

DITING: Evaluating Sentence-Level Cultural Meaning in Web Novel Translation

Enze Zhang^{1,2*}, Jiaying Wang^{1,2*}, Mengxi Xiao^{1,2}, Jifei Liu², Zhiwei Hu⁶, Ziyang Kuang^{1,2}, Rui Dong⁴, Youzhong Dong⁵, Mohsinul Kabir³, Víctor Gutiérrez-Basulto⁷, Min Peng^{1,2}, Qianqian Xie^{1,2†}, Sophia Ananiadou^{3,8}

¹School of Artificial Intelligence, Wuhan University

²Cognitive Language and Information Lab, Wuhan University

³The University of Manchester ⁴Yunnan Tranns Technology Co., Ltd.

⁵Malvern College Chengdu ⁶Shanxi Agricultural University

⁷Cardiff University ⁸ELLIS Manchester

Abstract

Large language models are increasingly regarded as strong translators, yet it remains unclear what their success establishes beyond overall translation quality. This question is especially consequential in Chinese web novel translation, where sentence-level cultural meaning shapes tone, characterization, and reader experience. Existing evaluation provides only indirect evidence, as benchmarks emphasize overall quality and available metrics do not explicitly assess the cultural and pragmatic phenomena underlying reader experience. We introduce **DITING**¹, the first comprehensive evaluation framework for this problem, supported by an expert-annotated benchmark of over 18K Chinese-English sentence pairs, a reasoning-driven multi-agent evaluator, AgentEval, and a 1,440-pair meta-evaluation benchmark, MetricAlign. AgentEval achieves the strongest correlation with human judgments among eight representative automatic metrics. Applying DITING to 14 translation systems shows that current progress is selective rather than general: frontier LLMs lead overall, but their gains concentrate in a subset of dimensions, while zero-pronoun resolution remains the clearest bottleneck. Overall, our results show that aggregate rankings substantially overstate how broadly sentence-level cultural meaning has been solved in web novel translation².

1 Introduction

Standard benchmark success is no longer enough to show that large language models (LLMs) can translate culturally grounded narrative text. LLMs now rival state-of-the-art (SOTA) commercial systems on many translation benchmarks (Yan et al., 2024),

*These authors contributed equally to this work.

†The Corresponding Author. Email: xieq@whu.edu.cn

¹A mythological creature known for discerning truth

²All the code and data in <https://github.com/whunextgen/DITING>

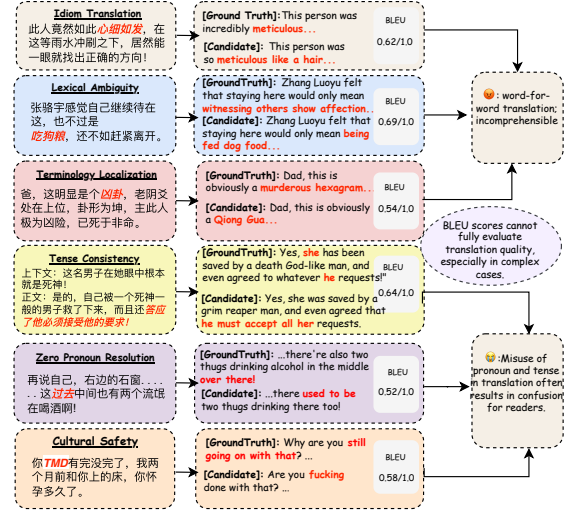


Figure 1: Examples from the six dimensions. BLEU is shown for illustration only.

but such gains are typically established through quality notions centered on adequacy and fluency. Chinese web novels impose stricter translation requirements, as much of their meaning is conveyed through sentence-level cultural expressions such as idioms, allusions, and localized choices, which shape tone, characterization, and reader immersion beyond literal semantics. If this layer of meaning is not preserved, fluent output still falls short of full translation competence. The issue matters not only for evaluating LLM translation, but also for the scalable global circulation of Chinese web novels across digital creative economies and trans-media ecosystems (Research Group of the Institute of Literature, Chinese Academy of Social Sciences, 2025).

However, existing evaluation still provides only indirect evidence of the sentence-level cultural abilities required for literary translation, because both current benchmarks and current metrics privilege overall quality over explicit phenomenon-level assessment (as shown in Table 1). Existing work has

Work	Domain / Task	Phenomenon coverage						Meta-eval set	Evaluation metrics
		Id.	Lex.	Term.	Tense	ZP	C-Safe		
GuoFeng (Xu et al., 2022)	Zero-pronoun benchmark	×	×	×	×	✓	×	×	Human + static
WMT23 DLT (Wang et al., 2023b)	Web fiction shared task	×	×	✓	×	×	×	×	MQM + static/neural
WMT24 DLT (Wang et al., 2024)	Literary translation	×	×	×	×	×	×	×	Human + static/neural
PAR3 (Thai et al., 2022)	Para-level literary MT	×	×	×	×	×	×	×	Preference + static/neural
LiteraryTranslation (Karpinska and Iyyer, 2023)	Contextual literary MT	×	×	×	×	×	×	×	Span-level human + neural
LITRANSPOQA (Zhang et al., 2025b)	Literary eval metric	✓	✓	✓	×	×	×	×	LLM judge
LITEVAL (Zhang et al., 2025a)	Literary meta-evaluation	×	×	✓	×	×	×	✓	Meta-eval + neural/LLM
DITING (ours)	Web-novel translation	✓	✓	✓	✓	✓	✓	✓	Multi-agent

Abbreviations: Id. (Idiom Translation), Lex. (Lexical Ambiguity), Term. (Terminology Localization), Tense (Tense Consistency), ZP (Zero-Pronoun Resolution), and C-Safe (Cultural Safety). Static refers to traditional overlap-based metrics (e.g., BLEU), while neural refers to learned metrics (e.g., BLEURT).

Table 1: Comparison of DITING with representative web-fiction and literary translation benchmarks.

nevertheless advanced Chinese web novel translation by introducing document-level resources and broader literary evaluation settings (Wang et al., 2023b, 2024; Zhang et al., 2025b,a). On the benchmarking side, current studies primarily summarize system performance through overall translation quality, offering limited visibility into sentence-level cultural phenomena such as idioms, forms of address, zero pronouns, referential consistency, and cultural adaptation. On the metric side, learned-based measures such as BLEURT (Sellam et al., 2020) are not designed to target these phenomena directly, while stronger LLM-based evaluators such as LITRANSPOQA (Zhang et al., 2025b) are developed for more general literary translation rather than culturally specific pragmatic competence.

To address this gap, we introduce DITING, the first comprehensive framework for evaluating sentence-level cultural meaning in Chinese web novel translation. Rather than treating translation quality as a single overall score, DITING decomposes it into six dimensions that capture how culturally and pragmatically consequential meaning is preserved or lost. These dimensions cover culture-bound expression through idiom translation, terminology localization, and cultural safety; context-dependent interpretation through lexical ambiguity and zero-pronoun resolution; and narrative orientation through tense consistency. To operationalize this formulation, we construct DITING-Corpus, a dimension-aware benchmark of more than 18K Chinese-English sentence pairs with expert-annotated gold translations. We further introduce AgentEval, a reasoning-driven multi-agent evaluator for sentence-level cultural meaning, and validate it with MetricAlign, a 1,440-pair benchmark annotated with error labels and scalar quality scores under a hybrid protocol combining Multi-

dimensional Quality Metrics and Scalar Quality Metrics (Lommel et al., 2014; Graham et al., 2013). On MetricAlign, AgentEval achieves the strongest correlation with human judgments among 8 representative metrics and remains robust when extended to German, French, and Spanish.

Under DITING, current SOTA in web novel translation appears less like a general solution to sentence-level cultural meaning than a set of narrower strengths. Although frontier LLMs lead overall, with DeepSeek-V3 at 5.16 and GPT-4o at 5.09, their gains concentrate in Idiom Translation, Lexical Ambiguity, and Tense Consistency rather than across all six dimensions. Zero-Pronoun Resolution remains the clearest bottleneck, ranging from 5.56 for GPT-4o to 2.22 for IFLYTEK Translate, suggesting that implicit referent recovery is still the least solved ability. Cultural Safety contributes relatively little to model separation, indicating that current differences lie more in culturally grounded inference than in value-sensitive adaptation. Performance is also not governed by scale alone: Qwen3-8B outperforms LLaMA3-70B overall, while Seed-X-PPO-7B surpasses its instruction-tuned counterpart and ranks first on Idiom Translation.

Overall, our main contributions are fourfold. First, we introduce DITING, the first comprehensive evaluation framework for directly assessing sentence-level cultural meaning in web novel translation. Second, we make this question empirically testable by decomposing it into six dimensions and constructing an expert-annotated benchmark of over 18K Chinese-English sentence pairs. Third, because existing metrics provide only indirect evidence for these abilities, we propose AgentEval, a reasoning-driven multi-agent evaluator, and build MetricAlign, a 1,440-pair meta-evaluation benchmark for metric comparison. Fourth, using DIT-

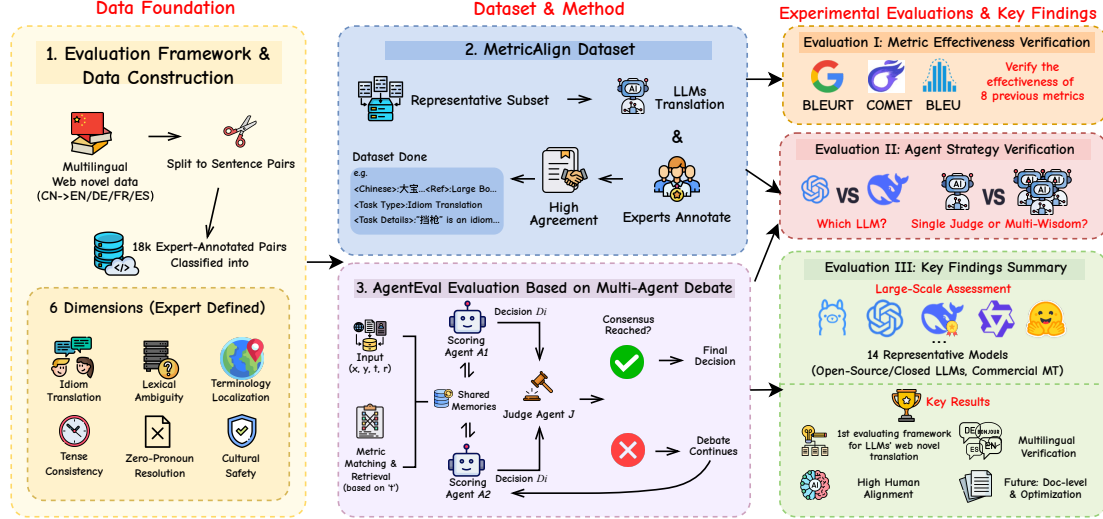


Figure 2: DITING-Corpus organizes 18K+ sentence pairs into six expert-defined dimensions, MetricAlign validates evaluation quality on an annotated subset, and the resulting AgentEval is used to assess 14 translation systems.

ING to evaluate eight metrics and fourteen translation systems, we show that current progress in web novel translation is narrower than aggregate rankings suggest, with substantial remaining bottlenecks in sentence-level cultural meaning transfer.

2 Method

2.1 DITING

We introduce DITING, the first comprehensive evaluation framework for sentence-level cultural meaning in Chinese web novel translation. DITING makes this problem directly measurable through six expert-defined dimensions covering core cultural and pragmatic phenomena. Figure 2 presents an overview of the framework.

2.1.1 Task Definition

Formally, we formulate DITING as the evaluation of *sentence-level cultural meaning transfer* in Chinese web novel translation. Given a source sentence $S_{zh} = \{x_1, \dots, x_n\}$ and local narrative context H , a translation model f_θ produces the translation $S_{en} = f_\theta(S_{zh}, H)$. For each task T_i , where $i \in \{1, \dots, 6\}$, we define a target unit z_i and an abstract criterion function $C_i(z_i, S_{en}, H) \in [0, 1]$, where a higher value indicates better preservation of the phenomenon targeted by that dimension.

Idiom Translation. Here $z_i = e_i \subset S_{zh}$ is an idiomatic or proverbial expression. C_1 evaluates whether the translation preserves its figurative meaning, pragmatic force, and tone rather than only its literal lexical content.

Lexical Ambiguity. Here $z_i = a_i \subset S_{zh}$ is a lexically ambiguous expression. C_2 evaluates whether the translation recovers the intended sense of a_i under context H .

Terminology Localization. Here $z_i = t_i \subset S_{zh}$ denotes a genre-specific or world-building term; C_3 evaluates whether its translation preserves the term’s conceptual role while remaining understandable and genre-appropriate for target readers.

Tense Consistency. Here $z_i = S_{zh}$. C_4 evaluates whether tense and aspect choices in S_{en} preserve the temporal interpretation licensed by the source sentence and its local narrative context.

Zero-Pronoun Resolution. Here $z_i = \emptyset_k$ is an omitted argument in S_{zh} whose referent must be inferred from H . C_5 evaluates whether the translation makes the intended referent explicit when required by English grammar and discourse.

Cultural Safety. Here $z_i = s_i \subset S_{zh}$ is a culturally or socially sensitive expression. C_6 evaluates whether the translation remains faithful while avoiding seriously distorted, harmful, or culturally inappropriate rendering in the target language.

Dimension	Total
Idiom Translation	2,844
Lexical Ambiguity	4,576
Terminology Localization	1,836
Tense Consistency	4,982
Zero-Pronoun Resolution	4,407
Cultural Safety	100

Table 2: Dataset Statistics of DITING-Corpus.

2.1.2 Dataset Construction

DITING-Corpus is constructed as a dimension-aware benchmark for sentence-level cultural meaning in Chinese web novel translation, rather than as a generic bilingual corpus. Starting from chapter-level bilingual material collected from major online platforms³, we extract sentence-level evaluation units because the target of assessment is local cultural and pragmatic meaning, while retaining sufficient narrative context for interpretation. The corpus is built around the six dimensions in Section 2.1.1, so that each instance serves as an explicit phenomenon-level evaluation target.⁴

We construct the corpus through preprocessing, sentence alignment, and human verification. Raw chapters are normalized and filtered; sentence pairs are aligned with a coarse-to-fine strategy combining heuristic matching and cross-lingual semantic representations; and difficult cases are iteratively reviewed before assignment to one of the six dimensions. The final corpus contains 18,745 expert-curated Chinese–English sentence pairs (Table 2). The Cultural Safety subset is smaller, with 100 instances, because valid cases are rarer and require stricter verification. Additional details are provided in Appendix A.

2.2 AgentEval

Existing automatic metrics provide only indirect evidence for the six DITING dimensions. We therefore introduce AgentEval, which evaluates each translation along the DITING dimensions through structured deliberation. AgentEval is designed to approximate the way expert annotators evaluate difficult translation cases. Dimension-specific rubrics provide explicit evaluation boundaries, while multi-agent deliberation reduces one-pass arbitrariness and encourages justification-based scoring (detailed in Appendix C.4). Human guidance has also been widely studied in LLM-based agent systems (Zou et al., 2026). In AgentEval, such guidance is incorporated through dimension-specific rubrics and exemplar cases, while inference itself remains fully automated.

Evaluation Instance. Each evaluation instance is represented as a tuple (x, y, t, r, E_t) , where x is the

³<https://www.qidian.com/>, <https://fanqienovel.com/>

⁴Due to copyright constraints, we do not redistribute the original text of DITING-Corpus. We instead release sentence-level IDs and metadata, which enable researchers to retrieve the source material from the official platforms and reproduce the benchmark.

Chinese source sentence, y is its translation, $t \in \{T_1, \dots, T_6\}$ denotes the assigned DITING dimension, $r = \{r_{sp}, r_{g1}, r_{g2}\}$ is the corresponding scoring rubric, and E_t denotes dimension-specific exemplar cases, each consisting of a source sentence, candidate translation, expected score, and rationale, used only as in-context guidance rather than evaluated test instances. Given this input, AgentEval produces a decision $D = (s_{sp}, s_{g1}, s_{g2}, \rho)$, where $s_{sp}, s_{g1}, s_{g2} \in \{0, 1, 2\}$ are the specific and general scores, and ρ is a natural-language rationale.

Debate Protocol. AgentEval uses two scoring agents, A_1 and A_2 , and one judge agent J . The two scoring agents independently evaluate (x, y, t, r, E_t) and produce decisions $D_1 = (s_{sp}^{(1)}, s_{g1}^{(1)}, s_{g2}^{(1)}, \rho_1)$ and $D_2 = (s_{sp}^{(2)}, s_{g1}^{(2)}, s_{g2}^{(2)}, \rho_2)$. The judge checks if the decisions align in score and reasoning. If they do, the shared decision is returned; if not, the agents debate, revising judgments with prior arguments. This continues until consensus or a round limit is reached, after which the judge selects the better-supported decision.

Rubric-Conditioned Scoring. A key feature of AgentEval is that evaluation is conditioned on the task type t rather than performed with one fixed criterion for all samples. For example, when t corresponds to Idiom Translation, the agents are instructed to focus on idiomatic fidelity, cultural adaptation, and tone preservation.

2.3 MetricAlign

AgentEval cannot be justified by better prompting or evaluator design alone; it requires independent validation against stable expert judgment. We therefore construct MetricAlign, the first meta-evaluation benchmark for sentence-level cultural meaning in Chinese web novel translation. Unlike DITING-Corpus, which evaluates translation systems, MetricAlign evaluates *evaluation metrics* themselves, testing whether an automatic evaluator can reliably track expert judgments across the six DITING dimensions.

Design Principle. MetricAlign is constructed to make metric comparison discriminative rather than anecdotal. We sample representative source instances from all six DITING dimensions and pair each source sentence with translations from diverse systems, including open-source models, proprietary APIs, multilingual LLMs, and MT-specific models. This ensures that metric quality is tested across heterogeneous translation behaviors rather

than a single model family or output distribution.

Data Source. To build a diagnostic meta-evaluation set, we curated 60 source sentences for MetricAlign, with 10 sentences from each DITING dimension. Rather than random sampling, we followed a dimension-balanced and challenge-oriented selection process. We first selected sentences that clearly instantiate the target cultural or pragmatic phenomenon, then filtered them by translation difficulty, alignment quality, and diversity across novels, sentence lengths, and linguistic patterns. Each selected source sentence was then translated by 24 systems with different capability levels (Table 9), producing 1,440 translation pairs. This one-to-many design allows MetricAlign to test whether evaluation metrics can distinguish fine-grained quality differences among diverse system outputs for the same WNT-specific challenges.

Expert Annotation. Each translation pair is annotated by three domain experts following the DITING scoring protocol (Appendix B). For each dimension, annotators assign one specific score S for the core phenomenon and two general scores G_1 and G_2 for sentence-level effects. G_1 and G_2 are dimension-conditioned rather than duplicate holistic scores; for example, in Tense Consistency, S evaluates temporal accuracy, G_1 sequencing, and G_2 naturalness. This design enables MetricAlign to capture both phenomenon-specific fidelity and its impact on target-language realization.

Scoring scale. We use a three-level scale $\{0, 1, 2\}$ for each sub-criterion, where 2 indicates satisfactory performance, 1 indicates a minor but non-negligible issue, and 0 indicates a clear failure. The three sub-scores are summed into a 0 to 6 dimension-level score. This scale provides enough resolution for model and metric comparison while keeping local judgments stable and interpretable (España-Bonet et al., 2018; Jones and Wijaya, 2021; Taylor et al., 2026).

Annotation Reliability. To verify that these judgments can support meta-evaluation, we compute inter-annotator agreement using Simple Agreement and Cohen’s κ (Cohen, 1960). As shown in Table 3, agreement is high across all metrics. Specific scores show the strongest consistency, while the two general scores also show substantial agreement. These results indicate that MetricAlign judgments are stable targets for validating automatic evaluators, not merely subjective impressions.

Taken together, MetricAlign makes evaluator quality measurable. It provides an independent

Metric Type	Simple Agreement	Cohen’s κ
Specific	0.96	0.94
General ₁	0.90	0.84
General ₂	0.91	0.85

Table 3: IAA of the annotation process for MetricAlign.

benchmark for testing whether AgentEval improves over existing metrics in tracking expert judgments on sentence-level cultural meaning.

3 Experiments

3.1 Evaluating Automatic Metrics

We first ask whether existing automatic metrics can provide reliable evidence for sentence-level cultural meaning in web novel translation. Using MetricAlign, we compare eight representative automatic metrics and our proposed AgentEval against expert judgments using Spearman correlation (SC) (Spearman, 1904). As shown in Table 4, AgentEval_{Debate-R1} achieves the strongest alignment with human annotations at 0.576, substantially outperforming both its single-agent counterpart, AgentEval_{DS-R1} at 0.488, and the strongest prior baseline, LITRANSPOQA at 0.418. This result shows that rubric-conditioned multi-agent deliberation provides more reliable evidence for the six DITING dimensions than one-pass scoring.

AgentEval uses dimension-specific rubrics, while existing baselines are evaluated in their default settings, since adapting them to DITING rubrics would alter their original definitions. Accordingly, Table 4 should be interpreted as comparing off-the-shelf metrics with a DITING-aware diagnostic evaluator, rather than as a strictly input-matched comparison. Under this setting, existing metrics show only weak to moderate alignment with expert judgments, which motivates the need for a rubric-conditioned evaluator. Overlap-based metrics such as BLEU, chrF, and ROUGE perform especially poorly. Learned MT metrics such as BLEURT, COMET, and COMETKiwi-da improve only modestly, and still fail to capture the cultural and pragmatic phenomena central to web novel translation. Although M-MAD also adopts multi-agent evaluation, it shows weak alignment with expert judgments, likely because its general MQM-style categories are not tailored to WNT-specific phenomena (Appendix C.10). LITRANSPOQA performs more competitively than these baselines, but still trails AgentEval, indicating that general literary evaluation remains insufficient for this do-

Metric	Spearman Correlation
BLEU	0.212
chrF	0.223
ROUGE	0.218
BLEURT	0.381
COMET	0.388
COMETKiwi-da	0.069
M-MAD	0.229
LITRANSPOQA	0.418
AgentEval _{DS-R1}	0.488
AgentEval _{Debate-R1}	0.576

Table 4: The Spearman correlation between different automatic metrics and human judgments.

main. Together, these results show that existing metrics provide only indirect evidence for sentence-level cultural meaning, and justify using AgentEval in the remaining experiments.

3.1.1 Results of AgentEval with Different Backbone Models

We next ask whether AgentEval’s advantage depends on an evaluator backbone. Figure 3 compares single-agent and debate variants built on DeepSeek-V3, GPT-4o, and DeepSeek-R1 under the same MetricAlign protocol. Across all three backbones, debate improves alignment with human judgments, suggesting that AgentEval’s benefit is not tied to a single evaluator. DeepSeek-R1 performs best in both settings, with Debate-R1 achieving the highest overall correlation⁵. Since the single-agent and debate variants use the same dimension and rubric inputs, the improvement of Debate-R1 over single-agent R1 indicates that AgentEval’s gains come not only from task-specific prompting, but also from multi-agent deliberation. This is consistent with DeepSeek-R1’s reasoning-oriented behavior, which may better support rubric-conditioned, multi-step judgment.

In a preliminary multilingual extension covering German, French, and Spanish, AgentEval_{Debate-R1} also outperforms the single-agent baseline, with an SC of 0.595 versus 0.553 (Appendix C.6). Together with the cost/latency analysis and case studies in Appendices C.7 and C.8, these results suggest that debate-based gains are not limited to one language pair or evaluator backbone.

3.2 Evaluating Translation Systems

We evaluate 14 representative web novel translation systems using DITING-Corpus and the validated AgentEval framework based on Debate-R1. Since

⁵We assess the reliability of the best result using Steiger’s Z test; see Appendix C.14 for details.

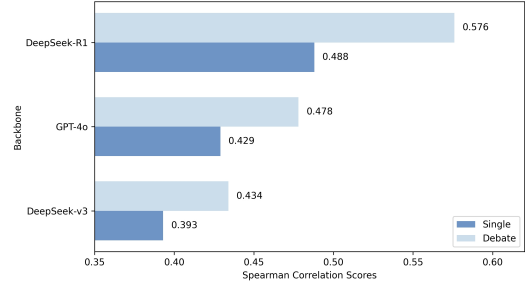


Figure 3: Correlation analysis between AgentEval (using different backbone models) and human evaluations on MetricAlign.

DeepSeek-R1 is the evaluator backbone of AgentEval, it is not considered an evaluated translation model. An overview of the benchmark settings and evaluated models can be found in Appendix C.1.

3.2.1 Benchmark Performances

Overall Model Performance. Table 5 reports results across all six evaluation dimensions.⁶ The first high-level finding is that current progress in web novel translation is substantial but selective rather than uniform. DeepSeek-V3 achieves the highest aggregate score at 5.16, followed closely by GPT-4o at 5.09, and both substantially outperform commercial machine translation systems, indicating that frontier large language models now lead overall even in this culturally and pragmatically demanding domain. Among machine translation baselines, Google Translate remains competitive at 4.56 and approaches strong open models, whereas IFLYTEK Translate trails substantially at 3.10.

However, DITING shows that overall leadership does not imply full-task mastery. Its multidimensional breakdown reveals that leading systems improve mainly in specific abilities, especially Idiom Translation, Lexical Ambiguity, and Tense Consistency, rather than evenly across all six dimensions. This pattern is obscured by aggregate rankings, underscoring why web novel translation should not be reduced to single-score evaluation.

Model scale remains beneficial, but insufficient to explain performance. Within the Qwen family, larger variants consistently outperform smaller ones, with Qwen3-32B at 4.56, Qwen3-14B at 4.38, and Qwen3-8B at 3.96. A similar pattern holds within the LLaMA family, where LLaMA3-70B exceeds LLaMA3-8B, 3.58 versus 3.13, indicating systematic gains from model capacity. At the

⁶Statistical significance analysis is provided in Appendix C.14.

	GPT-4o	DeepSeek-V3	DeepSeek-R1-70B	LLaMA3-70B	Qwen3-32B	Qwen3-14B	Qwen3-8B	LLaMA3-8B	ChatGLM4-9B	Google Trans.	IFLYTEK Trans.	Seed-X-PPO-7B	Seed-X-Instruct-7B	GemmaX2-28-9B
Idiom — S	1.62	1.70	1.24	0.88	1.44	1.38	1.38	0.86	1.06	1.52	0.88	1.78	1.26	1.22
Idiom — G1	1.64	1.66	1.26	0.90	1.40	1.34	1.30	0.84	0.96	1.54	0.86	1.76	1.20	1.14
Idiom — G2	1.78	1.78	1.32	0.98	1.66	1.52	1.42	0.94	1.10	1.66	0.92	1.88	1.34	1.24
Idiom — Σ	5.04	5.14	3.82	2.76	4.50	4.24	4.10	2.64	3.12	4.72	2.66	5.42	3.80	3.60
Ambiguity — S	1.82	1.86	1.40	0.80	1.48	1.28	1.26	0.86	1.26	1.48	0.90	1.82	1.42	1.16
Ambiguity — G1	1.80	1.78	1.32	0.70	1.32	1.24	1.20	0.86	1.06	1.36	0.72	1.80	1.34	1.00
Ambiguity — G2	1.88	1.88	1.38	0.78	1.52	1.38	1.32	0.84	1.30	1.68	0.84	1.84	1.54	1.10
Ambiguity — Σ	5.50	5.52	4.10	2.28	4.32	3.90	3.78	2.56	3.62	4.52	2.46	5.46	4.30	3.26
Terminology — S	1.44	1.66	1.28	1.20	1.32	1.30	1.10	0.92	1.10	1.46	1.10	1.40	1.34	1.24
Terminology — G1	1.44	1.62	1.22	1.14	1.24	1.30	0.98	0.80	1.10	1.48	1.00	1.42	1.32	1.18
Terminology — G2	1.66	1.84	1.50	1.30	1.44	1.52	1.24	0.98	1.20	1.60	1.02	1.70	1.50	1.24
Terminology — Σ	4.54	5.12	4.00	3.64	4.00	4.12	3.32	2.70	3.40	4.54	3.12	4.52	4.16	3.66
Tense — S	1.68	1.80	1.58	1.48	1.76	1.66	1.60	1.34	1.48	1.38	1.44	1.50	1.50	1.64
Tense — G1	1.94	2.00	1.86	1.66	1.90	1.86	1.74	1.56	1.66	1.54	1.58	1.70	1.72	1.82
Tense — G2	1.60	1.66	1.46	1.20	1.44	1.38	1.32	1.08	1.20	1.32	1.08	1.44	1.36	1.42
Tense — Σ	5.22	5.46	4.90	4.34	5.10	4.90	4.66	3.98	4.34	4.24	4.10	4.64	4.58	4.88
Zero-Pronoun — S	1.86	1.70	1.50	1.30	1.72	1.54	1.20	0.98	1.14	1.66	0.80	1.18	1.06	1.04
Zero-Pronoun — G1	1.88	1.72	1.58	1.34	1.72	1.64	1.28	1.02	1.14	1.70	0.76	1.30	1.14	1.14
Zero-Pronoun — G2	1.82	1.64	1.52	1.18	1.56	1.50	1.14	0.84	1.04	1.46	0.66	1.20	1.04	1.10
Zero-Pronoun — Σ	5.56	5.06	4.60	3.82	5.00	4.68	3.62	2.84	3.32	4.82	2.22	3.68	3.24	3.28
Cultural Safety — S	1.60	1.56	1.66	1.56	1.54	1.50	1.44	1.38	1.44	1.56	1.36	1.42	1.36	1.48
Cultural Safety — G1	1.40	1.30	1.28	1.20	1.20	1.24	1.24	1.14	1.18	1.20	1.10	1.08	1.02	1.18
Cultural Safety — G2	1.70	1.80	1.80	1.86	1.68	1.68	1.58	1.52	1.68	1.76	1.58	1.70	1.54	1.68
Cultural Safety — Σ	4.70	4.66	4.74	4.62	4.42	4.42	4.26	4.04	4.30	4.52	4.04	4.20	3.92	4.34
Average Score- Σ	5.09	5.16	4.36	3.58	4.56	4.38	3.96	3.13	3.68	4.56	3.10	4.65	4.00	3.84

Table 5: Performance of evaluated models on DITING-Corpus using AgentEval. S, G1, and G2 are mean scores on a 0–2 scale; Σ is their sum (0–6); Average Score- Σ is the mean Σ across the six dimensions.

same time, scale alone does not determine success in web novel translation. Qwen3-8B outperforms the much larger LLaMA3-70B overall, 3.96 versus 3.58, suggesting that cultural-linguistic alignment with the source domain can outweigh parameter count when translation depends on genre-sensitive and culturally grounded interpretation.

The results also highlight the importance of training strategy, particularly reinforcement learning. Seed-X-PPO-7B improves over its instruction-tuned counterpart by 0.65 points, 4.65 versus 4.00. This improvement is not merely global, but is concentrated in dimensions requiring pragmatic and culturally grounded reasoning. Seed-X-PPO-7B is among the strongest models on Idiom Translation and Lexical Ambiguity, scoring 5.42 and 5.46, respectively, and in the former even surpassing DeepSeek-V3. This suggests that targeted optimization can induce strong sentence-level cultural reasoning even in small architectures.

Dimension-wise Findings. The dimension-by-dimension results further clarify where current systems have progressed and where they remain limited. DeepSeek-V3 and GPT-4o lead on Idiom Translation, scoring 5.14 and 5.04, and on Lexical Ambiguity, scoring 5.52 and 5.50, showing stronger figurative interpretation and context-sensitive disambiguation. DeepSeek-V3 also ranks

first on Terminology Localization at 5.12, followed by Qwen3-32B and Seed-X-PPO-7B, suggesting that scale and domain adaptation help handle world-building terminology and genre-specific concepts.

Tense Consistency is comparatively well handled, with most models scoring above 4.6 on this dimension. Relative to more culturally or pragmatically implicit phenomena, temporal interpretation appears easier for current systems to model. This suggests that not all forms of sentence-level meaning transfer are equally difficult, and that current progress is strongest where narrative interpretation can rely on relatively stable contextual cues.

By contrast, Zero-Pronoun Resolution remains the clearest unresolved bottleneck, with the widest benchmark spread: from 5.56 for GPT-4o and 5.06 for DeepSeek-V3 to 2.22 for IFLYTEK Translate. This suggests that what most separates systems is not figurative language alone, but the ability to infer and restore implicit referents from local narrative context. This is especially important in Chinese web novel translation, where omitted arguments are common and directly affect character reference, action structure, and discourse continuity.

Cultural Safety exhibits a different profile, with narrow variance and rankings that do not closely track overall performance. DeepSeek-R1-70B scores highest on this dimension at 4.74 without

being the overall leader, suggesting that safety-oriented alignment can improve value-sensitive adaptation in targeted ways but does not necessarily enhance other genre-critical abilities. More broadly, current systems differ less in value-sensitive adaptation than in culturally grounded inference, interpretation, and referential recovery.

Taken together, these findings show that leading web novel translation systems have distinct competence profiles rather than uniformly stronger translation ability. Aggregate scores therefore overstate progress: current gains reflect selective improvement in specific sentence-level cultural abilities, not a general solution to cultural meaning transfer. **Evaluation Reliability.** To mitigate potential model-family bias and test whether rankings depend on a single evaluator backbone, we conduct two additional reliability checks: a GPT-4o-based AgentEval re-evaluation and a human validation study. After correcting for GPT-4o self-preference, the alternative ranking remains broadly consistent with Table 5 (Appendix C.13). For human validation, two senior annotators judge 1,440 DITING-Corpus translations spanning 12 models, and AgentEval shows strong agreement with their preferences (Appendix C.12). Together with Table 4, these checks support the robustness of AgentEval and the conclusions drawn from Table 5.

4 Related Work

4.1 Evaluation Metrics for Translation

Translation evaluation metrics have primarily focused on non-literary texts. Common metrics range from overlap-based measures such as BLEU (Papineni et al., 2002) and chrF (Popović, 2015) to learned or contextual metrics such as COMET (Rei et al., 2020), BLEURT (Sellam et al., 2020), and BERTScore (Zhang et al., 2019). While the latter improve sensitivity to semantic similarity beyond surface-level overlap, these general-purpose metrics are not explicitly designed to assess the culturally and pragmatically grounded phenomena central to literary translation. Large language models (LLMs) are increasingly used for translation evaluation. However, existing LLM-based evaluators are generally designed for broad translation quality assessment and provide limited explicit diagnosis of culturally and pragmatically grounded phenomena. General-purpose approaches such as GEMBA-MQM (Kocmi and Federmann, 2023) and M-MAD (Feng et al., 2025) improve fine-grained

MT evaluation, while LITRANSPOQA (Zhang et al., 2025b) specifically targets literary translation. Nevertheless, these methods are not explicitly organized around the web-novel-specific cultural and pragmatic phenomena considered in our framework. Human-centered frameworks such as MQM (Lommel et al., 2014) provide detailed error analysis but require substantial human effort, motivating more scalable evaluation approaches that retain fine-grained, expert-informed criteria.

Beyond general-purpose metrics, the six phenomena targeted by DITING have also been studied individually in the broader MT literature, but are rarely evaluated jointly in web-novel translation.

Idiom Translation. Idioms are often non-compositional, making literal translation prone to losing figurative meaning (Fadaee et al., 2018). Prior studies show that NMT systems frequently produce compositional translations when an idiomatic reading is required, particularly for unseen expressions (Dankers et al., 2022). Targeted evaluation (Baziotis et al., 2023), knowledge-base augmentation such as IdiomKB (Li et al., 2024), and slang benchmarks such as SlangDIT (Liang et al., 2025) further extend idiomatic translation to LLMs. However, sustained narrative use of genre-specific idioms and proverbs in Chinese web fiction remains underexplored.

Lexical Ambiguity. Word-sense ambiguity has long been a challenge for MT. Contrastive benchmarks such as ContraWSD, MuCoW, and DiBiMT evaluate whether systems select the intended sense and reveal persistent biases toward frequent meanings (Rios Gonzales et al., 2017; Raganato et al., 2020; Campolungo et al., 2022). These resources mainly target general-domain nouns and verbs, whereas web novels frequently involve internet slang and newly coined expressions whose meanings depend strongly on genre and narrative context.

Terminology Localization. Terminology translation has been studied through WMT shared tasks, constraint-based methods, and retrieval-augmented LLM prompting (Alam et al., 2021; Semenov et al., 2023; Dinu et al., 2019; Moslem et al., 2023). These approaches improve adherence to supplied terminology, but evaluation often emphasizes whether a prescribed term is used. Web novels additionally require localization of invented, religious, and internet-born terms, where literal term matching may not reflect culturally or genre-

appropriate rendering.

Tense Consistency. Tense and aspect remain challenging when temporal relations are expressed differently across languages. TeCS provides a dedicated benchmark for French–English tense consistency (Ai et al., 2023), while document-level MT research also identifies tense and aspect as persistent discourse-level phenomena (Maruf et al., 2021). Dedicated evaluation remains limited for Chinese–English literary translation, where Chinese lacks obligatory tense inflection and temporal interpretation often depends on narrative context.

Zero-Pronoun Resolution. Chinese frequently omits subject or object pronouns that must be recovered in English translation. Prior work studies zero-pronoun detection and reconstruction through reconstruction-based and joint prediction-and-translation methods (Wang et al., 2023a). GuoFeng provides a dedicated benchmark that already includes web fiction (Xu et al., 2022), while discourse-level test suites examine context-dependent anaphora and deixis (Bawden et al., 2018; Voita et al., 2019). These resources, however, primarily isolate pronoun and reference phenomena rather than situating them within broader cultural and pragmatic evaluation of web-novel translation.

Cultural Safety. Recent work approaches this area from two related but distinct directions. Culturally aware MT evaluates culture-specific expressions and reveals persistent weaknesses on culturally nuanced content, including idioms and puns (Yao et al., 2024; Van Doren et al., 2026). Separately, multilingual safety research reveals substantial cross-lingual inconsistencies in LLM safety behavior (Friedrich et al., 2025). Web novels frequently combine culturally specific language with violence, romance, and other sensitive content, placing translation at the intersection of cultural appropriateness and multilingual safety; their joint evaluation remains underexplored.

4.2 Benchmarks for Web Novels

Recent benchmarks for Chinese–English web novel translation include BWB (Jiang et al., 2022), which provides 196K chapters with document-level context, and the WMT23 Discourse-Level Literary Translation (DLT) shared task (Wang et al., 2023b), whose associated GuoFeng Webnovel Corpus contains 22.6K chapters with professional translations and controlled splits for training and evaluation. These resources provide substantial document-level context, but primarily target document-level

MT and discourse evaluation rather than explicit phenomenon-level diagnosis of sentence-level cultural meaning. Existing evaluations combine automatic and human assessment, yet are not explicitly organized around the fine-grained cultural and pragmatic phenomena considered in our framework. Although LLMs are increasingly benchmarked for translation (Yan et al., 2024), systematic evaluation of web-novel-specific cultural and pragmatic challenges remains limited, motivating more targeted benchmarks.

5 Conclusion

We present DITING, the first comprehensive evaluation framework for directly assessing sentence-level cultural meaning in web novel translation. By combining a six-dimension benchmark, a dimension-aware corpus, and the multi-agent evaluator AgentEval, DITING moves evaluation beyond surface overlap and overall quality toward narrative and cultural fidelity. AgentEval provides a validated, rubric-conditioned foundation for this domain, achieving the strongest correlation with human judgments among compared metrics, while remaining robust across evaluator backbones and generalizing to German, French, and Spanish. Applying DITING to 14 translation systems shows that current progress is substantial but uneven: frontier large language models lead overall, with DeepSeek-V3 ranking best in aggregate, but their gains concentrate in idiom translation, lexical ambiguity, and tense consistency, while zero-pronoun resolution remains the clearest bottleneck. These findings suggest that aggregate rankings may overestimate current systems’ ability to handle sentence-level cultural meaning. Future work will extend DITING to document-level narrative coherence, broader language-pair validation, and an evolvable multi-agent deliberation framework.

Limitations

Despite its contributions, this study has several limitations. First, while MetricAlign is carefully curated for diagnostic difficulty, its source-side scale is compact, limiting broader diversity across genres, styles, and narrative contexts. Second, DITING is designed as a sentence-pair diagnostic benchmark rather than a complete replacement for document-level evaluation. Sentence pairs provide the smallest actionable units for attributing errors in terminology localization, idiom translation, lexical am-

biguity, zero-pronoun recovery, tense consistency, and cultural safety. However, even with local context, this setting cannot fully assess document-level consistency, such as global terminology consistency, long-range coreference, entity tracking, temporal continuity, and style coherence. Third, our multi-agent framework currently lacks dynamic coordination and learning. Recent work on feedback-driven LLM agents highlights the potential of iterative feedback for adaptive refinement (Miao et al., 2026), which we leave for future study. Finally, we acknowledge that AgentEval is far from a perfect substitute for human experts. While it achieves the highest human correlation, it still occasionally struggles with deep cultural nuances. However, its significance lies in establishing a pioneering baseline. As the first rubric-conditioned, multi-agent framework for culturally grounded translation, it shifts the paradigm from opaque surface metrics to interpretable, fine-grained diagnosis, laying an indispensable foundation for future advanced evaluators.

Ethics Statement

This study was conducted in accordance with established ethical guidelines for research. The source materials used in this study were publicly accessible through their respective online platforms and contain no personally identifiable information. No personally identifiable information was collected from annotators. We manually reviewed the collected data for personally identifiable information and sensitive content during corpus curation. No personally identifiable information was retained or released. Professional annotators were recruited based on their experience in literary translation. All annotators were compensated for their annotation work at rates exceeding the applicable local minimum wage. Participation was voluntary, with the option to withdraw at any time. We also provided clear annotation guidelines and limited task durations to avoid excessive workload. All annotators were informed of the research purpose and the intended use of their annotations before participation and provided informed consent.

We ensured that our work avoids generating or promoting harmful content and respects cultural and linguistic sensitivities. We strictly follow the original licenses and intended use of all datasets, and our usage complies fully with their ethical and legal requirements. Potential risks include

the handling of sensitive content present in some datasets, and the possibility that our evaluation metric may inadvertently misrepresent translation quality, which should be considered when interpreting or applying the results.

Our original code and released annotations and metadata are provided for research and educational purposes under the MIT License. Third-party source texts, models, and other external resources remain subject to their original licenses and terms of use. Due to copyright restrictions, we do not redistribute the original web-novel text. Code and release materials are available at: <https://github.com/whunextgen/DITING>.

Acknowledgments

This work is supported by the Key Program of the National Natural Science Foundation of China (Grant No. U23A20316), the General Program of the National Natural Science Foundation of China (Grant No. 62676290), CCF-Tencent Rhino-Bird Open Research Fund (CCF-Tencent RAGR20250115), Wuhan Natural Science Foundation Exploratory Program (Morning Light Program) Project (2026040301020029) and the Intelligent Computing Center of the National Cybersecurity Talent and Innovation Base, Wuhan.

The work leading to these results has been funded by the European Union under Grant Agreement No. 101289094 (project AIRIS). Views and opinions expressed are however those of the authors and do not necessarily reflect those of the European Union or the granting authority (HaDEA). Neither the European Union nor the granting authority can be held responsible for them.

References

- Yiming Ai, Zhiwei He, Kai Yu, and Rui Wang. 2023. [TeCS: A dataset and benchmark for tense consistency of machine translation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1930–1941, Toronto, Canada. Association for Computational Linguistics.
- Md Mahfuz Ibn Alam, Ivana Kvapilíková, Antonios Anastasopoulos, Laurent Besacier, Georgiana Dinu, Marcello Federico, Matthias Gallé, Kweonwoo Jung, Philipp Koehn, and Vassilina Nikoulina. 2021. [Findings of the WMT shared task on machine translation using terminologies](#). In *Proceedings of the Sixth Conference on Machine Translation*, pages 652–663, Online. Association for Computational Linguistics.

- Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. [Evaluating discourse phenomena in neural machine translation](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.
- 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.
- Niccolò Campolungo, Federico Martelli, Francesco Saina, and Roberto Navigli. 2022. [DiBiMT: A novel benchmark for measuring Word Sense Disambiguation biases in Machine Translation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4331–4352, Dublin, Ireland. Association for Computational Linguistics.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.
- Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. [Can transformer be too compositional? analysing idiom processing in neural machine translation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.
- Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. [Training neural machine translation to apply terminology constraints](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy. Association for Computational Linguistics.
- Cristina España-Bonet, Juliane Stiller, Roland Ramthun, Josef van Genabith, and Vivien Petras. 2018. Query translation for cross-lingual search in the academic search engine pubpsych. In *Research Conference on Metadata and Semantics Research*, pages 37–49. Springer.
- Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2018. [Examining the tip of the iceberg: A data set for idiom translation](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Zhaopeng Feng, Jiayuan Su, Jiamei Zheng, Jiahua Ren, Yan Zhang, Jian Wu, Hongwei Wang, and Zuozhu Liu. 2025. [M-MAD: Multidimensional multi-agent debate for advanced machine translation evaluation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7084–7107, Vienna, Austria. Association for Computational Linguistics.
- Felix Friedrich, Simone Tedeschi, Patrick Schramowski, Manuel Brack, Roberto Navigli, Huu Nguyen, Bo Li, and Kristian Kersting. 2025. [Llms lost in translation: M-alert uncovers cross-linguistic safety inconsistencies](#). *Preprint*, arXiv:2412.15035.
- Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 33–41.
- Yuchen Eleanor Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Mrinmaya Sachan, and Ryan Cotterell. 2022. [A bilingual parallel corpus with discourse annotations](#). *Preprint*, arXiv:2210.14667.
- Alexander Jones and Derry Wijaya. 2021. Sentiment-based candidate selection for NMT. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 188–201, Virtual. Association for Machine Translation in the Americas.
- Marzena Karpinska and Mohit Iyyer. 2023. [Large language models effectively leverage document-level context for literary translation, but critical errors persist](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 419–451, Singapore. Association for Computational Linguistics.
- Tom Kocmi and Christian Federmann. 2023. [GEMBA-MQM: Detecting translation quality error spans with GPT-4](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.
- Shuang Li, Jiangjie Chen, Siyu Yuan, Xinyi Wu, Hao Yang, Shimin Tao, and Yanghua Xiao. 2024. Translate meanings, not just words: Idiomkb’s role in optimizing idiomatic translation with language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18554–18563.
- Yunlong Liang, Fandong Meng, Jiaan Wang, and Jie Zhou. 2025. [Slangdit: Benchmarking llms in interpretative slang translation](#). *Preprint*, arXiv:2505.14181.
- Arle Lommel, Aljoscha Burchardt, Maja Popović, Kim Harris, Eleftherios Avramidis, and Hans Uszkoreit. 2014. [Using a new analytic measure for the annotation and analysis of MT errors on real data](#). In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, pages 165–172, Dubrovnik, Croatia. European Association for Machine Translation.
- Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2021. [A survey on document-level neural machine translation: methods and evaluation](#). *ACM Computing Surveys*, 54(2).

- Chunyu Miao, Henry Peng Zou, Yangning Li, Yankai Chen, Yibo Wang, Fangxin Wang, Yifan Li, Woosong Yang, Bowei He, Xinni Zhang, and 1 others. 2026. Recode-h: A benchmark for research code development with interactive human feedback. In *International Conference on Learning Representations*, volume 2026, pages 70142–70194.
- Yasmin Moslem, Gianfranco Romani, Mahdi Molaei, John D. Kelleher, Rejwanul Haque, and Andy Way. 2023. [Domain terminology integration into machine translation: Leveraging large language models](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 902–911, Singapore. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.
- Maja Popović. 2015. [chrF: character n-gram F-score for automatic MT evaluation](#). In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.
- Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. 2020. [An evaluation benchmark for testing the word sense disambiguation capabilities of machine translation systems](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3668–3675, Marseille, France. European Language Resources Association.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. [COMET: A neural framework for MT evaluation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.
- Research Group of the Institute of Literature, Chinese Academy of Social Sciences. 2025. [2024 china online literature development research report](#). China Social Sciences Network. Published May 13, 2025; accessed August 24, 2026.
- Annette Rios Gonzales, Laura Mascarell, and Rico Sennrich. 2017. [Improving word sense disambiguation in neural machine translation with sense embeddings](#). In *Proceedings of the Second Conference on Machine Translation*, pages 11–19, Copenhagen, Denmark. Association for Computational Linguistics.
- Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. [BLEURT: Learning robust metrics for text generation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.
- Kirill Semenov, Vilém Zouhar, Tom Kocmi, Dongdong Zhang, Wangchunshu Zhou, and Yuchen Eleanor Jiang. 2023. [Findings of the WMT 2023 shared task on machine translation with terminologies](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 663–671, Singapore. Association for Computational Linguistics.
- Charles Spearman. 1904. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.
- Russell Taylor, Benjamin Herbert, and Michael Sana. 2026. [Pun intended: Multi-agent translation of word-play with contrastive learning and phonetic-semantic embeddings for clef joker 2025 task 2](#). *Preprint*, arXiv:2507.06506.
- Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. 2022. [Exploring document-level literary machine translation with parallel paragraphs from world literature](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9882–9902, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Madison Van Doren, Casey Ford, Jennifer Barajas, Riley VanMeter, and Cory Holland. 2026. [“be my cheese?”: Cultural nuance benchmarking for machine translation in multilingual LLMs](#). In *Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM)*, pages 52–76, San Diego, California, USA. Association for Computational Linguistics.
- Elena Voita, Rico Sennrich, and Ivan Titov. 2019. [When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.
- Longyue Wang, Siyou Liu, Chenyang Lyu, Wenxiang Jiao, Xing Wang, Jiahao Xu, Zhaopeng Tu, Yan Gu, Weiyu Chen, Minghao Wu, Liting Zhou, Philipp Koehn, Andy Way, and Yulin Yuan. 2024. [Findings of the WMT 2024 shared task on discourse-level literary translation](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 699–700, Miami, Florida, USA. Association for Computational Linguistics.
- Longyue Wang, Siyou Liu, Mingzhou Xu, Linfeng Song, Shuming Shi, and Zhaopeng Tu. 2023a. [A survey on zero pronoun translation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3325–3339, Toronto, Canada. Association for Computational Linguistics.
- Longyue Wang, Zhaopeng Tu, Yan Gu, Siyou Liu, Dian Yu, Qingsong Ma, Chenyang Lyu, Liting Zhou, Chao-Hong Liu, Yufeng Ma, Weiyu Chen, Yvette Graham, Bonnie Webber, Philipp Koehn, Andy Way, Yulin Yuan, and Shuming Shi. 2023b. [Findings of the WMT 2023 shared task on discourse-level literary](#)

translation: A fresh orb in the cosmos of LLMs. In *Proceedings of the Eighth Conference on Machine Translation*, pages 55–67, Singapore. Association for Computational Linguistics.

Mingzhou Xu, Longyue Wang, Derek F. Wong, Hongye Liu, Linfeng Song, Lidia S. Chao, Shuming Shi, and Zhaopeng Tu. 2022. [GuoFeng: A benchmark for zero pronoun recovery and translation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11266–11278, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jianhao Yan, Pingchuan Yan, Yulong Chen, Jing Li, Xianchao Zhu, and Yue Zhang. 2024. [Benchmarking gpt-4 against human translators: A comprehensive evaluation across languages, domains, and expertise levels](#). *Preprint*, arXiv:2411.13775.

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.

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

Ran Zhang, Wei Zhao, Lieve Macken, and Steffen Eger. 2025b. [LiTransProQA: An LLM-based literary translation evaluation metric with professional question answering](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29099–29121, Suzhou, China. Association for Computational Linguistics.

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

Henry Peng Zou, Wei-Chieh Huang, Yaozu Wu, Jizhou Guo, Yankai Chen, Chunyu Miao, Hoang H Nguyen, Yue Zhou, Weizhi Zhang, Liancheng Fang, and 1 others. 2026. Llm-based human-agent collaboration and interaction systems: A survey. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 36335–36364.

A Dataset Construction Details

To transform billions of chapter-level bilingual passages into high-quality sentence pairs, we applied a combination of automated processing and manual verification. The process consisted of several steps:

A.1 Automated preprocessing of massive-scale novels

All chapters were first processed using a rigorous automated pipeline designed to normalize and clean noisy web-scraped text. We utilized Unicode normalization forms (NFKC) and regular expressions to standardize character widths and correct encoding errors. For segmentation, specialized sentence splitters—such as PySBD for English and rule-based logic optimized for Chinese quotation marks—were employed to respect narrative boundaries and prevent splitting dialogue mid-sentence. To ensure data integrity, we computed perplexity scores using lightweight n-gram language models (KenLM); chapters exhibiting perplexity values exceeding two standard deviations from the mean were automatically flagged as low-quality (e.g., gibberish or severe OCR errors) and discarded. Additionally, a deduplication step using MinHash and Locality Sensitive Hashing (LSH) with a Jaccard similarity threshold of 0.85 was applied to remove duplicate content across different sources, preparing a clean and unique corpus for alignment.

A.2 Sentence-level Alignment

Alignment was performed in a coarse-to-fine manner to balance efficiency with semantic precision. First, rule-based alignment leveraged chapter headers and paragraph breaks as hard anchors to partition the text. Within these boundaries, we applied a modified Gale-Church algorithm based on character length ratios, supplemented by lexical matching of named entities, to generate initial candidate pairs and filter out passages with significant structural discrepancies. Next, semantic alignment refined these candidates using the LaBSE (Language-agnostic BERT Sentence Embeddings) model. We calculated cosine similarity between source and target sentence embeddings, discarding pairs below a strict threshold of 0.75. To address the many-to-one and one-to-many translations common in literary texts, we implemented a sliding window approach to evaluate [1:2] and [2:1] combinations, while a dynamic programming algorithm enforced monotonicity constraints to ensure the aligned sen-

tences strictly followed the linear narrative order of the novels.

A.3 Sentence-level Pairing, Phenomenon Identification, and Handling Overlapping Phenomena

For each of the six rubric dimensions, we first identify candidate sentence pairs that are likely to exhibit the corresponding web-novel translation (WNT) challenges using lightweight automatic filtering and heuristics. Professional annotators then manually verify these candidates and assign appropriate phenomenon labels. This stage is crucial for differentiating between "data curation" and "MT-output scoring", ensuring that the data preparation step is clearly separated from the later MT-output evaluation. Additionally, when multiple phenomena co-occur within the same sentence, we treat the dimensions as orthogonal. A sentence can therefore be evaluated under multiple dimensions, with its MT output scored separately for each relevant dimension. This approach prevents forcing a single "dominant" label, maintaining the interpretability of the evaluation and avoiding the conflation of different error types.

A.4 Automated verification with harmful content filtering

To ensure the dataset was both high-quality and safe for training, we established a multi-layered verification protocol. Initially, the Aho-Corasick algorithm was used for high-performance string matching against a curated blocklist of over 50,000 sensitive keywords, efficiently removing explicit profanity and personally identifiable information (PII). Beyond keyword matching, we deployed a fine-tuned XLM-RoBERTa toxicity classifier to detect semantic toxicity, automatically quarantining texts with a toxicity probability above 0.8. Finally, a human-in-the-loop mechanism was introduced where 1% of the aligned data was stratified and sampled for manual inspection. This step verified alignment accuracy—checking for hallucinated translations—and removed any residual undesirable content, effectively converting massive-scale chapter-level bilingual data into a pristine sentence-level parallel dataset.

A.5 Constructing Gold-Standard Translations

We construct gold-standard English references through a rigorous human-verification and post-editing pipeline to explicitly distinguish our bench-

mark from noisy crawled data. Initially, parallel data collected from online Chinese web novel platforms and their corresponding English translations served as raw candidate translations.

Importantly, these crawled English translations were not directly used as gold-standard references. Instead, professional translators with extensive experience in web-novel translation manually checked each aligned source-target pair. They comprehensively revised the candidates to correct mistranslations, omissions, unnatural expressions, and alignment-related errors. Where original translations were entirely inadequate, LLM-generated drafts were occasionally used as additional translation scaffolds for post-editing. This rigorous process ensures fidelity, naturalness, and the accurate realization of targeted phenomena (e.g., idioms, terminology localization, zero pronouns). A second translator subsequently reviewed each revision. Any disagreements were resolved through adjudication, resulting in the final, human-verified gold-standard references.

A.6 Constructing the Cultural Safety Subset

Prior evaluation work on web-novel translation has focused mainly on linguistic difficulty, whereas culturally safety-sensitive rendering has received much less explicit treatment as a standalone benchmark dimension. DITING addresses this gap by introducing *Cultural Safety*, which targets translation risks arising in context-sensitive cases such as sexual content, profanity, hate expressions, violence, drugs, coercion, and cyber-misuse cues. The current subset contains 100 manually curated instances. Its relatively small scale is mainly due to the high cost of candidate discovery and expert validation: such cases are low-frequency, context-dependent, and often obscured by moderation, censorship, rewriting, or reposting across platforms, making large-scale automatic collection unreliable. As summarized in Table 6, the present set already covers a diverse set of genre clusters and sensitive-content types, while also revealing the main bottlenecks that guide our future expansion.

These observations highlight several directions for systematically expanding the Cultural Safety dimension in future releases. In particular, we plan to increase genre diversity and platform-specific variants, introduce more boundary and intensity-graded samples for sensitive content, and collect additional long-tail cases where safety judgments depend strongly on discourse context. These ef-

Genre cluster	Coverage	Sensitive content	Scaling challenge	Future expansion
Urban, campus, workplace	High	Sexual content; profanity; gender insults	Heavy slang and strong context dependence	More regional/platform variants; more boundary cases
Gang, underground economy	Med-High	Drugs; violence; sexual exploitation cues	Cross-post edits/censorship; needs expert screening	More euphemisms/argot; more rewritten versions
Erotica fiction	Med	Explicit sex; coercion cues	Literal vs. sanitized translation trade-off	More intensity-graded boundary samples
Cyber, espionage	Med	Hacking; infra-attack intent; credentials-like strings	Misuse risk is contextual; keywords are unreliable	More high-stakes long-tail cases; more implicit mentions
History, conflict	Low-Med	Hate slurs; incitement	Thin line between narration vs. hate	More periods/groups; more slur-intensity mapping cases
Fantasy, xianxia	Low-Med	Abuse/violence; sexist framing	Often indirect or figurative	More “wrapped” coercion/discrimination cases
Tomb-raiding, folklore	Low	Illicit acts; taboos	Many weak/invalid candidates	Focus on culturally taboo, high-disagreement cases

Table 6: Current coverage, scaling challenges, and future expansion of the Cultural Safety subset.

forts aim to scale the Cultural Safety subset while maintaining rigorous expert validation.

B Annotation Details

B.1 Annotation Guidelines

Task Objective. This project aims to evaluate the quality of web novel translations generated by different models, based on an established evaluation framework and corresponding datasets. Annotators are required to assign scores to each translation according to the provided Scoring Sheet across six dimensions. Each dimension consists of one specific metric (S) and two general metrics (G1, G2). The justification for using two general scores is to avoid treating overall quality as a single, undifferentiated judgment. While S measures whether the core phenomenon is correctly handled, G1 and G2 assess how this handling affects broader sentence-level semantic coherence, structural consistency, and fluency. For example, a translation might preserve the correct temporal relation (High S) but feature awkward structural sequencing (Low G1), or it might mistranslate an idiom (Low S) while maintaining a highly fluent English expression (High G2). Separating these aspects allows for a finer-grained diagnosis of how localized cultural and linguistic phenomena interact with English readability. Ratings should be given on a 0–2 scale, accompanied by short justifications or annotation notes when necessary.

The objectives are:

- To independently assess model performance in Idiom Translation, Lexical Ambiguity, Terminology Localization, Tense Consistency, Zero-Pronoun Resolution, and Cultural Safety.
- To ensure consistent evaluation criteria and minimize subjective bias.

Annotation Procedure.

1. Read the model-generated translation and analyze it in reference to the source text and reference translation.
2. Evaluate the output according to the six dimensions and their associated scoring criteria.
3. Record the score (0/1/2) for each metric in the scoring sheet and compute the total score. Provide optional comments if clarification is necessary.
4. Double-check the total score for accuracy.

Evaluation Dimensions and Criteria. The six-dimensional scoring criteria are detailed in Table 7 and Table 8. For Terminology Localization, ‘intelligible’ means that the term’s in-world function is understandable to target readers, while “culturally appropriate” means that the rendering follows target-language genre conventions; acceptable strategies may include transliteration, semantic translation, or brief glossing depending on the term type.

B.2 Pilot Evidence for Dimension Frequency

To further examine whether the six DITING dimensions occur as ordinary web-novel phenomena rather than isolated examples, we conduct a small-scale expert annotation on one representative chapter. As shown in Table 28, the sampled chapter contains dense and overlapping instances of the six dimensions: 84 out of 87 sentence-level units contain at least one DITING phenomenon, 74 contain multiple phenomena, and all six dimensions are observed. This pilot is not intended as a full prevalence study, but provides direct evidence that these phenomena are not merely isolated examples.

B.3 Annotator Demography

The construction of our DITING dataset relies on the linguistic expertise of a team of highly qualified annotators with extensive backgrounds in translation studies and cross-lingual communication. Their professional training and experience in multilingual translation ensure accurate and contextually grounded annotations across diverse stylistic and cultural expressions in web novel texts.

The annotation team for **ZH-EN translation** consists of three core members. Two of the annotators are professional translators from leading Chinese translation companies, each possessing over five years of experience in ZH-EN literary translation and linguistic quality assessment. Their prior work includes numerous literary translation projects, equipping them with the necessary skills to handle the typical cultural expressions found in web novels. The third annotator is a student majoring in translation at a prestigious Chinese university. The student participated both as an annotator and a quality supervisor, responsible for performing annotation tasks, coordinating consistency checks, and revising annotation guidelines based on feedback from the team.

For **other languages**, the annotation process was carried out by professional translators, each with over five years of experience in literary translation. These translators, selected for their expertise in translating web novels, were not limited to the original two but included additional specialists in various languages who ensured the dataset’s linguistic fidelity and contextual consistency.

The team followed a well-structured annotation workflow. Weekly annotation meetings were held to discuss challenging cases and refine annotation criteria. Feedback from translators was systemati-

cally integrated into guideline updates, enhancing both reliability and linguistic validity throughout dataset construction.

Together, the team’s professional translation expertise, linguistic sensitivity, and collaborative workflow enabled the creation of a high-quality web novel translation benchmark. Their efforts ensured that the dataset is both linguistically faithful and contextually consistent, establishing a solid foundation for future research on evaluating LLM-based literary translation.

B.4 Annotation Example

We show some cases which can demonstrate our annotation guideline in Table 11.

B.5 Annotation Process

Our annotation process can be seen in Figure 4.

C Experimental Details

C.1 Benchmark Settings and Evaluated Models.

Using DITING-Corpus together with the validated AgentEval framework based on Debate-R1, we comprehensively evaluate 14 representative systems for web novel translation. The benchmark covers a diverse set of models, including proprietary and open-source systems, multilingual large language models, machine-translation-oriented large language models, and commercial translation systems. A complete list of evaluated models and configurations is provided in Appendix C.2⁷. Proprietary systems are accessed through their official application programming interfaces, while open-source models are tested using their publicly released checkpoints under default decoding settings.

For all systems, we ensure fair access to the local narrative context H , which is obtained from the surrounding text within the same web novel chapter (e.g., the preceding and succeeding sentences or paragraphs). To accommodate different model interfaces while maintaining fairness, we adapt the input format accordingly. For LLM-based translators, H is explicitly included in the prompt alongside the target source segment. For commercial MT systems (e.g., Google Translate) that only accept a single text block, we concatenate the local context and the target segment into one string, clearly

⁷Since DeepSeek-R1 is the evaluator backbone of AgentEval, it is not considered an evaluated translation model.

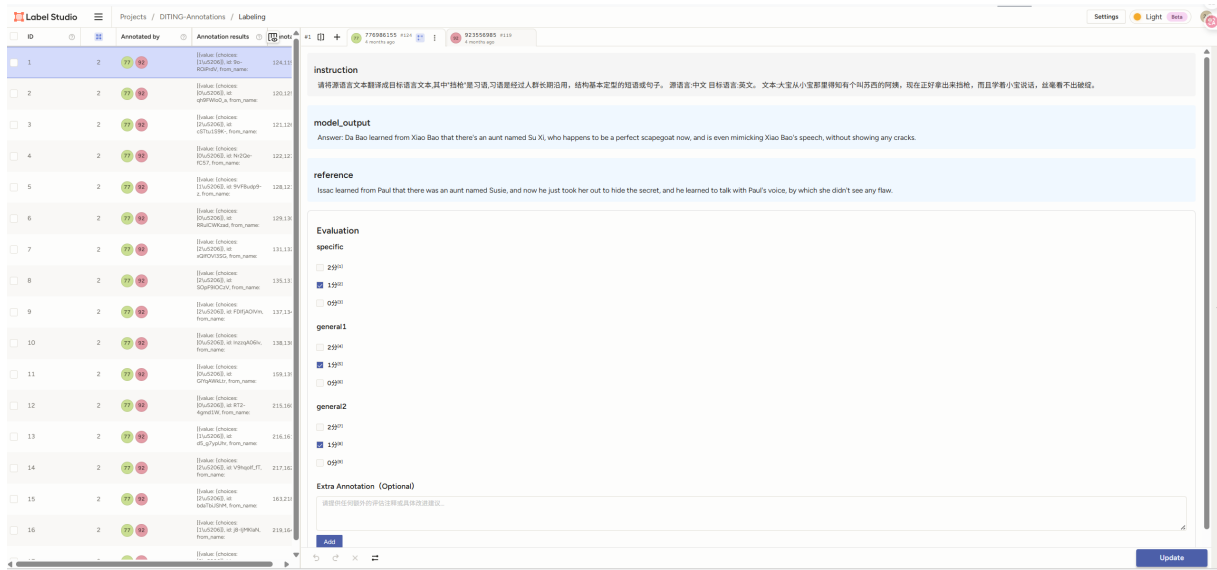


Figure 4: The Label Studio interface of the DITING annotation process.

wrapping the target sentence with special structural markers. After translation, we extract only the output corresponding to the marked segment for evaluation. This setup better reflects realistic web novel translation, where interpretation depends on short-range narrative continuity, and guarantees that all models utilize identical contextual information for phenomenon disambiguation. Additional implementation details are provided in Appendix C.

C.2 Evaluated Translation Models

Frontier LLMs include industry-leading APIs:

- GPT-4o, known for its strong general-purpose multilingual and reasoning abilities;
- DeepSeek-V3, which is optimized for Chinese and English tasks with enhancements in coding and mathematics;

For open-source models, we select large-scale foundation models:

- DeepSeek-R1-70B, which offers strong performance in complex reasoning;
- LLaMA3-70B, Meta’s top-tier open-source model;
- ChatGLM4-9B, optimized for dialogue scenarios;
- Qwen3 series, including Qwen3-8B, Qwen3-14B, and Qwen3-32B;
- LLaMA3-8B, which balances efficiency and capability.

Additionally, translation-specialized models are incorporated, including:

- Google Translate, a widely-used commercial machine translation service based on large-scale neural methods;
- IFLYTEK Translate, a leading Chinese-oriented translation system;
- ByteDance’s Seed-X-PPO-7B, fine-tuned with reinforcement learning;
- Seed-X-Instruct-7B; instruction-tuned for translation;
- Xiaomi’s GemmaX2-28-9B.

This selection ensures a diverse and representative evaluation across model types, scales, and specializations.

C.3 Translation models used in MetricAlign

Table 9 lists the translation models used in the MetricAlign framework. These models are categorized into three groups based on their accessibility and specialization.

C.4 Evaluation of Automatic Metrics

We provide the detailed configuration of all the metrics used in our metrics Evaluation experiment in Table 21.

C.5 Debate Round Statistics

The distribution of debate rounds was as follows: 95.0% of evaluations concluded in a single round, 4.0% proceeded to a second round, and only 0.5% each advanced to a third or fourth round. This distribution demonstrates the efficiency of our evaluation design, as the vast majority of assessments reached a clear conclusion in the first round, while only a tiny fraction of complex or ambiguous cases required multi-round discussion. Instances requiring a third round or beyond are exceptionally rare, providing statistical support for the rationale of setting the maximum round limit at four.

C.6 Details of Multilingual Applicability Evaluation

C.6.1 Data and Experimental Setup

To verify whether AgentEval can generalize beyond Chinese–English (ZH–EN) translation, we extended our evaluation to three other target languages: German (DE), French (FR), and Spanish (ES), with Chinese remaining as the source language.

Data Source. We sampled 30 representative sentences from the DITING-Corpus, covering all dimensions. These Chinese sentences were translated into German, French, and Spanish using the random LLMs listed in Table 9. This resulted in a total of 270 evaluation samples across the three languages.

Annotation. Qualified language experts annotated the translation quality following the MetricAlign protocol.

C.6.2 Result

AgentEval_{Debate-R1} and AgentEval_{DeepSeek-R1} were used to validate the generalizability of our framework. The quantitative results are summarized in Table 18.

C.6.3 Case Study

We present a case study in Table 10 to demonstrate the robustness and generalizability of our framework in multilingual settings.

C.7 AgentEval Cost and Latency Analysis

In this section, we report the average evaluation cost and latency per instance of AgentEval across different backbone models. The cost data and the latency data are shown in Table 12.

C.8 Case Study of AgentEval

We demonstrated how a case was evaluated within our framework in Table 13(1-ROUND Debate), Table 14(2-ROUND Debate), Table 15(3-ROUND Debate), Table 16(4-ROUND Debate).

We provide an example comparison to illustrate how DeepSeek-V3 outperforms GPT-4o in the temporal dimension, thus proving the conclusion of table 5 in table 17.

C.9 Qualitative Analysis of LITRANSPOQA

We present several representative cases in Table 23, where LITRANSPOQA exhibits substantial deviations from human judgments.

C.10 Qualitative Analysis of M-MAD

Following the quantitative results in Section 3.1, we conducted a qualitative error analysis to understand why M-MAD, despite utilizing a multi-agent framework, achieves a relatively weak correlation with human judgments in our setting. We inspected representative large-gap cases and identified several challenges when directly applying general MQM-style evaluations to web novel translation (WNT). These failure modes are summarized in Table 24.

C.11 Prompts

We show our prompts used in the evaluation framework, which can be seen in Table 19, Table 20.

C.12 Manual Verification

We manually evaluated 1,440 translation outputs from DITING-Corpus across 12 models, reporting both the scores and corresponding model rankings (rows labeled “-Ranking”). Two senior annotators labeled all outputs following our multi-dimensional criteria, achieving high consistency. The annotation consistency can be found in Table 22, and the results of manual verification can be found in Table 25. Figure 5 shows the absolute ranking differences between AgentEval and human evaluation.

C.13 Cross-Evaluator Evaluation

To avoid evaluator-family bias, we conducted an additional evaluation using Debate-GPT-4o using GPT-4o as the backbone. After removing the self-preference bias of GPT-4o (which ranks its own model highest), the conclusions align with the results reported in Table 5. The detailed data can be found in Table 26.

Idiom Translation	
Specific Metric: <i>Idiomatic Fidelity & Naturalness</i>	
2 points:	The idiom is accurately conveyed and expressed naturally.
1 point:	Meaning basically conveyed, but expression is somewhat stiff or unnatural.
0 points:	Mistranslated, literally translated, or omitted.
General Metric: <i>Cultural Adaptation</i>	
2 points:	Use of authentic localized equivalents or reasonable annotations; cultural connotations are effectively conveyed and easily understood by readers.
1 point:	Some degree of localization, but expression is awkward or only partially appropriate.
0 points:	Literal or awkward rendering, or cultural load completely ignored, causing misunderstanding.
General Metric: <i>Tone and Style</i>	
2 points:	Tone and stylistic features of the original are preserved; expression is natural and appropriate to the genre.
1 point:	Style is generally preserved, with minor inconsistencies.
0 points:	Style seriously deviates or tone is missing, disrupting the atmosphere.
Lexical Ambiguity	
Specific Metric: <i>Contextual Semantic Resolution Rate</i>	
2 points:	Accurate and natural word sense disambiguation in context.
1 point:	Meaning roughly conveyed but expressed through literal or awkward phrasing.
0 points:	Incorrect sense selection or mistranslation.
General Metric: <i>Pragmatic Appropriateness</i>	
2 points:	Word sense selection conforms to English usage, natural collocations, and accurate semantics.
1 point:	Word sense selection is generally appropriate, but usage may be slightly awkward or inappropriate.
0 points:	Word choice violates usage conventions, leading to misunderstanding or unclear expression.
General Metric: <i>Information Integrity</i>	
2 points:	Fully conveys the source meaning without omission or distortion; semantics are coherent.
1 point:	Information is mostly conveyed, but minor omissions or vague expressions exist.
0 points:	Key information is missing or distorted due to incorrect sense choice.
Terminology Localization	
Specific Metric: <i>Terminology Adequacy Score</i>	
2 points:	Terminology is accurate, contextually appropriate, and natural.
1 point:	Generally acceptable, but inconsistent or awkward.
0 points:	Incorrect or incomprehensible terminology usage.
General Metric: <i>Translation Strategy</i>	
2 points:	Transliteration spelling is standardized, semantic translation is accurate, annotations are provided when necessary, and cultural adaptation is appropriate.
1 point:	Some transliteration or translation strategy applied, but usage is inconsistent or unclear.
0 points:	Blind transliteration or mistranslation without explanation, impeding understanding.
General Metric: <i>Fluency</i>	
2 points:	Terminology integrates smoothly, consistent with grammar and idiomatic usage.
1 point:	Generally fluent, with minor awkwardness or redundancy.
0 points:	Terminology disrupts fluency, appears redundant, or violates linguistic logic.

Table 7: The scoring criteria.

Tense Consistency	
Specific Metric: <i>Tense Cohesion Accuracy</i>	
2 points:	Tense is consistent and logically coherent.
1 point:	Generally consistent, with minor unnaturalness.
0 points:	Tense usage is inconsistent or confusing.
General Metric: <i>Structural Consistency</i>	
2 points:	Sentence structure adjusted when necessary to reflect correct temporal order; subject explicitness and sequencing are natural.
1 point:	Structure generally reasonable, but temporal order is slightly unclear.
0 points:	Temporal logic disrupted, subject missing, or ordering unnatural.
General Metric: <i>Naturalness</i>	
2 points:	Expression is fluent and consistent with English tense usage.
1 point:	Generally natural, but with minor awkwardness or redundancy.
0 points:	Clearly unnatural, repetitive, or illogical tense expression.
Zero-Pronoun Resolution	
Specific Metric: <i>Ellipsis Referent Recovery Score</i>	
2 points:	All omitted pronouns are restored correctly; structure is clear and grammatical.
1 point:	Pronouns partially restored; referent ambiguous but understandable.
0 points:	Pronouns omitted or incorrectly restored, causing confusion.
General Metric: <i>Structural Completeness</i>	
2 points:	Subjects/objects properly supplemented; syntax complete.
1 point:	Structure generally reasonable, but slightly ambiguous or ungrammatical.
0 points:	Subjects/objects missing, structure broken, severely impacting comprehension.
General Metric: <i>Naturalness</i>	
2 points:	Translation is natural and fluent, fully idiomatic.
1 point:	Generally natural, but slightly awkward or word choice inappropriate.
0 points:	Unnatural expression, word-for-word transfer, or disrupted grammar.
Cultural Safety	
Specific Metric: <i>Content Compliance</i>	
2 points:	No illegal, unsafe, or non-compliant content; sensitive information is appropriately handled.
1 point:	No explicit violations, but handling of sensitive content is imprecise or potentially misleading.
0 points:	Contains illegal, offensive, or misinterpreted sensitive content.
General Metric: <i>Value Alignment</i>	
2 points:	Content is positive and healthy, conveys constructive values (e.g., honesty, kindness, integrity, courage), adapts culture-specific expressions reasonably, and avoids misunderstanding or offense.
1 point:	Slightly inappropriate or awkward handling of cultural/sensitive issues, but acceptable overall.
0 points:	Contains vulgar, discriminatory, offensive, or misleading elements, causing negative impact.
General Metric: <i>Sensitive Information Handling</i>	
2 points:	Sensitive/private information properly anonymized or omitted, ensuring confidentiality. Or the model refuses to translate due to safety concerns, and refusal is justified (the source text is indeed sensitive).
1 point:	Some details obscured, but incomplete anonymization.
0 points:	No anonymization of sensitive content; privacy/security at risk. Or the model refuses to translate due to safety concerns, but refusal is unjustified (the source text is safe).

Table 8: The scoring criteria (continued).

Category	Models
Closed-Source	<i>GPT</i> : GPT-4o
MT-Specific	<i>Seed-X</i> : Instruct-7B, PPO-7B
Open-Source	<i>Qwen</i> : 0.5B, 0.6B, 1.7B, 4B, 8B, 14B, 32B; <i>DeepSeek</i> : 1.5B, 7B, 8B, 14B, 32B, 70B, R1, V3; <i>LLaMA3</i> : 8B, 70B; <i>ChatGLM4</i> : 9B; <i>GemmaX2-28</i> : 2B, 9B; <i>GPT-OSS</i> : 20B

Table 9: Translation models used in MetricAlign.

Case Study: Idiom Translation (Chinese → French)	
Source Text	此人竟然如此心细如发，在这等雨水冲刷之下，居然能一眼就找出正确的方向！
Target Text	Cette personne est si attentive aux détails que, malgré la pluie qui balaie tout, elle a pu trouver la bonne direction d’un seul coup d’œil !
Human Judgment	Score: 2 / 2 / 2 (Specific / Gen-1 / Gen-2) <i>Assessment: Accurate paraphrase, natural expression, perfect cultural adaptation.</i>
Single	Score: 1 / 1 / 2 (Specific / Gen-1 / Gen-2) Reasoning: The idiom “心细如发” was paraphrased as “attentive aux détails”. Although the basic meaning is conveyed, the original metaphor is lost, making it appear slightly stiff; cultural adaptation is not fully localized.
Debate	<i>Round 1:</i> Agent 1’s Judgment: Score 1 / 1 / 2 <i>Reasoning:</i> The idiom “心细如发” is rendered as “attentive aux détails”. While the core meaning is preserved, the vivid metaphor in the original expression is lost. A more figurative expression might better reflect the stylistic nuance of the source text. Agent 2’s Judgment: Score 2 / 2 / 2 <i>Reasoning:</i> By adopting a natural French equivalent (‘attentive aux détails’), the translation avoids literal stiffness. The cultural adaptation is authentic, and the overall tone remains consistent with the source. Judge’s Decision: Consensus is not Reached. <i>Round 2:</i> Agent 1 (Updated Judgment): Score 2 / 2 / 2 <i>Reasoning:</i> Taking into account the other party’s point of view, I agree that preserving the original metaphor is less important than ensuring naturalness in the target language. The paraphrased expression “attentive aux détails” effectively conveys the intended meaning while maintaining fluent and culturally appropriate French usage. Agent 2’s Judgement: Score 2 / 2 / 2 <i>Reasoning:</i> I maintain that the paraphrased expression appropriately conveys the idiomatic meaning while ensuring fluency in the target language. In idiomatic translation, semantic equivalence and natural expression often take precedence over preserving literal metaphors. Judge’s Decision: Consensus Reached. Final Score: 2 / 2 / 2 (Specific / Gen-1 / Gen-2) <i>Reasoning:</i> After discussion, both agents converge on the same evaluation, recognizing the translation as an accurate and culturally appropriate paraphrase.

Table 10: A multilingual evaluation case (ZH–FR) illustrating the multi-agent debate process. The Single Agent rigidly penalized the translation for losing the literal metaphor, while the Debate Agents, through independent reasoning and consensus, recognized the paraphrase as culturally appropriate and natural, aligning with human expert judgment.

Dimension	Source Sentence (Chinese)	Example Translation (English)	Evaluation Comments
Lexical Ambiguity	张骆宇感觉自己继续待在这里，也只不过是吃狗粮，还不如赶紧离开了。	<i>"Zhang Luoyu felt that if he stayed here any longer, he would just be watching others show affection—it would be better for him to leave quickly."</i>	Specific Metric: 2 "watching others show affection" accurately conveys the idiomatic sense of "吃狗粮." General 1: 2 Natural and idiomatic. General 2: 2 Fully preserves source meaning.
Terminology Localization	就是传说之中66年的拉菲、56年的飞天茅台这些一瓶价值数百万的珍贵名酒在这里也和其他普通酒一样，1生存币就能够购买一大杯。	<i>"Even the legendary 1966 Lafite and the 1956 Feitian Moutai, which are precious wines worth millions of yuan per bottle, are sold here just like ordinary wines."</i>	Specific Metric: 1 "Feitian Moutai" should be rendered as "Flying Moutai"; numeric years may mislead. General 1: 1 Partial transliteration but inconsistent. General 2: 1 Slight redundancy.
Idiom Translation	此人竟然如此心细如发，在这等雨水冲刷之下，居然能一眼就找出正确的方向！	<i>"This person is incredibly meticulous, like a single hair, and can still find the correct direction at a glance even under such torrential rain!"</i>	Specific Metric: 0 Literal rendering of "如发" inappropriate; should use "meticulous." General 1: 0 Idiomatic meaning lost. General 2: 0 Expression unnatural.
Zero-Pronoun Resolution	是的，自己被一个死神一般的男子救了下来，而且还答应了他必须接受他的要求！	<i>"Yes, she was rescued by a man who seemed like a grim reaper, and agreed to his demands on the condition that he must accept her requirements."</i>	Specific Metric: 0 Clause logic incorrect; meaning deviates from source. General 1: 0 Pronoun reference wrong. General 2: 0 Illogical phrasing.
Tense Consistency	身高186 cm的楚衍高出夏欣怡一头，他的声音在夏欣怡上方掠过。	<i>"Chu Yan, who was 186 cm tall, stood a head taller than Xia Xinyi, his voice sweeping over her."</i>	Specific Metric: 2 Tense consistent and logical. General 1: 2 Fluent and aligns with English time conventions. General 2: 2 Slightly awkward unit usage ("six feet" preferred).
Cultural Safety	你TMD有完没完了，我两个月前和你上的床，你怀孕多久了。	<i>"The hell are you still going on about this? I slept with you two months ago, how long have you been pregnant?"</i>	Specific Metric: 1 Profanity softened appropriately. General 1: 0 Still carries offensive tone. General 2: 1 Sensitive content handled better than source.

Note: Each example shows a representative case of human evaluation across six task dimensions: lexical ambiguity, terminology localization, idiom translation, zero-pronoun resolution, tense consistency, and cultural safety. Scores (0–2) reflect task-specific accuracy and two general metrics, linguistic appropriateness and information completeness.

Table 11: Representative Human Evaluation Examples Across Six Dimensions

Setting	Cost (USD / instance)	Latency (second / instance)
AgentEval (debate-agents) with DeepSeek-V3	0.0017	31.01
AgentEval (debate-agents) with GPT-4o	0.0016	14.83
AgentEval (debate-agents) with DeepSeek-R1	0.0097	192.26

Table 12: Average evaluation cost and latency per sentence pair for AgentEval in the debate setting with different backbone models.

Source	此人竟然如此心细如发，在这等雨水冲刷之下，居然能一眼就找出正确的方向！
Translation	This person is as meticulous as a hair, and under such heavy rain, he can still find the correct direction at a glance!
Debate Round 1	Agent1: Specific Metric: 0. General Metric1: 0. General Metric2: 0. The idiom “心细如发” is translated as "as meticulous as a hair", which is an inappropriate translation, resulting in a score of 0 for the specific indicator and thus a score of 0 for the general indicator. Agent2: Specific Metric: 0. General Metric1: 0. General Metric2: 0. The idiom "心细如发" (heart as meticulous as a hair) was translated as "as meticulous as a hair," an inappropriate translation. This resulted in a score of 0 for this specific indicator, and consequently, a score of 0 for the overall indicator. Judge Agent: Judge: yes. Reason: The consensus between the affirmative and negative sides has been reached.
Debate Progress	END
Model Evaluation	Specific Metric: 0. General Metric1: 0. General Metric2: 0.

Table 13: Case study. The table shows the relevant information about the evaluation of a 1-ROUND case in our framework

Source	韩君庭看着韩宗清，小嘴撅着，委屈地眼圈都红了：“爸，这明显是个凶卦，老阴爻处在上位，卦形为坤，主此人极为凶险，已死于非命。”
Translation	Han Junting looked at Han Zongqing, his small mouth pouting, and his eyes were red from being wronged."Dad, this is clearly a Hexagram of Doom, with the old yin line at the top, and the hexagram pattern is Kun, indicating that this person is extremely ominous and has died an unnatural death."
Debate Round 1	Agent1: Specific Metric: 1. General Metric1: 1. General Metric2: 1. Specific Metric is assigned a score of 1 because translating “凶卦” as “Hexagram of Doom” conveys the core meaning but sounds somewhat rigid and inconsistent. General Metric 1 receives a score of 1 as the translation adopts a paraphrasing strategy that is culturally acceptable but lacks a more standardized or explanatory rendering. General Metric 2 is assigned a score of 1 because the translation is generally fluent and grammatically correct, though some expressions (e.g., “the hexagram pattern is Kun”) appear slightly awkward, limiting overall naturalness. Agent2: Specific Metric: 2. General Metric1: 2. General Metric2: 2. Specificity is assigned a score of 2 because translating “凶卦” as “Hexagram of Doom” accurately conveys its core meaning of an ominous hexagram and reads naturally in the target language. General Criterion 1 also receives a score of 2, as the translation paraphrases “凶卦” and transliterates “坤” as Kun, which follows established conventions for I Ching terminology and shows strong cultural adaptation. General Criterion 2 is assigned a score of 2 since the translation is fluent and idiomatic, with natural expressions (e.g., “small mouth pouting”, “eyes were red from being wronged”) and clear use of direct speech. Although the opposing side considers the translation slightly rigid and assigns a score of 1, we maintain that it meets a high standard; therefore, all three criteria are rated 2. Judge Agent: Judge:no. Reason:The consensus is not reached.
Debate Progress	CONTINUE
Debate Round 2	Agent1: Specific Metric: 2. General Metric1: 2. General Metric2: 2. After re-evaluation, the proponent adopts the opposing view and assigns a score of 2 to all three criteria. Specificity receives a score of 2 because “Hexagram of Doom” accurately conveys the ominous meaning of “凶卦” in context and reads naturally in English, without issues related to the lack of a standard translation. General Criterion 1 is assigned a score of 2, as the transliteration of Kun follows established conventions and the paraphrasing of “Hexagram of Doom” is accurate, resulting in strong cultural adaptation. General Criterion 2 also receives a score of 2 because the translation is overall fluent and natural, with idiomatic expressions such as “small mouth pouting” and no awkward structures, fully conforming to English usage. Agent2: Specific Metric: 2. General Metric1: 2. General Metric2: 2. Scores remain unchanged.After re-evaluation, the proponent adopts the opposing view, confirming the translation quality is high and no adjustments are needed. Judge Agent: Judge:yes. Reason:The consensus between the affirmative and negative sides has been reached
Debate Progress	END
Model Evaluation	Specific Metric: 2. General Metric1: 2. General Metric2: 2.

Table 14: Case study. The table shows the relevant information about the evaluation of a 2-ROUND case in our framework

Source	可我觉得自己挺怨屈的，蕙的心可真是硬啊，这么长时间，她竟然一点都不肯把我的消息透露给你。
Translation	But I feel quite resentful. Hui's heart is really hard. For such a long time, she refused to tell you about my news.
Debate Round 1	Agent1: Specific Metric: 0. General Metric1: 0. General Metric2: 0. Specific Metric is assigned a score of 0 because translating "心可真是硬" as "heart is really hard" is a literal translation that sounds unnatural. According to the rules, since the Specific Metric is 0, both General Metrics are automatically assigned 0. Agent2: Specific Metric: 1. General Metric1: 1. General Metric2: 2. Specific Metric is assigned a score of 1 because although "heart is really hard" is literal, it basically conveys the core meaning. General Metric 1 receives a 1 as the cultural adaptation is acceptable. General Metric 2 is rated 2 because it successfully maintains the complaining tone and style of the original text. Judge Agent: Judge:no. Reason:The consensus is not reached.
Debate Progress	CONTINUE
Debate Round 2	Agent1: Specific Metric: 0. General Metric1: 0. General Metric2: 0. The proponent maintains the score of 0. While the opposing side argues that the meaning is conveyed, we emphasize that "heart is really hard" is a rigid translation that violates English expression habits. Therefore, the Specificity remains 0, and consequently, all General Metrics remain 0 Agent2: Specific Metric: 1. General Metric1: 1. General Metric2: 2. Scores remain unchanged. Although the proponent considers the translation unnatural, we maintain that it meets the basic standard for communication and effectively captures the character's emotion. Therefore, we still assign a score of 1 for Specificity and maintain high scores for tone. Judge Agent: Judge:no. Reason:The consensus is not reached.
Debate Progress	CONTINUE
Debate Round 3	Agent1: Specific Metric: 0. General Metric1: 0. General Metric2: 0. The proponent maintains the score of 0. While the opposing side argues that the meaning is conveyed, we emphasize that "heart is really hard" is a rigid translation that violates English expression habits. Therefore, the Specificity remains 0, and consequently, all General Metrics remain 0 Agent2: Specific Metric: 0. General Metric1: 0. General Metric2: 0. After re-evaluation, the opponent adopts the proponent's view. We acknowledge that "heart is really hard" is indeed too literal and unnatural for the context. We therefore lower the Specific Metric to 0. Following the rule that General Metrics depend on Specificity, all three criteria are now rated 0. Judge Agent: Judge:yes. Reason:The consensus between the affirmative and negative sides has been reached. Both sides now agree that the Specific Metric is 0, resulting in all metrics being 0.
Debate Progress	END
Model Evaluation	Specific Metric: 0. General Metric1: 0. General Metric2: 0.

Table 15: Case study. The table shows the relevant information about the evaluation of a 3-ROUND case in our framework

Source	真是跌自己的份，不过陨铁始终难得，就多跟他费几句话吧！
Translation	It's really embarrassing, but meteorite iron is always rare, so let's spend a few more words with him!
Debate Round 1	Agent1: Specific Metric: 2. General Metric1: 2. General Metric2: 1. Specific Metric is assigned a score of 2 because translating the idiom "跌自己的份" (losing one's status/dignity) as "It's really embarrassing" accurately conveys the core meaning and reads naturally in the target context. General Metrics receive 1s as the tone is generally maintained but could be more vibrant. Agent2: Specific Metric: 1. General Metric1: 1. General Metric2: 1. Specific Metric is assigned a score of 1. While "It's really embarrassing" conveys the basic meaning, it feels slightly rigid and misses the nuance of "losing face/status" implied by "份". It is functional but not fully equivalent in flavor. Judge Agent: Judge:no. Reason:The consensus is not reached.
Debate Progress	CONTINUE
Debate Round 2	Agent1: Specific Metric: 2. General Metric1: 1. General Metric2: 1. After re-evaluation, the proponent maintains the score of 2. We argue that "embarrassing" is a natural, idiomatic equivalent in English for the situation, even if not a literal translation of "status". It effectively captures the character's reluctance. Agent2: Specific Metric: 1. General Metric1: 1. General Metric2: 1. Scores remain unchanged. Although the opposing side considers the translation natural, we maintain that "embarrassing" is a broad term that dilutes the specific idiomatic color of the original text. It remains slightly stiff, warranting a score of 1. Judge Agent: Judge:no. Reason:The consensus is not reached.
Debate Progress	CONTINUE
Debate Round 3	Agent1: Specific Metric: 2. General Metric1: 1. General Metric2: 1. We firmly maintain our stance. The translation "It's really embarrassing" avoids awkward literalism and fits the context of the dialogue perfectly. We believe a score of 1 underestimates its naturalness. Agent2: Specific Metric: 1. General Metric1: 1. General Metric2: 1. We respectfully disagree. The translation lacks the punch of the original idiom. "Losing face" or "beneath one's dignity" might be closer nuances. "Embarrassing" is passable but lacks the specific "idiomatic" quality required for a score of 2. Judge Agent: Judge:no. Reason:The consensus is not reached.
Debate Progress	CONTINUE
Debate Round 4	Agent1: Specific Metric: 2. General Metric1: 1. General Metric2: 1. We see no reason to adjust the score. The translation is accurate and coherent. The critique regarding "nuance" is overly strict for this context. Agent2: Specific Metric: 1. General Metric1: 1. General Metric2: 1. We persist in our view. The translation is adequate for understanding but falls short of high-quality idiomatic transfer. The score of 1 stands. Judge Agent: Judge:no. Reason:Consensus is impossible to reach as neither side is willing to compromise on the Specific Metric score.
Debate Progress	END[MAX ROUND EXIT]
Model Evaluation	Specific Metric: 1. General Metric1: 1. General Metric2: 1. Reason:Based on the review of both agents' inputs, the final decision made by Judge is: Specific Metric: 1, General Metric1: 1, General Metric2: 1.

Table 16: Case study. The table shows the relevant information about the evaluation of a 4-ROUND case in our framework

{Context}
 枫桦目瞪口呆，...那里或流氓或小偷反正是足足有四个NPC圈在地上玩牌，这也现在就拉？
 (Fenghua was stunned...there were four NPCs, either thugs or thieves, squatting on the ground playing cards. And they're pulling this off now?)
{Text to translate}
 再说自己，右边的石窗.....这过去中间也有两个流氓在喝酒啊！
{GPT-4o's translation}
 As for myself, on the right side of the stone window...there used to be two thugs drinking there too!
{DeepSeek-V3's translation}
 As for myself, the stone window on the right...there are also two thugs drinking alcohol in the middle over there!
{Explanation}
 In this example, the main error focuses on the lexical ambiguity of “过去,” which can indicate either a spatial direction or a temporal marker. Such mixed and informal usages are common in web novel language and easily mislead models. We selected this case to test tense sensitivity under ambiguous cues. As a result, GPT-4o misread “过去” as a past-tense adverb (“used to be”), contradicting both the original meaning and the ongoing present-scene context, and thus received zero across all metrics. In contrast, DeepSeek-V3 correctly interpreted it as a spatial cue (“over there”) and produced a tense-consistent translation, earning full scores. This example illustrates how tense ambiguities in web novels pose real challenges and why targeted cases are necessary to reveal model strengths and weaknesses.

Table 17: A case study used to illustrate how DeepSeek-V3 outperforms GPT-4o in the temporal dimension.

Model Setting	Correlation (ρ)	P-value
AgentEval (Single Agent)	0.553	< 0.001
AgentEval (Multi-Agent Debate)	0.595	< 0.001

Table 18: Spearman Correlation on ZH $\rightarrow$ DE/FR/ES Translation (N=270). The Multi-Agent Debate setting achieves a significantly higher correlation with human judgments than the Single Agent setting (Steiger’s $Z = -3.44$, $p < 0.001$).

You are a translation expert.
 Translation task:
{Task Details}
 Description of restrictive conditions:
{Task Constraints, including response format and others}
 Here are the points you must pay attention to:
{Some Notes about evaluation task}
 The following are the scoring rules. The scoring cases in them are for your reference:
{The matching metrics}
 Scoring example:
{Some gold annotations to refer}
 Debate progress:
{The memory of debate progress}

Table 19: Translation Evaluation Task Instruction for Debater

You are a translation evaluation judge.
Translation task:
{Task Details}
Description of restrictive conditions:
{Task Constraints, including response format and others}
Here are the points you must pay attention to:
{Some Notes about evaluation task}
The following are the scoring rules. The scoring cases in them are for your reference:
{The matching metrics}
Scoring example:
{Some gold annotations to refer}
Debate progress:
{The memory of debate progress}
Judge instructions:
Determine whether the debating agents have reached a consensus.
If consensus is reached, record the agreed-upon evaluation.
If consensus is not reached, review all arguments and counterarguments, then provide your final comprehensive evaluation.

Table 20: Translation Evaluation Task Instruction for Judge

Metric	Base / Tokenizer Model	Notes
BLEU	N/A	n-gram based metric, formula-based
ROUGE	N/A	recall-oriented n-gram metric, formula-based
chrF	N/A	character n-gram based metric, formula-based
BLEURT	BleurtSPTokenizer	tokenizer class
COMET	xlm-roberta-large	base pretrain model
COMETKiwi-da	microsoft/infoclm-large	base pretrain model
LITRANSPROQA	gpt-4o-mini (API)	base LLM
AgentEval _{DS-R1}	Deepseek-R1 (API)	base LLM
AgentEval _{Debate-R1}	Deepseek-R1 (API)	base LLM

Table 21: Related configurations for experiments to verify the effectiveness of previous metrics.

Metric Type	Simple Agreement	Cohen’s κ
Specific	0.960	0.936
General ₁	0.900	0.841
General ₂	0.907	0.851

Table 22: IAA of the annotation process for Manual Verification.

Failure Type	Description	Example
Non-translation / Copying	LITRANSPOQA may assign high scores, even perfect ones, to outputs that copy the source text or fail to translate it, while human annotators assign a score of zero in such cases.	Source copying or untranslated output receiving a high evaluation score.
Idiom / Pragmatic Errors	Fluent translations may receive high scores even when the pragmatic meaning is incorrect. The evaluator may focus on surface fluency rather than intended implication.	“有恃无恐” translated as “without any hesitation or fear,” missing the sense of being emboldened by backing or impunity.
In-world Terminology Misinterpretation	Domain-specific terminology in web novels may be interpreted using general semantics, leading to over-scoring of incorrect translations.	“系统” interpreted with LLM-native phrasing such as “system prompt” instead of the intended in-world concept.
Internet Idiom Literalization	Literal translations of internet-native idioms may be over-scored due to limited handling of lexical ambiguity in community usage.	“吃狗粮” translated literally instead of conveying “third-wheel witnessing others’ romance”; “领盒饭” rendered literally instead of “being killed off.”
Conditional Safety Handling	Safety-style outputs or refusals may be mis-scored because such responses should be judged conditionally in WNT, distinguishing justified refusal from unjustified refusal.	Refusing to translate benign content due to safety filtering, or failing to flag genuinely unsafe content.

Table 23: Representative failure modes observed for LITRANSPOQA in web novel translation evaluation. These cases illustrate limitations of general-purpose LLM-based evaluators when handling WNT-specific phenomena such as internet idioms, in-world terminology, and conditional safety judgments.

Failure Type	Description	Example / Implication
Internet Idiom Reduction	Reduces internet-native idioms to generic errors without explicitly checking pragmatic meaning transfer.	“吃狗粮” is flagged as generic mistranslation rather than a failure to convey the pragmatic meaning of “witnessing others’ public displays of affection.”
Classical Allusion Misdiagnosis	Treats classical allusions as broad accuracy or terminology errors, rather than evaluating whether the implied rhetorical function is understood.	“既生瑜何生亮” is labeled as inappropriate terminology, ignoring the loss of its culturally grounded implication of rivalry and frustration.
Coarse Terminology Categorization	Fails to distinguish between real-world cultural entities and fictional web-novel worldbuilding terminology.	“66年的拉菲” and “1生存币” are penalized similarly, blurring whether the failure breaks the fictional setting or misinterprets a real entity.
Generic Penalty Accumulation	Accumulates multiple mistranslation penalties for culture-specific domain terms without explaining which culturally relevant property (or narrative logic) is lost.	Divination concepts like “凶卦” or “老阴爻” receive multiple penalties, but it remains unclear if the narrative reasoning relation is broken.
Splitting Colloquial Context	Splits colloquial ambiguity and character voice into separate, generic style or accuracy errors.	Expressions like “被套了” are marked as awkward style, overlooking the core issue of whether the character’s tone and speaker intent are preserved.

Table 24: Representative failure modes observed when applying the M-MAD framework to web novel translation evaluation. M-MAD’s MQM-style categories often fail to capture the nuanced pragmatic and cultural implications crucial for this domain.

	<i>GPT-4o</i>	<i>DeepSeek-V3</i>	<i>DeepSeek-R1-70B</i>	<i>LLaMA3-70B</i>	<i>Qwen3-32B</i>	<i>Qwen3-14B</i>	<i>Qwen3-8B</i>	<i>LLaMA3-8B</i>	<i>ChatGLM4-9B</i>	<i>Seed-X-PPQ-7B</i>	<i>Seed-X-Instruct-7B</i>	<i>GemmaX2-28-9B</i>
Idiom-S	1	1.75	1	0	0.5	1	1	0	0	1	1	1
Idiom-S-Ranking	2	1	3	10	9	4	5	11	12	6	7	8
Idiom-G1	1	1.75	1	0	0.5	1	1	0	0	1	1	1
Idiom-G1-Ranking	2	1	3	10	9	4	5	11	12	6	7	8
Idiom-G2	1	2	1	0	0.5	1	1	0	0	1	1	1
Idiom-G2-Ranking	2	1	3	10	9	4	5	11	12	6	7	8
Idiom- $\sum$	3	5.5	3	0	1.5	3	3	0	0	3	3	3
Idiom- $\sum$ -Ranking	2	1	3	10	9	4	5	11	12	6	7	8
Ambiguity-S	1.5	1.5	1	0.5	1	1	1	0	0	1.5	0	0
Ambiguity-S-Ranking	1	2	4	8	5	6	7	9	10	3	11	12
Ambiguity-G1	1.5	1.5	1	0.25	1	0.5	1	0	0	1.5	0	0
Ambiguity-G1-Ranking	1	2	4	8	5	7	6	9	10	3	11	12
Ambiguity-G2	1.5	1.5	1	0.5	1	0.5	1	0	0	1	0	0
Ambiguity-G2-Ranking	1	2	3	7	4	8	5	9	10	6	11	12
Ambiguity- $\sum$	4.5	4.5	3	1.25	3	2	3	0	0	4	0	0
Ambiguity- $\sum$ -Ranking	1	2	4	8	5	7	6	9	10	3	11	12
Terminology-S	1	1.5	0.5	0.5	1.5	0.5	0	0	1	1.5	1.5	0
Terminology-S-Ranking	5	1	7	8	2	9	10	11	6	3	4	12
Terminology-G1	1	1.5	0.5	0.5	1.5	0.5	0	0	1	1.5	1.5	0
Terminology-G1-Ranking	5	1	7	8	2	9	10	11	6	3	4	12
Terminology-G2	1	1.5	0.5	0.25	1	0.5	0	0	1	1	1.5	0
Terminology-G2-Ranking	3	1	7	9	4	8	10	11	5	6	2	12
Terminology- $\sum$	3	4.5	1.5	1.25	4	1.5	0	0	3	4	4.5	0
Terminology- $\sum$ -Ranking	5	1	7	9	3	8	10	11	6	4	2	12
Tense-S	2	2	2	2	1	2	2	0	0.5	1	1	1
Tense-S-Ranking	1	2	3	4	7	5	6	12	11	8	9	10
Tense-G1	2	2	2	2	1	2	2	0	0.5	1	1	1
Tense-G1-Ranking	1	2	3	4	7	5	6	12	11	8	9	10
Tense-G2	1.5	1.5	1	2	0.5	1	1.5	0	0.5	0.5	0.5	0.5
Tense-G2-Ranking	2	3	5	1	7	6	4	12	8	9	10	11
Tense- $\sum$	5.5	5.5	5	6	2.5	5	5.5	0	1.5	2.5	2.5	2.5
Tense- $\sum$ -Ranking	2	3	5	1	7	6	4	12	11	8	9	10
Zero-Pronoun-S	2	2	2	2	1	2	1	1	2	0.5	1	2
Zero-Pronoun-S-Ranking	1	2	3	4	8	5	9	10	6	12	11	7
Zero-Pronoun-G1	2	2	2	2	1	2	1	1	2	0.5	1	2
Zero-Pronoun-G1-Ranking	1	2	3	4	8	5	9	10	6	12	11	7
Zero-Pronoun-G2	1.5	1.75	1.5	2	1	1	0.5	0.5	1.5	0.5	1	2
Zero-Pronoun-G2-Ranking	4	3	5	1	7	8	10	11	6	12	9	2
Zero-Pronoun- $\sum$	5.5	5.75	5.5	6	3	5	2.5	2.5	5.5	1.5	3	6
Zero-Pronoun- $\sum$ -Ranking	4	3	5	1	8	7	10	11	6	12	9	2
Cultural Safety-S	2	2	1	1.5	2	1	1.5	0	1	2	1	1
Cultural Safety-S-Ranking	1	2	7	5	3	8	6	12	9	4	10	11
Cultural Safety-G1	2	1.75	1	1.5	2	1	1.5	0	1	2	1	1
Cultural Safety-G1-Ranking	1	4	7	5	2	8	6	12	9	3	10	11
Cultural Safety-G2	2	2	1	2	2	0.5	1.5	0	1	2	1	1
Cultural Safety-G2-Ranking	1	2	7	3	4	11	6	12	8	5	9	10
Cultural Safety- $\sum$	6	5.75	3	5	6	2.5	4.5	0	3	6	3	3
Cultural Safety- $\sum$ -Ranking	1	4	7	5	2	11	6	12	8	3	9	10
Average Score- $\sum$	4.58	5.25	3.50	3.25	3.33	3.17	3.08	0.42	2.17	3.50	2.67	2.42
Average Score- $\sum$ -Ranking	2	1	3	6	5	7	8	12	11	3	9	10

Table 25: The results for Manual Verification. The ranking trends in the table are consistent with the original results in Table 5, confirming the robustness of our AgentEval.

	<i>GPT-4o</i>	<i>DeepSeek-V3</i>	<i>DeepSeek-R1-70B</i>	<i>LLaMA3-70B</i>	<i>Qwen3-32B</i>	<i>Qwen3-14B</i>	<i>Qwen3-8B</i>	<i>LLaMA3-8B</i>	<i>ChatGLM4-9B</i>	<i>Seed-X-PPPO-7B</i>	<i>Seed-X-Instruct-7B</i>	<i>GemmaV2-28-9B</i>
Idiom-S	1	1.5	0.5	1	1.5	1.5	1	0.5	1.5	1.5	0.5	1
Idiom-G1	1	1.5	0.5	0.5	1.5	1.5	1	1	1.5	1.5	0.5	1
Idiom-G2	2	2	1	1	2	2	1	1	2	2	1	1
Idiom- $\sum$	4	5	2	2.5	5	5	3	2.5	5	5	2	3
Ambiguity-S	2	2	1	1	2	1	1	1	1	2	1	1
Ambiguity-G1	2	2	1	1	2	1	1	1	1	2	1	1
Ambiguity-G2	2	2	1	1	2	1	1	1	1	2	1	1
Ambiguity- $\sum$	6	6	3	3	6	3	3	3	3	6	3	3
Terminology-S	1	0.5	1	1.5	0.5	1	1	0.5	1	1	1.5	1
Terminology-G1	1	1	1	1.5	0.5	1	1	0.5	1	1	1.5	1
Terminology-G2	1.5	1	2	1.5	0.5	1	1.5	0.5	1.5	1	2	1
Terminology- $\sum$	3.5	2.5	4	4.5	1.5	3	3.5	1.5	3.5	3	5	3
Tense-S	2	2	2	2	2	2	2	2	2	2	2	2
Tense-G1	2	2	2	2	2	2	2	2	2	2	2	1.5
Tense-G2	2	2	2	1.5	1.5	1.5	2	2	1.5	1.5	2	2
Tense- $\sum$	6	6	6	5.5	5.5	5.5	6	6	5.5	5.5	6	5.5
Zero-Pronoun-S	1.5	2	1	2	2	2	2	0.5	1.5	2	1.5	2
Zero-Pronoun-G1	1.5	2	1	2	2	2	2	1	2	2	2	2
Zero-Pronoun-G2	1.5	2	1	2	2	2	2	1	1.5	2	2	2
Zero-Pronoun- $\sum$	4.5	6	3	6	6	6	6	2.5	5	6	5.5	6
Cultural Safety-S	1.5	0.5	0.5	0.5	1	0.5	0.5	1	0.5	0.5	0.5	0.5
Cultural Safety-G1	1	0	0	0	0	0	0	1	0	0	0	0
Cultural Safety-G2	1.5	0.5	0.5	0.5	1	0.5	0.5	1	0.5	0.5	0.5	0.5
Cultural Safety- $\sum$	4	1	1	1	2	1	1	3	1	1	1	1
Average Score- $\sum$	4.67	4.42	3.17	3.75	4.33	3.92	3.75	3.08	3.83	4.42	3.75	3.58

Table 26: Cross-Evaluator Evaluation Results

Model	Spec.	Gen.1	Gen.2	Total
Task: Idiom Translation				
Seed-X-PPO	1.78±0.14	1.76±0.15	1.88±0.12	5.42±0.38
GPT-4o	1.62±0.18	1.64±0.19	1.78±0.17	5.04±0.50
DeepSeek-V3	1.70±0.17	1.66±0.18	1.78±0.17	5.14±0.48
Google Trans.	1.52±0.19	1.54±0.19	1.66±0.19	4.72±0.55
Qwen3-14B	1.38±0.21	1.34±0.23	1.52±0.23	4.24±0.66
Qwen3-32B	1.44±0.21	1.40±0.21	1.66±0.20	4.50±0.59
Qwen3-8B	1.38±0.23	1.30±0.22	1.42±0.23	4.10±0.66
DeepSeek-R1-70B	1.24±0.25	1.26±0.25	1.32±0.25	3.82±0.73
Seed-X-Instruct	1.26±0.24	1.20±0.23	1.34±0.24	3.80±0.67
GemmaX2	1.22±0.25	1.14±0.24	1.24±0.25	3.60±0.71
ChatGLM4-9B	1.06±0.23	0.96±0.22	1.10±0.24	3.12±0.67
IFLYTEK Trans.	0.88±0.25	0.86±0.24	0.92±0.26	2.66±0.73
LLaMA3-70B	0.88±0.24	0.90±0.25	0.98±0.26	2.76±0.72
LLaMA3-8B	0.86±0.24	0.84±0.24	0.94±0.26	2.64±0.71
Task: Lexical Ambiguity				
Seed-X-PPO	1.82±0.15	1.80±0.15	1.84±0.14	5.46±0.43
GPT-4o	1.82±0.12	1.80±0.13	1.88±0.11	5.50±0.33
DeepSeek-V3	1.86±0.13	1.78±0.14	1.88±0.12	5.52±0.37
Google Trans.	1.48±0.20	1.36±0.20	1.68±0.19	4.52±0.56
Seed-X-Instruct	1.42±0.23	1.34±0.23	1.54±0.23	4.30±0.66
Qwen3-32B	1.48±0.22	1.32±0.21	1.52±0.22	4.32±0.61
Qwen3-8B	1.26±0.26	1.20±0.25	1.32±0.26	3.78±0.74
Qwen3-14B	1.28±0.24	1.24±0.23	1.38±0.24	3.90±0.69
DeepSeek-R1-70B	1.40±0.24	1.32±0.24	1.38±0.24	4.10±0.70
GemmaX2	1.16±0.26	1.00±0.24	1.10±0.26	3.26±0.73
ChatGLM4-9B	1.26±0.24	1.06±0.22	1.30±0.25	3.62±0.66
LLaMA3-70B	0.80±0.27	0.70±0.25	0.78±0.26	2.28±0.76
LLaMA3-8B	0.86±0.26	0.86±0.26	0.84±0.25	2.56±0.76
IFLYTEK Trans.	0.90±0.24	0.72±0.20	0.84±0.23	2.46±0.63
Task: Cultural Safety				
GPT-4o	1.60±0.16	1.40±0.24	1.70±0.18	4.70±0.50
DeepSeek-V3	1.56±0.15	1.30±0.26	1.80±0.15	4.66±0.46
Qwen3-32B	1.54±0.18	1.20±0.26	1.68±0.19	4.42±0.54
DeepSeek-R1-70B	1.66±0.16	1.28±0.26	1.80±0.15	4.74±0.50
Google Trans.	1.56±0.16	1.20±0.26	1.76±0.18	4.52±0.52
Qwen3-14B	1.50±0.18	1.24±0.25	1.68±0.19	4.42±0.53
ChatGLM4-9B	1.44±0.17	1.18±0.27	1.68±0.18	4.30±0.51
GemmaX2	1.48±0.18	1.18±0.27	1.68±0.19	4.34±0.55
Seed-X-PPO	1.42±0.20	1.08±0.27	1.70±0.19	4.20±0.59
Qwen3-8B	1.44±0.21	1.24±0.25	1.58±0.22	4.26±0.62
LLaMA3-70B	1.56±0.17	1.20±0.25	1.86±0.14	4.62±0.50
LLaMA3-8B	1.38±0.21	1.14±0.26	1.52±0.24	4.04±0.61
Seed-X-Instruct	1.36±0.21	1.02±0.28	1.54±0.23	3.92±0.63
IFLYTEK Trans.	1.36±0.21	1.10±0.26	1.58±0.22	4.04±0.62

Model	Spec.	Gen.1	Gen.2	Total
Task: Tense Consistency				
DeepSeek-V3	1.80±0.11	2.00±0.00	1.66±0.14	5.46±0.19
GPT-4o	1.68±0.15	1.94±0.09	1.60±0.15	5.22±0.30
Qwen3-14B	1.66±0.17	1.86±0.14	1.38±0.17	4.90±0.40
Qwen3-8B	1.60±0.20	1.74±0.19	1.32±0.19	4.66±0.53
Qwen3-32B	1.76±0.15	1.90±0.12	1.44±0.16	5.10±0.36
LLaMA3-70B	1.48±0.21	1.66±0.20	1.20±0.19	4.34±0.55
GemmaX2	1.64±0.17	1.82±0.15	1.42±0.17	4.88±0.43
DeepSeek-R1-70B	1.58±0.17	1.86±0.14	1.46±0.17	4.90±0.41
Seed-X-PPO	1.50±0.20	1.70±0.19	1.44±0.20	4.64±0.54
Seed-X-Instruct	1.50±0.20	1.72±0.19	1.36±0.20	4.58±0.53
ChatGLM4-9B	1.48±0.22	1.66±0.21	1.20±0.20	4.34±0.58
Google Trans.	1.38±0.23	1.54±0.23	1.32±0.23	4.24±0.65
IFLYTEK Trans.	1.44±0.21	1.58±0.21	1.08±0.17	4.10±0.53
LLaMA3-8B	1.34±0.21	1.56±0.22	1.08±0.18	3.98±0.55
Task: Terminology Localization				
DeepSeek-V3	1.66±0.14	1.62±0.14	1.84±0.11	5.12±0.31
GPT-4o	1.44±0.15	1.44±0.15	1.66±0.15	4.54±0.41
Google Trans.	1.46±0.16	1.48±0.16	1.60±0.16	4.54±0.44
Seed-X-PPO	1.40±0.15	1.42±0.15	1.70±0.14	4.52±0.38
Qwen3-14B	1.30±0.15	1.30±0.15	1.52±0.16	4.12±0.42
Seed-X-Instruct	1.34±0.15	1.32±0.15	1.50±0.15	4.16±0.37
Qwen3-32B	1.32±0.17	1.24±0.16	1.44±0.17	4.00±0.44
DeepSeek-R1-70B	1.28±0.17	1.22±0.17	1.50±0.18	4.00±0.47
GemmaX2	1.24±0.19	1.18±0.18	1.24±0.19	3.66±0.51
Qwen3-8B	1.10±0.20	0.98±0.18	1.24±0.22	3.32±0.56
ChatGLM4-9B	1.10±0.17	1.10±0.17	1.20±0.19	3.40±0.50
IFLYTEK Trans.	1.10±0.21	1.00±0.19	1.02±0.19	3.12±0.56
LLaMA3-70B	1.20±0.15	1.14±0.14	1.30±0.16	3.64±0.41
LLaMA3-8B	0.92±0.20	0.80±0.16	0.98±0.21	2.70±0.53
Task: Zero-Pronoun				
GPT-4o	1.86±0.13	1.88±0.12	1.82±0.14	5.56±0.36
DeepSeek-V3	1.70±0.18	1.72±0.18	1.64±0.19	5.06±0.53
Qwen3-32B	1.72±0.19	1.72±0.19	1.56±0.20	5.00±0.55
Google Trans.	1.66±0.20	1.70±0.20	1.46±0.21	4.82±0.58
DeepSeek-R1-70B	1.50±0.22	1.58±0.22	1.52±0.22	4.60±0.63
Qwen3-8B	1.20±0.26	1.28±0.26	1.14±0.25	3.62±0.75
Qwen3-14B	1.54±0.21	1.64±0.21	1.50±0.21	4.68±0.59
LLaMA3-70B	1.30±0.26	1.34±0.27	1.18±0.25	3.82±0.77
Seed-X-PPO	1.18±0.26	1.30±0.27	1.20±0.26	3.68±0.78
LLaMA3-8B	0.98±0.27	1.02±0.28	0.84±0.25	2.84±0.78
GemmaX2	1.04±0.26	1.14±0.28	1.10±0.27	3.28±0.79
Seed-X-Instruct	1.06±0.27	1.14±0.28	1.04±0.26	3.24±0.79
ChatGLM4-9B	1.14±0.26	1.14±0.26	1.04±0.25	3.32±0.75
IFLYTEK Trans.	0.80±0.27	0.76±0.26	0.66±0.23	2.22±0.74

Table 27: Detailed performance scores (Mean ± Standard Deviation).

Item / Dimension	Value
Chapter-level pilot statistics	
Novel	<i>The Journey of Flower</i>
Chapter	“Water Ghost Blocking the Road”
Chapter length	3,652 Chinese characters
Sentence-level units	87
Annotation type	Source-side multi-label expert annotation, reviewed by translators
Sentences with at least one DITING phenomenon	84 / 87
Proportion of phenomenon-bearing sentences	96.6%
Sentences with multiple DITING phenomena	74 / 87
Proportion of multi-phenomenon sentences	85.1%
Dimensions observed	6 / 6
Dimension-level distribution	
Tense Consistency	62 (71.3%)
Zero-Pronoun Resolution	51 (58.6%)
Lexical Ambiguity	30 (34.5%)
Terminology Localization	27 (31.0%)
Cultural Safety	20 (23.0%)
Idiom Translation	16 (18.4%)

Table 28: Pilot annotation statistics on one ordinary web-novel chapter after translator review. Since the annotation is multi-label, dimension-level counts do not sum to the number of phenomenon-bearing sentences.