idiomatic comparisons, and colloquial evaluative phrases. For instance, *Three cobblers* [三个臭皮匠] invokes a proverb about collective wisdom rather than a literal description of three craftsmen. Similarly, expressions such as [吃软不吃硬] require functional or pragmatic translation rather than word-by-word rendering.

Religious Culture. Religious Culture refers to expressions grounded in religious belief systems, ritual practices, moral-philosophical traditions, cosmology, spiritual institutions, and culturally embedded concepts of fate, virtue, worship, or transcendence. These expressions may originate from organized religions, folk beliefs, or philosophical traditions that function as cultural frames for interpreting behavior and events. Examples include references to Confucianism, Buddhist concepts, ritual practices, ancestral worship, temple culture, and religion-derived moral vocabulary. Such expressions require careful translation because their

intended meaning may involve both doctrinal content and broader cultural associations.

Ecological Culture. Ecological Culture covers expressions related to geography, climate, seasonal cycles, agricultural time, local environments, calendrical systems, and culture-specific ways of conceptualizing nature. These expressions are often rooted in the interaction between a community and its natural environment, and may encode culturally specific temporal or spatial knowledge. Examples include solar terms, seasonally marked customs, geography-based expressions, and environmentally grounded metaphors. For instance, *Grain Rain* [谷雨] is not simply a weather description, but one of the traditional Chinese solar terms associated with seasonal change and agricultural timing.

Label Assignment Principle. Each instance receives a single primary cultural label. The label is assigned based on the role of the culture-loaded span in the local sentence context, rather than on the expression in isolation. For example, an expression involving a traditional building may be labeled as Material Culture when the architectural object itself is central, but as Social Culture if the sentence emphasizes an institutional or social practice associated with that building. Similarly, an idiomatic expression containing an ecological image is labeled as Linguistic Culture if the main translation challenge lies in its non-compositional figurative meaning. This single-label design improves diagnostic consistency and enables category-conditioned evaluation in CURE, while acknowledging that some expressions may involve secondary cultural dimensions.

G Error Taxonomy

Omission. Omission occurs when a culture-loaded expression is entirely unrepresented in the target text, leading to the loss of a concept that cannot be recovered through context.

> **Example.** 关系户 (guānxihù) denotes a socially embedded form of nepotism.
> **Source:** 他在单位里是个典型的关系户，很多事情明明不符合规定…
> **System Output:** He has **special connections** at work, and many things are approved…
> **Analysis:** The output replaces a specific cultural referent with a vague descriptor ("special connections"), failing to instantiate the institutionalized nature of the original term. Since the defining concept is absent rather than misinterpreted, it is categorized as Omission.

Literalization. Literalization arises from surface-level lexical correspondence that fails to trigger the idiomatic or conventional meaning recognized in the source culture.

> **Example.** 穿小鞋 (chuān xiǎoxié) metaphorically describes covert retaliation or obstruction.
> **Source:** 自从那次意见不合之后，他就在工作中不断给我穿小鞋…
> **System Output:** …he has been making me **wear small shoes** at work…
> **Analysis:** The word-for-word mapping results in an opaque expression for English readers. The failure to decode the figurative layer while preserving surface lexical components identifies this as Literalization.

Sense Error. Sense Error occurs when a translation selects a contextually incompatible interpretation of a culture-loaded term, introducing a factual or referential inaccuracy.

> **Example.** 老油条 (lǎoyóutiáo) figuratively refers to a sophisticated, street-smart veteran.
> **Source:** 他是个老油条，在这个行业里混了二十多年…
> **System Output:** He is an **old oil dealer** who has worked in this industry…
> **Analysis:** Unlike Literalization, which is often nonsensical, the system here selects a plausible but incorrect literal sense, introducing a false referent. This active selection of the wrong meaning classifies it as a Sense Error.

Neutralization. Neutralization captures the general intent of the source while stripping it of culturally specific connotations or evaluative intensity.

> **Example.** 官僚作风 (guānliáo zuòfēng) encodes a specific cultural critique of self-serving bureaucracy.
> **Source:** …说到底还是官僚作风在作祟，层层推诿…
> **System Output:** …because of **the way organizations operate**, with responsibilities passed around…
> **Analysis:** The translation preserves the situational logic but flattens the cultural salience into a generic description. The loss of cultural specificity without losing the core message defines Neutralization.

Mis-substitution. Mis-substitution involves replacing a source expression with a target-culture analogue that introduces misleading or non-equivalent sociocultural framing.

> **Example.** 关系户 (guānxihù) refers to a person who benefits from personal connections or nepotistic access.
> **Source:** 这是个典型的关系户问题，在一些单位里已经成了公开的秘密…
> **System Output:** This is a typical case of an **old boys' club**…
> **Analysis:** While "old boys' club" is culturally

related, it shifts the reference to a Western social structure that does not fully align with the systemic implications of the Chinese term. This inappropriate analogy constitutes Mis-substitution.

Pragmatic Shift. Pragmatic Shift occurs when the interpersonal stance, tone, or evaluative force of an expression is altered, even if the referential meaning remains intact.

> **Example.** 会来事 (huì láishì) implies social adeptness without inherently carrying moral condemnation.
> **Source:** 他这个人挺会来事的，和领导、同事之间的关系处理得不错。
> **System Output:** He is very good at **pleasing people**...
> **Analysis:** The output injects a negative connotation of insincerity ("pleasing people") not present in the original. The distortion occurs at the level of speaker attitude, identifying it as a Pragmatic Shift.

Over-interpretation. Over-interpretation arises when the translation introduces explicit explanations or value judgments that exceed the semantic encoding of the source.

> **Example.** 走形式 (zǒu xíngshì) criticizes superficial compliance.
> **Source:** 这个检查不过是走形式，真正的问题并没有得到解决...
> **System Output:** This inspection is merely **pretending to comply with bureaucratic requirements**...
> **Analysis:** The system adds explanatory material ("pretending", "bureaucratic requirements") that narrows the interpretive scope and injects extra authorial judgment, exceeding the original's informational content.

Additional Japanese Examples. The following examples illustrate that culture-loaded errors also arise in Chinese-to-Japanese translation, especially when shared scripts or apparent lexical similarity obscure differences in cultural realization.

> **Sense Error.** 地龙 refers here to an underground heating flue rather than an animal.
> **Source:** 这种老房子没有集中供暖，全靠地下的地龙来维持室温。
> **System Output:** この古い家には集中暖房がなく、地下のミミズによって室温を保っている。
> **Analysis:** The output selects the animal-related sense of the term and renders it as "earthworm," producing an incorrect referent. This is categorized as a Sense Error.

> **Neutralization.** 官僚作风 encodes a culturally specific critique of bureaucratic formalism and buck-passing.
> **Source:** 说到底还是官僚作风在作祟，层层推诿。
> **System Output:** 局のところ、のやり方が原因で、任がたらい回しにされている。
> **Analysis:** The translation preserves the general idea of organizational dysfunction but removes the evaluative force of bureaucratic formalism. The culturally salient criticism is flattened into a generic description, making this a case of Neutralization.

> **Mis-substitution.** 关系户 refers to a person who benefits from personal connections or nepotistic access.
> **Source:** 这是个典型的关系户问题，在一些单位里已经成了公开的秘密。
> **System Output:** これは典型的な天下りで、一部のではすでに公然の秘密になっている。
> **Analysis:** Although 天下り is also a Japanese socio-institutional term, it specifically evokes post-retirement placement of bureaucrats and does not match the broader Chinese notion of connection-based privilege. This inappropriate target-culture analogue constitutes Mis-substitution.

> **Pragmatic Shift.** 会来事 describes social tact or interpersonal adeptness without necessarily implying moral blame.
> **Source:** 他这个人挺会来事的，和领导、同事之间的关系处理得不错。
> **System Output:** 彼はかなりごますりがうまく、上司や同僚とのもうまく理している。
> **Analysis:** The Japanese phrase ごますりがうまい introduces a stronger negative implication of flattery or ingratiation. Since the referential meaning is partially preserved but the interpersonal stance changes, this is a Pragmatic Shift.

Metric	Pearson r	Spearman ρ
CURE	**0.78**	**0.81**
COMET	0.46	0.43
BLEU	0.31	0.28

Table 21: Correlation between human cultural-correctness judgments and automatic evaluation metrics.

H CURE: Design, Implementation, and Human Alignment

This appendix provides a consolidated description of the design scope, implementation details, and human alignment analysis of CURE. In contrast to the main paper, which focuses on motivation and empirical results, this appendix clarifies the operational boundaries of the metric, its evaluation pipeline, and the empirical basis for its alignment with human judgments.

All CURE evaluations are conducted with grok-4.1-fast as the reference evaluator. We use deterministic decoding and fixed prompt templates across all experiments to ensure reproducibility. The prompt templates specify the decision criteria for each evaluation component, including the Semantic Validator, Fidelity Critic, and Clarity Critic.

All prompts are held constant across systems, target languages, and evaluation settings. Full prompt templates are provided in Appendix L.

H.1 Evaluation Scope and Boundary Conditions

CURE is designed as a diagnostic evaluation metric targeting the realization of culture-loaded meaning at the span level. Given a source sentence, a system translation, and a human-authored Cultural Explication, CURE evaluates whether the culturally grounded meaning defined by the explication is successfully instantiated in the translation.

CURE does not aim to assess sentence-level fluency, overall translation quality, or adequacy outside the annotated culture-loaded span. It is therefore not intended to replace general-purpose metrics such as BLEU or COMET, but rather to complement them by providing targeted signals specific to cultural translation phenomena. Errors unrelated to cultural meaning, including grammatical infelicities or stylistic differences outside the annotated span, are explicitly out of scope.

H.2 CURE Evaluation Pipeline and Scoring

CURE adopts a deterministic, multi-stage evaluation pipeline implemented via a set of specialized judging agents. Each evaluation instance consists of a tuple (x, y, e), where x denotes the source sentence, y the system translation, and e the corresponding Cultural Explication.

Evaluation Pipeline. The evaluation proceeds in three sequential stages:

1. **Semantic Validation.** A semantic validator determines whether the culturally grounded meaning specified in the explication e is realized in the translation y. This stage produces a binary validity decision $v \in \{0, 1\}$. If $v = 0$, the instance is deemed culturally incorrect and receives an overall CURE score of zero.

2. **Quality Assessment.** For instances with $v = 1$, two independent critics assess complementary aspects of translation quality: (i) a *fidelity critic*, evaluating faithfulness to the source intent, and (ii) a *clarity critic*, evaluating interpretability for target-language readers. Each critic produces a normalized score in $[0, 1]$.

3. **Score Aggregation.** The final CURE score is computed as a weighted combination of the fidelity and clarity scores, conditioned on the cultural category of the instance.

Category-conditioned Scoring. To account for systematic differences in how cultural meaning is expressed across categories, CURE employs two category-conditioned scoring protocols. Let s_f denote the fidelity score and s_c the clarity score. The final quality score S is computed as:

$$S = \alpha_C \cdot s_f + \beta_C \cdot s_c,$$

where α_C and β_C are category-dependent weights satisfying $\alpha_C + \beta_C = 1$.

Specifically, we define the following configurations:

- **Protocol A** (fidelity-oriented): $\alpha_C = 0.7$, $\beta_C = 0.3$
- **Protocol B** (clarity-oriented): $\alpha_C = 0.4$, $\beta_C = 0.6$

Protocol A is applied to cultural categories where preserving source-side meaning and referential precision is critical, while Protocol B is applied to categories where interpretability and pragmatic accessibility for target-language readers play a more prominent role. The assignment of protocols is deterministic and fixed across all experiments.

Determinism and Reproducibility. All judging agents operate under deterministic decoding configurations with fixed prompt templates and no in-context example variation. CURE does not involve any learned parameters or training procedures; all scores are produced at inference time using fixed configurations, ensuring reproducibility across runs.

H.3 Agreement between CURE and Human Judgments

To quantify the alignment between CURE and human evaluation, we conduct a correlation analysis on a manually evaluated subset of the CulT-Eval benchmark. We uniformly sample **1,000 instances** from the full dataset of 7,959 sentence pairs, ensuring coverage across cultural categories and translation systems.

Each sampled instance is independently assessed by expert annotators following the guidelines described in Appendix A. Annotators provide both (i) a binary judgment of cultural correctness and (ii) a graded acceptability score on a five-point Likert

scale, reflecting the degree to which the culturally grounded meaning is realized in the translation.

Table 21 reports Pearson correlation, measuring linear agreement, and Spearman rank correlation, measuring consistency in relative ranking, between human judgments and automatic evaluation metrics. CURE quality scores exhibit strong correlation with human judgments under both measures, substantially outperforming sentence-level metrics such as BLEU and COMET.

The higher Spearman correlation indicates that CURE is particularly effective at preserving the relative ordering of translations with respect to cultural correctness, even when absolute score differences are small. In contrast, sentence-level metrics show weaker alignment with human judgments, suggesting limited sensitivity to culture-specific translation errors.

Overall, these results support the use of CURE as a reliable diagnostic metric for evaluating cultural translation quality, while underscoring its complementary role alongside existing automatic evaluation methods.

I Cultural Explication Quality Control

Cultural Explications serve as the semantic anchor for CURE. To reduce annotation-induced bias, annotators were instructed to write concise, factual, and context-grounded definitions that specify the intended cultural referent without prescribing a single target-language wording. Explications avoid unnecessary stylistic preferences, unsupported background information, and translation-specific phrasing.

During quality control, explications were checked for three criteria: (i) semantic sufficiency, whether the explanation identifies the intended culture-loaded meaning; (ii) non-prescriptiveness, whether it allows legitimate target-language paraphrases or localization strategies; and (iii) contextual fit, whether it matches the source sentence rather than a dictionary-level meaning alone.

I.1 Weight Configuration and Robustness Considerations

The weighting parameters α_C and β_C are introduced to operationalize different evaluation emphases across cultural categories, rather than to optimize metric performance. Specifically, Protocol A adopts a fidelity-oriented configuration ($\alpha_C = 0.7$, $\beta_C = 0.3$), while Protocol B adopts a clarity-oriented configuration ($\alpha_C = 0.4$, $\beta_C = 0.6$). These values are chosen as representative points to reflect contrasting evaluation priorities, and are fixed across all experiments.

Importantly, these parameters are not tuned on evaluation data, nor are they intended to be treated as learnable hyperparameters. Our goal is not to identify an optimal weighting scheme, but to provide a transparent and interpretable mechanism for balancing complementary evaluation dimensions.

We further observe that moderate variations of the weighting parameters within a reasonable range (e.g., ± 0.1) do not alter the qualitative trends reported in the paper, including relative comparisons across systems and categories. This suggests that the reported findings are not sensitive to the specific numerical choices of α_C and β_C, but instead reflect more general properties of cultural translation difficulty.

J Model and Inference Configurations

J.1 Evaluated Systems

All experiments are conducted on translation outputs produced by a set of multiple competitive systems, including both LLM–based translators and strong neural machine translation baselines. These systems are selected to represent a range of commonly used translation paradigms in contemporary multilingual translation settings.

CURE is applied post hoc as an evaluation metric and does not depend on any internal model states, training data, or architectural details of the evaluated systems. All systems are treated as black-box generators, and only their final translation outputs are used for evaluation.

The same set of system outputs is reused across all experimental analyses in the paper, including metric comparison, category-wise evaluation, ablation studies, and the human–metric alignment analysis

J.2 Inference Settings

All translation outputs used in our experiments are generated under deterministic decoding settings, with stochastic sampling disabled. Specifically, decoding is performed with temperature set to 0 and beam size set to 1, ensuring stable and reproducible outputs across runs.

For all systems, the maximum generation length is capped at 256 tokens. No system-specific decoding heuristics, reranking strategies, or post-

processing rules are applied beyond standard translation prompts.

The inference configurations described above are held fixed across all experiments. In particular, the subset of 1,000 instances used for the human–metric alignment analysis is sampled from translations generated under the same deterministic settings.

This unified experimental setup ensures that differences observed across evaluation metrics can be attributed to metric behavior rather than variations in model configuration or inference strategy.

Averaging over Runs. All reported results are computed from a single generation run per system. As all translation outputs are generated under deterministic decoding settings, repeated runs produce identical outputs. Consequently, no averaging across random seeds or multiple runs is required.

Metric scores are first computed at the instance level and then aggregated by taking the arithmetic mean over all evaluated instances. This procedure is applied consistently across all systems and evaluation metrics reported in the paper.

K Difficulty Variations across Types of Culture-loaded Expressions

This appendix analyzes how translation difficulty varies across different types of culture-loaded expressions. Rather than comparing evaluation methods or system performance, the goal of this analysis is to characterize intrinsic differences associated with the nature of cultural expressions themselves, using both human judgments and CURE as complementary difficulty signals.

K.1 Analysis Setup

To control for differences in category size, we conduct a stratified analysis on a balanced subset of the CulT-Eval benchmark. Specifically, we first randomly sample **500 instances** from the full dataset of 7,959 sentence pairs, and then retain **100 instances** for each cultural category defined in Section 3.2. This procedure ensures equal representation across cultural types while avoiding bias toward high-frequency categories.

For each instance, we analyze (i) human judgments of cultural correctness using the five-point acceptability ratings , and (ii) CURE scores produced under the fixed configurations described in Appendix 5. Scores are averaged within each category to reflect relative translation difficulty across cultural expression types.

Cultural Type	#Inst.	Human	CURE
Linguistic idioms	100	3.18	2.84
Social conventions	100	3.39	3.05
Institutional concepts	100	3.53	3.19
Historical / cultural references	100	3.61	3.28
Material culture	100	3.76	3.42
Overall	500	3.49	3.16

Table 22: Average human and CURE scores across different types of culture-loaded expressions on a stratified sample of 500 instances. Lower scores indicate higher translation difficulty.

K.2 Difficulty by Cultural Type

Table 22 reports the number of instances and average scores for each cultural category in the stratified sample. Lower scores indicate higher translation difficulty.

K.3 Discussion

Consistent patterns of difficulty emerge across cultural expression types under both human judgments and CURE scores. Linguistic idioms receive the lowest average scores, reflecting their reliance on non-compositional meaning and context-dependent interpretation. Expressions related to social conventions and institutional concepts occupy an intermediate range, where meanings are culturally grounded but partially recoverable from context.

In contrast, material culture and concrete cultural references receive higher scores, suggesting that expressions with explicit or referential content are generally easier to translate accurately. Importantly, we emphasize that these differences reflect intrinsic properties of cultural expressions rather than deficiencies of any particular translation system or evaluation method.

Overall, the alignment of difficulty patterns across human judgments and CURE suggests that variation in translation difficulty is primarily driven by the nature of the cultural expressions themselves, providing additional context for interpreting evaluation results reported in the main paper.

K.4 Difficulty by Linguistic and Cultural Expression Form

Beyond broad cultural categories, translation difficulty also varies systematically with the linguistic and cultural form of the expression. In this analysis, we further group culture-loaded instances according to four commonly discussed forms: colloquial

Form	Human	CURE
Colloquial sayings	3.52	3.21
Literary expr.	3.34	3.02
Chengyu	3.09	2.76
CSIs	3.61	3.29
Overall	**3.39**	**3.07**

Table 23: Average human and CURE scores across expression forms on a stratified sample of 500 instances, with 125 instances per form. Lower scores indicate higher translation difficulty.

sayings, literary expressions, idiomatic chengyu, and culture-specific items (CSI). These forms differ in terms of compositionality, conventionalization, and reliance on shared cultural knowledge, which in turn affects translation difficulty.

Using the same stratified subset, we analyze human judgments and CURE scores for instances belonging to each form. Scores are averaged within each group to reflect relative translation difficulty, with lower scores indicating higher difficulty.

Table 23 summarizes the results. Idiomatic chengyu receive the lowest average scores under both human judgment and CURE, reflecting their highly non-compositional semantics and strong dependence on cultural convention. Literary expressions also exhibit relatively high difficulty, as their meanings are often conveyed indirectly through stylistic or rhetorical devices. In contrast, colloquial sayings and culture-specific items tend to receive higher scores, suggesting that expressions with clearer pragmatic functions or referential anchors are generally more accessible in translation.

As in previous analyses, we emphasize that these differences reflect intrinsic properties of expression forms rather than shortcomings of any particular translation system.

L Prompt

Prompt Template: Culture-Mining Agent (Data Extraction)

System Instruction: You are a bilingual data mining expert. Your task is to extract "Culture-Loaded Sentence Pairs" from the provided raw bilingual text.
IMPORTANT CONSTRAINTS:

- You must identify sentences containing Idioms, Slang, or Culture-Specific Items (e.g., history, food, traditions).
- You must extract both the Chinese source and the English translation.
- Ignore common sentences that lack specific cultural depth.
- Output must be a strictly valid JSON list.

User Input:
```
[User] Extract cultural pairs from the following text chunk:
Raw Text: {RAW_TEXT_CHUNK}
Return format: [ { "src": "Chinese sentence...", "tgt": "English sentence...", "focus_term": "The specific cultural word" } ]
```

Figure 6: The prompt template for the Culture-Mining.

Prompt Template: Fine-grained Taxonomy Classifier

System Instruction: You are a Cultural Linguist. Your task is to classify a specific Chinese term into one of the **Five Cultural Categories**.
Taxonomy Definitions:

1. **Ecological Culture:** Terms related to specific animals, plants, geography, climate, or natural phenomena unique to the region (e.g., 梅雨, 熊猫, 黄河).
2. **Material Culture:** Terms related to food, clothing, architecture, artifacts, or daily necessities (e.g., 旗袍, 饺子, 炕, 四合院).
3. **Social Culture:** Terms related to institutions, titles, festivals, customs, history, or social hierarchy (e.g., 高考, 春节, 尚书, 关系户).
4. **Religious Culture:** Terms related to beliefs, mythology, philosophy (Confucianism, Taoism, Buddhism), or taboos (e.g., 阴阳, 菩萨, 玉帝).
5. **Linguistic Culture:** Terms involving idioms (Chengyu), metaphors, slang, proverbs, or witticisms (e.g., 吃软不吃硬, 摸鱼, 破防).

User Input:
```
[User] Context Sentence: {SOURCE_SENTENCE} Target Term: {FOCUS_TERM}
Task: 1. Analyze the meaning of the term in context. 2. Assign strictly ONE category from the list above. 3. Provide a short explanation.
Return Format JSON: { "term": "{FOCUS_TERM}", "category": "Social Culture", "reason": "It refers to a specific historical government position." }
```

Figure 7: Prompt template for fine-grained cultural classification.

Prompt Template: Core Dispatcher (Protocol Routing)

System Instruction: You are a taxonomy expert in cultural linguistics. Your task is to classify the "Cultural Category" of a specific term within a source sentence.
IMPORTANT DEFINITIONS:

- **Protocol A (Fact-Centric):**
  - Specific Concepts: Concrete institutions, historical artifacts, unique objects, technical items.
  - Proper Nouns: Names of unique places, people, organizations, festivals.
  - **Core Logic:** Referential Precision. The translation must point to the EXACT same entity.
  - **Tie-Breaker:** If a term has a specific history/material existence, choose A (e.g., "Forbidden City").
- **Protocol B (Style-Centric):**
  - Figurative Language: Idioms (Chengyu), metaphors, allegories.
  - Slang/Pop Culture: Buzzwords, internet memes, dialect words used for effect.
  - **Core Logic:** Pragmatic Equivalence. Imagery and tone are more important than literal words.
  - **Tie-Breaker:** If the term describes a situation/feeling rather than an object, choose B (e.g., "eating vinegar").

User Input:
`[User] Analyze the cultural term inside the brackets "[]" in the source sentence.`
`[Input] Source: {SOURCE_WITH_BRACKETS}`
`[Task] Classify the term based on its primary function in this specific context. Return ONLY one label: "Protocol A" or "Protocol B".`

Figure 8: The Core Dispatcher prompt. It routes instances to the correct evaluation protocol (Fact-Centric or Style-Centric) based on linguistic features and functional context.

Prompt Template: Stage I - Semantic Validator (The Gate)

System Instruction: You are a strict Semantic Validator. Your ONLY goal is to detect "Hallucinations" (Error Type B3) or "Severe Mis-substitutions" (Error Type B5).
CRITICAL GROUND TRUTH RULE:

- You must rely EXCLUSIVELY on the provided [Cultural Explication].
- If the Explication says X, and the Model translates it as Y (where Y!=X), it is INVALID.
- Ignore your own internal knowledge if it conflicts with the Explication.

NEGATIVE CONSTRAINTS (What NOT to check):

- DO NOT check for fluency, grammar, or style.
- DO NOT check for "neutralization" (loss of flavor). A boring but factually correct translation is VALID.
- DO NOT penalize literal translations here, provided they refer to the correct concepts.

User Input:
`[User] [Data] Protocol: {PROTOCOL_LABEL} Source: {SOURCE} Cultural Term: {TERM} Cultural Explication (Ground Truth): {EXPLICATION} Model Hypothesis: {HYPOTHESIS}`
`[Task] 1. Ignore whether the translation is elegant. 2. Check ONLY if the semantic meaning matches the Explication. 3. If the hypothesis invents a new entity (Hallucination) or refers to a completely wrong concept, mark INVALID.`
`[Output format] Reasoning: [Brief analysis] Decision: VALID or INVALID`

Figure 9: The Semantic Validator prompt. It acts as a strict "Validity Gate" ($\mathbb{I}_{valid}$), filtering out hallucinations by anchoring evaluation to the Cultural Explication.

Prompt Template: Stage II-A - Fidelity Critic (Nuance Analysis)

System Instruction: You are a Translation Critic focusing on "Fidelity" and "Pragmatic Equivalence". Your job is to score how well the Model Hypothesis captures the intended nuance and force, strictly according to the active Protocol.
REFERENCE USAGE RULE:

- The [Reference Anchor] is just ONE possible translation.
- **DO NOT** penalize the hypothesis for using different words than the reference.
- **DO** judge based on whether the *meaning* and *effect* are equivalent.

DYNAMIC GUIDELINE (Protocol-Dependent): {DYNAMIC_INSTRUCTION}
SCORING SCALE:

- **5 (Perfect):** Captures full nuance, tone, and imagery.
- **3 (Acceptable):** Core meaning present, but significant nuance lost (e.g., too generic).
- **1 (Failure):** Severe mistranslation or complete loss of meaning.

User Input:
`[User] [Data] Protocol: {PROTOCOL_LABEL} Source: {SOURCE} Reference Anchor: {REFERENCE} Model Hypothesis: {HYPOTHESIS}`
`[Task] Evaluate Fidelity (1-5). Focus: Does the hypothesis capture the {PROTOCOL_LABEL} constraints?`
`[Output format] Reasoning: [Critique] Score: [1-5]`

Figure 10: The Fidelity Critic prompt. It dynamically adjusts criteria (e.g., penalizing literalism in Protocol B vs. generalization in Protocol A) to assess translation nuance.

Prompt Template: Stage II-B - Clarity Critic (Communicative Intelligibility)

System Instruction: You are a Target Audience Evaluator simulating an English native reader who has NO prior knowledge of Chinese culture. Your goal is to assess "Communicative Intelligibility".
SCORING PHILOSOPHY (Thick Translation):

- **Bonus (+):** Reward **Explicitation** (e.g., in-text glosses, brief explanations, appositives like "Kang, a heated brick bed"). Reward **Transparency** (rephrasing for clarity).
- **Penalty (-):** Penalize **Opaque Terms** (Pinyin without context). Penalize **Confusion** (if the reader would ask "What does that mean?").

SCORING SCALE:

- **5 (Crystal Clear):** Seamlessly bridged. The naive reader fully understands.
- **3 (Gist Only):** Reader gets the general idea but misses the cultural specific.
- **1 (Incomprehensible):** Complete communication breakdown.

User Input:
`[User] [Data] Source: {SOURCE} Model Hypothesis: {HYPOTHESIS}`
`[Task] Evaluate Clarity (1-5) for a non-Chinese reader. Did the translator build a bridge for the reader, or leave them confused?`
`[Output format] Reasoning: [Analysis] Score: [1-5]`

Figure 11: The Communicator prompt. It rewards explicitation strategies consistent with Appiah's "Thick Translation" theory.