CULTURE-CMT: BENCHMARKING CONTEXTUALIZED SOCIAL MEDIA TRANSLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Social-media comments are interactions, yet they are often translated and evaluated as isolated text. This mismatch is especially problematic when meaning depends on surrounding discussion, community-specific slang, or culturally situated expressions. We introduce **CULTURE-CMT**, a Chinese–English benchmark for contextualized social-media comment translation that restores the post and conversational context naturally available to users. Its 572 manually verified examples are organized into **Slang**, **Context**, and **Both**, enabling diagnostic analysis of distinct and interacting sources of translation difficulty. To support scalable evaluation, we further develop a retrieval-augmented cultural-effectiveness Judger that combines a calibrated scoring model with dynamically retrieved cultural knowledge. CULTURE-CMT reveals several non-trivial patterns. Providing interaction context improves most model configurations, but does not eliminate culturally ineffective translations. Examples requiring both contextual grounding and culturally situated language remain substantially more difficult than either source of difficulty alone. Moreover, larger models are not consistently better, and thinking produces highly model-dependent gains, including occasional degradation. In contrast, a 27B model adapted on benchmark-disjoint social-media data achieves the lowest observed invalidity rate among the evaluated configurations under thinking inference with context. Together, these results suggest that realistic social-media translation is not simply a matter of scaling models or reasoning longer, but of learning to translate meaning as it is constructed within evolving online interactions.



Figure 1: Interaction context can be necessary to resolve ambiguous social-media expressions and recover the intended translation. The expression "Ziwei" is ambiguous in isolation but is correctly resolved as Alexander Zverev when the original post context is available.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Social media platforms are a critical deployment scenario for cross-lingual communication. Unlike formal genres such as news and encyclopedic writing, user-generated content (UGC) is rich in slang, community-specific expressions, and rapidly evolving non-standard forms (Baldwin et al., 2013; Michel & Neubig, 2018; Sun et al., 2024). Recent benchmarks have begun to study this setting systematically. CULTURE-MT (Wu et al., 2026) focuses on culturally grounded expressions and adaptation, while CSM-MTBench (Zhao et al., 2026) distinguishes longer, information-rich posts from short, reactive snippets. Related work further shows that slang, neologisms, tone, and other emerging online expressions remain challenging for strong LLMs (Mei et al., 2024; Huang et al., 2025; Liang et al., 2025). Beyond these linguistic challenges, however, **comments are interactions rather than isolated utterances**: they respond to posts, add information, express stances, and often leave shared information implicit. Interactional context has been shown to help recover implicit emotion and sociopragmatic meaning (Lee & Lau, 2020; Zhang et al., 2023), while context-aware MT benchmarks diagnose discourse-level phenomena (Jwalapuram et al., 2020; Fernandes et al., 2023). Our focus is complementary: we preserve the native post–comment and parent–reply structure of social-media interactions and study it jointly with culturally situated language.

This interaction context can directly change how a comment should be understood and translated. As shown in Figure 1, consider the comment 除了紫薇其他人我都看，紫薇斗志全无 ("I can watch everyone except Ziwei. Ziwei has no fighting spirit."). Without the surrounding discussion, "紫薇" ("Ziwei") is ambiguous and may naturally be associated with the well-known character from the Chinese television drama *My Fair Princess*. Given the surrounding discussion of the men's round of 16 at the Rome Masters, however, it is correctly resolved as a community-specific nickname for Alexander Zverev. For platform users, such context is part of the information naturally available when reading the comment, rather than an additional hint. A small qualitative pilot over 15 randomly sampled comments further suggests that restoring post and conversational context can resolve some translation errors, but does not eliminate culturally ineffective outputs. Some failures instead involve slang or community-specific expressions, while context can occasionally lead models toward an incorrect interpretation. This raises a broader question: **once the original interaction context is restored, what still limits social-media translation?**

To study this question, we introduce **CULTURE-CMT**, a benchmark for *contextualized machine translation* that evaluates the cultural effectiveness of Chinese–English social-media comment translation. Each instance preserves the target comment together with its available post title, post content, and parent comment where applicable, covering both top-level comments and replies. Motivated by our preliminary observations and prior work on slang and emerging internet language (Wuraola et al., 2024; Sun et al., 2024; Huang et al., 2025), we categorize examples into **Slang**, **Context**, and **Both** to analyze slang-related difficulty, contextual dependence, and their combination. We then use CULTURE-CMT as a stress test for three common routes to improving LLM-based translation: **Scale**, **Reasoning**, and **Adaptation**. This framing is motivated by recent evidence that specialized translation models can rival much larger general-purpose systems (Guo et al., 2025; Cheng et al., 2025; Zheng et al., 2026), while explicit reasoning can benefit challenging interpretative language tasks (Wang et al., 2023; Liang et al., 2025).

Our results reveal a non-trivial pattern. Larger models are not consistently better, and thinking provides highly model-dependent gains. A 27B model adapted on benchmark-disjoint social-media data achieves the lowest observed invalidity rate among the evaluated configurations under thinking inference with context. Beyond aggregate performance, the difficulty breakdown shows that **Both** remains harder than **Slang** or **Context** alone among the analyzed models, highlighting the challenge of combining contextual grounding with culturally situated language. Together, these results suggest that progress in realistic social-media translation requires more than simply increasing model scale or reasoning compute.

Our contributions are threefold:

- We introduce **CULTURE-CMT**, a Chinese–English social-media comment translation benchmark that preserves the interaction context naturally available to users.

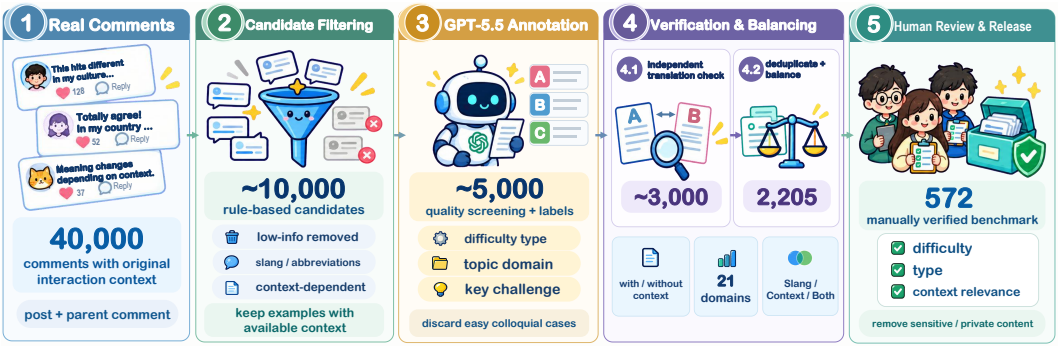

Figure 2: Construction pipeline of CULTURE-CMT, from 40K contextualized social-media comments to 572 manually verified benchmark examples.

- We provide a diagnostic evaluation framework that separates **Slang**, **Context**, and **Both** difficulties, together with a retrieval-augmented cultural-effectiveness Judger validated against human annotations.
- We compare **Scale**, **Reasoning**, and **Adaptation**, documenting non-monotonic scaling, model-dependent reasoning gains, and clear benefits from targeted adaptation of the same backbone.

2 Benchmark Construction

2.1 Task Definition

CULTURE-CMT evaluates Chinese–English social-media comment translation under the interaction context naturally available to users. Each instance contains a target comment c and its available context, including the post title t, post content p, and, for replies, the parent comment r. For a top-level comment, $\mathcal{C} = \{t, p\}$; for a reply, $\mathcal{C} = \{t, p, r\}$.

The task is to generate an English translation

$$y = f(c, \mathcal{C})$$

that preserves the intended meaning of c in its original interaction context.

For diagnostic analysis, we categorize examples into three difficulty types: **Slang**, involving internet slang, abbreviations, community-specific expressions, or other culturally situated usages; **Context**, where interpretation depends on the surrounding interaction; and **Both**, where both sources of difficulty are present.

2.2 Construction Pipeline

CULTURE-CMT is constructed from naturally occurring social-media comments through a multi-stage pipeline (Figure 2). All source data were collected between April 12 and May 12, 2026. Starting from 40,000 comments with their original interaction context, the pipeline consists of candidate filtering, model-assisted annotation, difficulty verification, deduplication and balancing, and human review. The resulting benchmark contains 572 manually verified examples.

Candidate Filtering. We first apply rule-based quality filters to malformed, overly short or long, template-like, and low-information comments. We then identify candidates containing either social-media-specific language, such as slang, abbreviations, and community-specific expressions, or potential contextual dependence. Candidates are required to have their original post context available. This stage yields approximately 10,000 examples.

Model-assisted Annotation. We use GPT-5.5 to perform structured annotation over the candidate pool, including quality screening, difficulty type, topical domain, and the primary translation

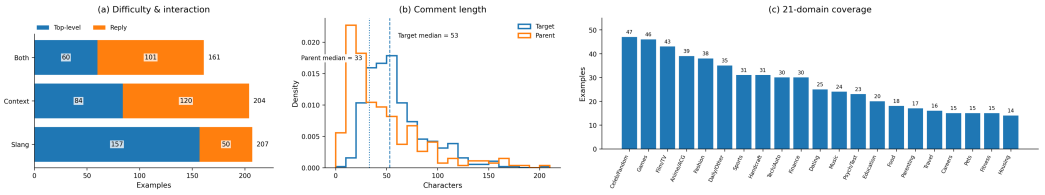

Figure 3: Dataset composition by difficulty type, interaction structure, comment length, and topical domain.

challenge. Approximately 5,000 examples are selected for subsequent verification. The annotation prompt and the 21 domain labels are provided in Appendix B.1 and Appendix B.5, respectively.

Difficulty Verification and Balancing. We next verify translation difficulty using independently generated translations. Candidates are assessed for substantive semantic, cultural, slang-related, or contextual variation across translations. For **Context** and **Both** examples, we additionally compare translations generated with and without interaction context to confirm contextual dependence. Approximately 3,000 examples remain after verification.

We then apply character-level n-gram deduplication and balance the candidate pool across the 21 topical domains and three difficulty types, resulting in 2,205 candidates for human review. The corresponding prompts are provided in Appendix B.2, Appendix B.3, and Appendix B.4.

Human Review. Three graduate-student reviewers independently examine the 2,205 candidates, covering translation difficulty, difficulty-type validity, and context relevance, with cross-checking across reviewers. The review also includes screening for sensitive, inappropriate, private, and personally identifiable information. After human verification, 572 examples remain and constitute CULTURE-CMT.

2.3 Dataset Statistics

Figure 3 summarizes CULTURE-CMT along four dimensions: difficulty type, interaction structure, topical domain, and comment length.

The benchmark contains 207 **Slang**, 204 **Context**, and 161 **Both** examples, together with 301 top-level comments and 271 replies. Their interaction structures differ noticeably across difficulty types: **Slang** is dominated by top-level comments (157/207), whereas replies account for a larger share of **Context** (120/204) and **Both** (101/161). This pattern reflects the greater conversational dependence of the latter two categories.

CULTURE-CMT spans 21 social-media domains. The largest categories are Celebrities/Fandom (47), Games (46), and Film & TV (43), while the smallest domain still contains 14 examples, providing broad topical coverage across diverse social-media settings.

Target comments are concise overall but exhibit a long-tailed length distribution, with a median of 53 characters and a mean of 58.9. Available parent comments are shorter, with a median of 33 characters and a mean of 45.7. Together, these statistics reflect the short-form yet highly variable nature of social-media interaction.

3 Evaluation Protocol

CULTURE-CMT evaluates whether a translation preserves not only the literal meaning of a social-media comment, but also its culturally situated meaning, affect, and communicative intent. Following the cultural-effectiveness criterion of CULTURE-MT (Wu et al., 2026), each translation receives a four-level score

$$s \in \{0, 1, 2, 3\},$$

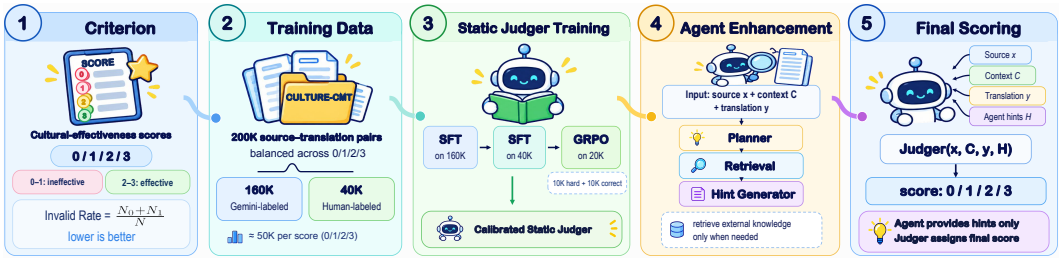

Figure 4: Overview of the CULTURE-CMT evaluation framework. The cultural-effectiveness Judger is trained through two-stage SFT and GRPO, and is augmented with retrieved cultural knowledge at inference time.

where scores 0–1 indicate culturally ineffective translations and scores 2–3 indicate culturally effective ones. We adopt the CULTURE-MT scoring rubric, summarized together with our annotation procedure in Appendix E.2.

Our primary metric is the cultural invalidity rate (**Invalid Rate**),

$$R_{\text{invalid}} = \frac{N_0 + N_1}{N},$$

where N_k is the number of examples receiving score k and $N = 572$ is the benchmark size. Missing translation outputs are assigned score 0. Lower values indicate better performance.

3.1 Cultural-effectiveness Judger

To support scalable evaluation, we train a Qwen3.5-27B cultural-effectiveness Judger (Figure 4). Its training corpus contains 200K source–translation pairs from public social-media data, balanced across the four score levels. Among them, 160K are labeled by Gemini 3.1 Pro and 40K are manually annotated under the CULTURE-MT criterion (Wu et al., 2026). The Judger training data are disjoint from the 40K source pool used to construct CULTURE-CMT. We additionally reserve development sets of 1,000–2,000 examples during training and a separate 432-pair human-annotated set for final Judger evaluation. Data separation details are provided in Appendix E.1.

Training proceeds in three stages. We first perform SFT on the 160K Gemini-labeled examples and then continue SFT on the 40K human-labeled examples. We subsequently initialize GRPO from the resulting checkpoint using 20K human-labeled examples, comprising 10K examples that remain challenging after SFT and 10K correctly judged examples. The GRPO reward combines accuracy and format objectives: correct binary and four-way judgments receive rewards of 1 and 0.5, respectively, while the format reward encourages analysis followed by a final score. Training uses Relax (Relax Contributors, 2026); full configurations are provided in Appendix E.3.

The resulting Judger predicts

$$\hat{s} = f_\theta(x, \mathcal{C}, y),$$

where x is the source comment, $\mathcal{C}$ is its interaction context, and y is the candidate translation.

3.2 Agent-enhanced Judger

Social-media evaluation frequently requires knowledge of rapidly evolving slang, memes, fandom terminology, named entities, and platform-specific expressions. We therefore augment the trained Judger with a retrieval agent that supplies relevant cultural knowledge at inference time.

Given $(x, \mathcal{C}, y)$, the evaluation process is

$$\begin{aligned} Q &= g_{\text{plan}}(x, \mathcal{C}, y), \\ D &= R(Q), \\ H &= g_{\text{hint}}(x, \mathcal{C}, y, Q, D), \\ \hat{s} &= f_\theta(x, \mathcal{C}, y, H), \end{aligned}$$

where g_{plan} identifies knowledge needed for evaluation and generates search queries, R retrieves relevant evidence, and g_{hint} converts the evidence into concise, translation-specific judging hints.

Retrieved evidence serves as auxiliary knowledge for the calibrated Judger, which produces the final 0–3 cultural-effectiveness score. This design combines a learned evaluation criterion with dynamically retrieved cultural information, allowing the evaluator to accommodate emerging social-media usages at inference time. Full agent prompts are provided in Appendix C.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Models. We evaluate 33 model configurations spanning multiple families and scales, including Qwen3.5-9B, 27B and 35B-A3B (Team, 2026), Qwen3.6-27B(Google, 2026a), Qwen3.8-27B and Qwen3.8-Max (Qwen Team, 2026), GLM-5.2 and GLM-5.3 (GLM-5-Team et al., 2026), Gemini 3.1 Pro (Google, 2026c) and Gemini 3.7 Flash (Google, 2026b), Gemma-4-31B and Gemma-4-26B-A4B (Team et al., 2026a), Claude Opus 5 (Anthropic, 2026), GPT-5.6 (OpenAI, 2026), and Kimi K3 (Team et al., 2026b). Where both inference modes are available, we evaluate *thinking* and *non-thinking* configurations of the same model. Cross-model results are interpreted descriptively, as model families differ in architecture and training, while thinking and non-thinking configurations use their native inference settings.

We also evaluate **Qwen3.6-27B-SFT**, a 27B model adapted for social-media translation using 1.2M training examples strictly disjoint from CULTURE-CMT. This provides a task-adapted comparison alongside the general-purpose models. Details of data construction and supervised fine-tuning are provided in Appendix D.

Evaluation. We follow the cultural-effectiveness evaluation protocol described in Section 3. Our primary metric is the proportion of translations receiving scores 0 or 1, denoted as the **Invalid Rate**; lower values indicate better cultural effectiveness. We additionally report the proportion of score-3 translations and the mean score. Judger inference uses greedy decoding (temperature 0). Contextualized and comment-only translations are evaluated with their corresponding with-context and without-context Judger prompts, so each condition reflects the complete translation-and-evaluation setting available under that context regime.

4.2 EVALUATION OF THE CULTURAL-EFFECTIVENESS JUDGER

We assess the Judger against human annotations on translations of final CULTURE-CMT examples.

Data and Metrics. For each of the 33 model configurations, we sample 25 source–translation pairs: 5 **Slang**, 10 **Context**, and 10 **Both**, yielding 825 candidates. After human annotation, we sample by the adjudicated 0–3 labels to construct an approximately score-balanced 432-pair evaluation set. This set is reserved for final Judger evaluation. Appendix E.2 describes the double-annotation and adjudication procedure.

We report binary accuracy and Cohen's κ for the ineffective/effective partition ($\{0, 1\}$ versus $\{2, 3\}$), together with four-way accuracy and Spearman's ρ over the original scores. The balanced evaluation set enables comparison across evaluators under comparable score coverage.

Results. The combined agent enhancement increases observed binary accuracy by 1.16 and 4.17 percentage points for the SFT and GRPO-aligned Judgers, respectively, while improving all four reported metrics for both backbones (Figure 5). GRPO alone improves ordinal agreement (ρ: 0.5889 to 0.6131), but reduces binary and four-way accuracy. The Agent Judger with the GRPO-aligned backbone achieves the highest binary accuracy (77.31%), κ (0.5498), and ρ (0.6554), whereas the SFT-based Agent Judger slightly leads in four-way accuracy. The main translation experiments use the GRPO-based Agent Judger. Its agreement with humans is assessed on the separate 432-pair evaluation set described above.

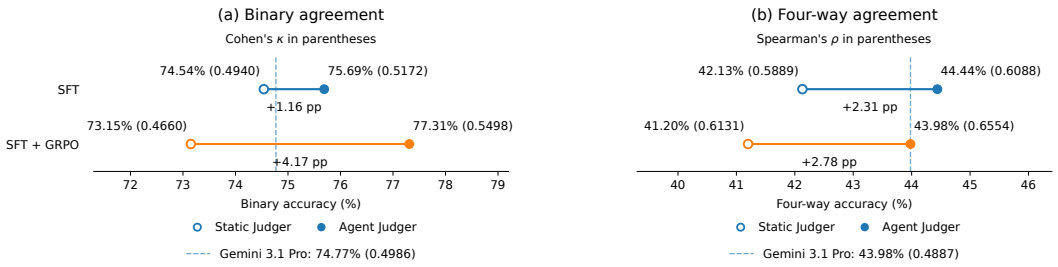

Figure 5: Agreement with human judgments for static and agent-enhanced Judgers ($N = 432$). (a) Binary accuracy with Cohen's κ in parentheses; (b) four-way accuracy with Spearman's ρ in parentheses. Gains are computed from unrounded accuracies.

4.3 Overall Results and Difficulty Profiles

Table 1 summarizes the complete score distributions under both contextualized and comment-only settings. We report the proportions of scores 0–3, the cultural invalidity rate (**Invalid**↓, i.e., scores 0 and 1), and the mean score. To quantify the effect of interaction context, we define

$$\Delta_{\text{Invalid}} = R^{\text{no-ctx}}_{\text{Invalid}} - R^{\text{ctx}}_{\text{Invalid}}, \qquad \Delta_{\text{Mean}} = \mu^{\text{ctx}} - \mu^{\text{no-ctx}},$$

where positive values indicate better scores in the contextualized condition. Translation and Judger prompts both change between conditions. Within each inference mode, models are ordered by contextualized **Invalid**↓.

Effect of Interaction Context. Most configurations achieve lower Invalid Rates and higher mean scores when interaction context is available, although the magnitude of the improvement varies substantially across models. For example, non-thinking GLM-5.2 improves from 30.1% to 21.7% Invalid, while thinking GPT-5.6 Terra improves by 8.7 percentage points. By contrast, thinking Gemini 3.1 Pro has the same Invalid Rate in both settings, with only a small increase in mean score. These results show that interaction context is broadly useful, but models differ considerably in how effectively they exploit it.

4.4 Difficulty Profiles under Thinking Inference

We next describe difficulty profiles for the evaluated thinking configurations with context.

Figure 6 shows Invalid rates on the **Slang**, **Context**, and **Both** subsets for representative models. Across the models shown, **Both** is consistently the most challenging subset. Even systems that perform relatively well on **Slang** or **Context** individually often exhibit substantially higher Invalid rates on **Both**, indicating that the combination of contextual dependence and socially situated language remains difficult. For example, Qwen3.6-27B-SFT has an Invalid Rate of 26.7% on Both, compared with 15.9% on Slang and 12.3% on Context. GLM-5.2 shows a similar profile, with 17.9%, 12.7%, and 26.7% on the three subsets, respectively.

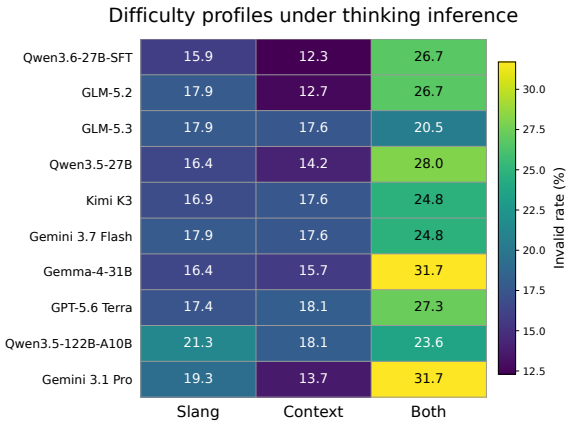

Figure 6: Invalid rates (%) by difficulty under thinking inference with context. Subset sizes are 207 (Slang), 204 (Context), and 161 (Both).

The figure also shows that similar overall performance can mask different difficulty profiles. Some models are relatively stronger on **Context** but weaker on **Slang**, while others show the opposite tendency. For instance, Qwen3.5-27B performs competitively on **Context** (14.2%) but remains much weaker on

Table 1: Overall CULTURE-CMT results with and without interaction context ($N = 572$ in each condition). Missing outputs receive score 0; gains use unrounded values. Lower **Invalid** is better; higher **Mean** is better.

	With Context						Without Context						Context Gain	
Model	**0**	**1**	**2**	**3**	**Invalid↓**	**Mean↑**	**0**	**1**	**2**	**3**	**Invalid↓**	**Mean↑**	Δ_{Inv} ↑	Δ_{Mean} ↑
Non-thinking inference														
Qwen3.8-Max	0.2	20.1	25.5	54.2	**20.3**	**2.337**	0.3	23.6	27.6	48.4	**24.0**	**2.241**	+3.7	+0.096
GLM-5.2	1.6	20.1	30.6	47.7	21.7	2.245	1.0	29.0	29.7	40.2	30.1	2.091	**+8.4**	+0.154
Qwen3.8-27B	0.3	21.3	29.9	48.4	21.7	2.264	0.7	24.7	26.6	48.1	25.3	2.220	+3.7	+0.044
Claude Opus 5	0.7	22.2	28.8	48.3	22.9	2.247	0.5	27.3	27.8	44.4	27.8	2.161	+4.9	+0.086
GPT-5.6 Terra	1.0	21.9	25.7	51.4	22.9	2.274	1.0	27.6	28.7	42.7	28.7	2.129	+5.8	+0.145
Gemma-4-31B	1.2	22.7	26.0	50.0	24.0	2.248	1.4	29.7	30.2	38.6	31.1	2.061	+7.2	**+0.187**
Qwen3.6-27B-SFT (ours)	1.2	23.6	29.7	45.5	24.8	2.194	1.2	29.4	29.9	39.5	30.6	2.077	+5.8	+0.117
Qwen3.5-27B	1.2	24.1	30.2	44.4	25.3	2.178	1.9	31.1	30.9	36.0	33.0	2.010	+7.7	+0.168
Qwen3.5-397B	0.9	24.5	28.8	45.8	25.3	2.196	1.9	29.2	26.4	42.5	31.1	2.094	+5.8	+0.101
Qwen3.5-122B-A10B	1.9	26.0	29.7	42.3	28.0	2.124	3.5	30.1	29.2	37.2	33.6	2.002	+5.6	+0.122
Gemma-4-26B-A4B	1.0	29.0	24.7	45.3	30.1	2.142	1.9	31.8	30.4	35.8	33.7	2.002	+3.7	+0.140
Qwen3.6-27B	2.3	32.3	28.7	36.7	34.6	1.998	3.1	33.6	26.6	36.7	36.7	1.969	+2.1	+0.030
Qwen3.5-35B-A3B	3.8	31.1	29.4	35.7	35.0	1.969	5.1	35.7	29.2	30.1	40.7	1.843	+5.8	+0.126
MiniMax M3	3.8	32.7	30.8	32.7	36.5	1.923	3.8	32.9	28.8	34.4	36.7	1.939	+0.2	-0.016
Qwen3.5-9B	10.7	34.1	30.4	24.8	44.8	1.694	6.6	39.7	33.6	20.1	46.3	1.671	+1.6	+0.023
Thinking inference														
Qwen3.6-27B-SFT (ours)	0.5	17.1	27.6	54.7	**17.7**	**2.365**	0.7	22.2	28.0	49.1	22.9	2.255	+5.2	+0.110
GLM-5.2	0.5	18.0	28.3	53.1	18.5	2.341	0.2	23.3	28.3	48.3	23.4	2.247	+4.9	+0.094
GLM-5.3	0.7	17.8	26.0	55.4	18.5	2.362	0.9	19.2	24.5	55.4	**20.1**	**2.344**	+1.6	+0.017
Qwen3.5-27B	0.2	18.7	30.1	51.0	18.9	2.320	0.3	23.4	30.8	45.5	23.8	2.213	+4.9	+0.107
Kimi K3	0.3	19.1	26.7	53.8	19.4	2.341	0.7	21.5	28.8	49.0	22.2	2.260	+2.8	+0.080
Gemini 3.7 Flash	1.0	18.7	25.0	55.2	19.8	2.344	0.5	22.9	25.5	51.0	23.4	2.271	+3.7	+0.073
Gemma-4-31B	1.0	19.4	25.0	54.5	20.5	2.330	0.5	24.7	26.7	48.1	25.2	2.224	+4.7	+0.107
GPT-5.6 Terra	0.5	19.9	27.4	52.1	20.5	2.311	1.0	28.1	27.4	43.4	29.2	2.131	**+8.7**	**+0.180**
Qwen3.5-122B-A10B	0.3	20.5	30.1	49.1	20.8	2.280	0.2	26.4	33.2	40.2	26.6	2.135	+5.8	+0.145
Gemini 3.1 Pro	0.0	20.8	24.3	54.9	20.8	2.341	0.0	20.8	26.7	52.4	20.8	2.316	+0.0	+0.024
Qwen3.5-397B	0.2	21.2	30.4	48.3	21.3	2.267	0.5	22.2	31.5	45.8	22.7	2.226	+1.4	+0.042
Qwen3.5-35B-A3B	1.0	21.3	32.9	44.8	22.4	2.213	1.6	26.9	32.7	38.8	28.5	2.087	+6.1	+0.126
Qwen3.6-27B	0.3	22.7	28.1	48.8	23.1	2.253	2.4	26.4	29.2	42.0	28.8	2.107	+5.8	+0.147
Qwen3.8-27B	0.2	23.1	30.2	46.5	23.3	2.231	1.2	26.9	30.6	41.3	28.1	2.119	+4.9	+0.112
Gemma-4-26B-A4B	0.9	23.3	26.9	49.0	24.1	2.240	1.0	27.4	28.3	43.2	28.5	2.136	+4.4	+0.103
Claude Opus 5	1.2	23.1	26.2	49.5	24.3	2.240	1.0	24.3	29.4	45.3	25.3	2.189	+1.0	+0.051
Qwen3.8-Max	0.5	24.3	28.3	46.9	24.8	2.215	0.3	25.5	26.9	47.2	25.9	2.210	+1.0	+0.005
Qwen3.5-9B	0.5	29.2	35.8	34.4	29.7	2.042	1.4	35.5	34.8	28.3	36.9	1.900	+7.2	+0.142

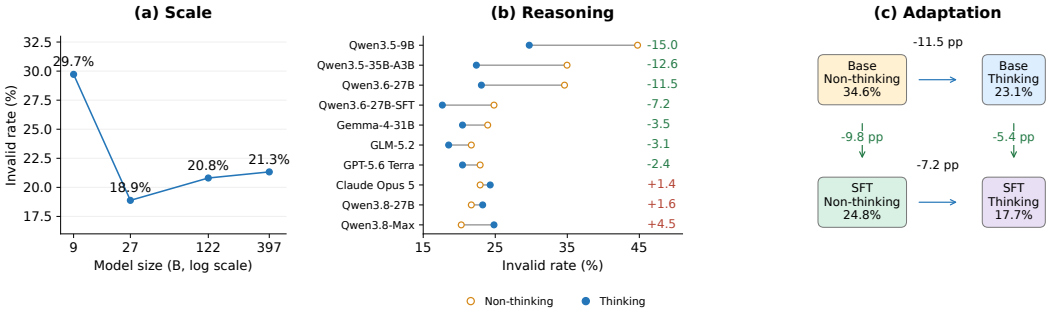

Figure 7: Scale, reasoning, and adaptation under the contextualized setting. **(a) Scale:** Invalid Rates across the evaluated Qwen3.5 sizes are non-monotonic. **(b) Reasoning:** thinking/non-thinking differences vary across models. **(c) Adaptation:** SFT of the same 27B backbone lowers observed Invalid Rates in both modes.

Both (28.0%). By contrast, Kimi K3 exhibits a more balanced profile across the three subsets. The three-way partition thus makes model-specific difficulty profiles visible beyond a single aggregate score.

Our adapted Qwen3.6-27B-SFT has its lowest Invalid Rate on **Context**, while its higher error rate on **Both** identifies the combined difficulty as a remaining challenge under thinking inference.

4.5 Scale, Reasoning, and Adaptation

We next compare three common routes for improving LLM-based social-media translation: **Scale**, **Reasoning**, and **Adaptation**. Figure 7 summarizes their performance under the contextualized setting.

Scale. Within Qwen3.5 under thinking inference, Invalid decreases from 29.7% at 9B to 18.9% at 27B, while the 122B-A10B and 397B variants obtain 20.8% and 21.3%. The observed ranking is therefore non-monotonic in total parameter count. For mixture-of-experts models, total and active parameter counts also differ.

Reasoning. Thinking inference lowers Invalid by 15.0, 12.6, and 11.5 percentage points for Qwen3.5-9B, Qwen3.5-35B-A3B, and Qwen3.6-27B, respectively. Claude Opus 5, Qwen3.8-27B, and Qwen3.8-Max instead have higher Invalid Rates under thinking inference. The benefit of switching to thinking mode therefore varies across models.

Adaptation. SFT of the same Qwen3.6-27B backbone lowers Invalid from 34.6% to 24.8% without thinking and from 23.1% to 17.7% with thinking. The adapted model achieves the lowest observed contextualized Invalid Rate among the evaluated configurations.

Scope. CULTURE-CMT is a targeted stress test constructed from social-media data collected between April 12 and May 12, 2026. Accordingly, our results characterize the evaluated benchmark and model configurations rather than population-level platform error rates. The scale and reasoning analyses are descriptive comparisons across existing configurations, while the adaptation comparison captures the full effect of the proposed SFT procedure.

5 Conclusion

We introduced **CULTURE-CMT**, a 572-example benchmark for Chinese–English social-media comment translation that preserves the post and conversational context naturally available to users. The benchmark supports diagnostic analysis through its **Slang**, **Context**, and **Both** subsets under the cultural-effectiveness rubric inherited from CULTURE-MT. We further develop a Qwen3.5-27B cultural-effectiveness Judger augmented with retrieved cultural information and evaluate its agreement with human annotations. Across the evaluated configurations, interaction context generally improves cultural effectiveness, while the **Both** subset remains the most challenging among the models analyzed. Performance is non-monotonic across model sizes, the benefit of thinking varies substantially by model, and targeted adaptation of the same 27B backbone lowers Invalid Rates in both inference modes. Overall, these findings highlight the importance of jointly modeling interaction context and culturally situated language, and suggest targeted adaptation as a promising direction for improving social-media translation.

AI Use Statement

We used generative AI tools to assist with manuscript organization and grammatical editing, as well as figure preparation and consistency checks during revision. The model-assisted data screening, annotation, translation generation and refinement, and reasoning-trace generation used in this study are described in Section 2.2, Section 3.1, and Appendix D. The authors manually reviewed the AI-assisted revisions for correctness and consistency with the reported results and take full responsibility for the final content, including the text, figures, analyses, and conclusions.

References

Anthropic. Introducing claude opus 5. https://www.anthropic.com/news/claude-opus-5, 2026. Accessed: 2026-09-25.

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. How Noisy Social Media Text, How Diffrnt Social Media Sources? In *Proceedings of the Sixth Inter-*

national Joint Conference on Natural Language Processing (IJCNLP), 2013. URL https://aclanthology.org/I13-1041/.

Shanbo Cheng, Yu Bao, Qian Cao, Luyang Huang, Liyan Kang, et al. Seed-X: Building Strong Multilingual Translation LLM with 7B Parameters, 2025. URL https://arxiv.org/abs/2507.13618.

Patrick Fernandes, Kayo Yin, Emmy Liu, André Martins, and Graham Neubig. When does translation require context? A data-driven, multilingual exploration. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 606–626, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.36. URL https://aclanthology.org/2023.acl-long.36/.

GLM-5-Team, :, Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, Chenzheng Zhu, Congfeng Yin, Cunxiang Wang, Gengzheng Pan, Hao Zeng, Haoke Zhang, Haoran Wang, Huilong Chen, Jiajie Zhang, Jian Jiao, Jiaqi Guo, Jingsen Wang, Jingzhao Du, Jinzhu Wu, Kedong Wang, Lei Li, Lin Fan, Lucen Zhong, Mingdao Liu, Mingming Zhao, Pengfan Du, Qian Dong, Rui Lu, Shuang-Li, Shulin Cao, Song Liu, Ting Jiang, Xiaodong Chen, Xiaohan Zhang, Xuancheng Huang, Xuezhen Dong, Yabo Xu, Yao Wei, Yifan An, Yilin Niu, Yitong Zhu, Yuanhao Wen, Yukuo Cen, Yushi Bai, Zhongpei Qiao, Zihan Wang, Zikang Wang, Zilin Zhu, Ziqiang Liu, Zixuan Li, Bojie Wang, Bosi Wen, Can Huang, Changpeng Cai, Chao Yu, Chen Li, Chengwei Hu, Chenhui Zhang, Dan Zhang, Daoyan Lin, Dayong Yang, Di Wang, Ding Ai, Erle Zhu, Fangzhou Yi, Feiyu Chen, Guohong Wen, Hailong Sun, Haisha Zhao, Haiyi Hu, Hanchen Zhang, Hanrui Liu, Hanyu Zhang, Hao Peng, Hao Tai, Haobo Zhang, He Liu, Hongwei Wang, Hongxi Yan, Hongyu Ge, Huan Liu, Huanpeng Chu, Jia'ni Zhao, Jiachen Wang, Jiajing Zhao, Jiamin Ren, Jiapeng Wang, Jiaxin Zhang, Jiayi Gui, Jiayue Zhao, Jijie Li, Jing An, Jing Li, Jingwei Yuan, Jinhua Du, Jinxin Liu, Junkai Zhi, Junwen Duan, Kaiyue Zhou, Kangjian Wei, Ke Wang, Keyun Luo, Laiqiang Zhang, Leigang Sha, Liang Xu, Lindong Wu, Lintao Ding, Lu Chen, Minghao Li, Nianyi Lin, Pan Ta, Qiang Zou, Rongjun Song, Ruiqi Yang, Shangqing Tu, Shangtong Yang, Shaoxiang Wu, Shengyan Zhang, Shijie Li, Shuang Li, Shuyi Fan, Wei Qin, Wei Tian, Weining Zhang, Wenbo Yu, Wenjie Liang, Xiang Kuang, Xiangmeng Cheng, Xiangyang Li, Xiaoquan Yan, Xiaowei Hu, Xiaoying Ling, Xing Fan, Xingye Xia, Xinyuan Zhang, Xinze Zhang, Xirui Pan, Xu Zou, Xunkai Zhang, Yadi Liu, Yandong Wu, Yanfu Li, Yidong Wang, Yifan Zhu, Yijun Tan, Yilin Zhou, Yiming Pan, Ying Zhang, Yinpei Su, Yipeng Geng, Yong Yan, Yonglin Tan, Yuean Bi, Yuhan Shen, Yuhao Yang, Yujiang Li, Yunan Liu, Yunqing Wang, Yuntao Li, Yurong Wu, Yutao Zhang, Yuxi Duan, Yuxuan Zhang, Zezhen Liu, Zhengtao Jiang, Zhenhe Yan, Zheyu Zhang, Zhixiang Wei, Zhuo Chen, Zhuoer Feng, Zijun Yao, Ziwei Chai, Ziyuan Wang, Zuzhou Zhang, Bin Xu, Minlie Huang, Hongning Wang, Juanzi Li, Yuxiao Dong, and Jie Tang. Glm-5: from vibe coding to agentic engineering, 2026. URL https://arxiv.org/abs/2602.15763.

Google. Qwen3.6-27b: Flagship-level coding in a 27b dense model. https://www.alibabacloud.com/blog/603063, 2026a. Accessed: 2026-09-24.

Google. Gemini 3.7 flash model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-7-Flash-Model-Card.pdf, 2026b. Accessed: 2026-09-13.

Google. Gemini 3.1 pro: Best for complex tasks and bringing creative concepts to life. https://deepmind.google/models/gemini/pro/, 2026c. Accessed: 2026-09-13.

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

Chen Huang, Junkai Luo, Xinzuo Wang, Wenqiang Lei, and Jiancheng Lv. Can large language models understand internet buzzwords through user-generated content. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.acl-long.632/.

Prathyusha Jwalapuram, Barbara Rychalska, Shafiq Joty, and Dominika Basaj. Can your context-aware MT system pass the DiP benchmark tests? Evaluation benchmarks for discourse phenomena in machine translation, 2020. URL https://arxiv.org/abs/2004.14607.

Sophia Yat Mei Lee and Helena Yan Ping Lau. An Event-Comment Social Media Corpus for Implicit Emotion Analysis. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC 2020)*, 2020. URL https://aclanthology.org/2020.lrec-1.203/.

Yunlong Liang, Fandong Meng, Jiaan Wang, and Jie Zhou. SlangDIT: Benchmarking LLMs in interpretative slang translation, 2025. URL https://arxiv.org/abs/2505.14181.

Lingrui Mei, Shenghua Liu, Yiwei Wang, Baolong Bi, and Xueqi Cheng. SLANG: New concept comprehension of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2024. URL https://arxiv.org/abs/2401.12585.

Paul Michel and Graham Neubig. MTNT: A Testbed for Machine Translation of Noisy Text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018. URL https://aclanthology.org/D18-1050/.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No Language Left Behind: Scaling Human-Centered Machine Translation, 2022. URL https://arxiv.org/abs/2207.04672.

OpenAI. Gpt-5.6: Frontier intelligence that scales with your ambition. https://openai.com/zh-Hans-CN/index/gpt-5-6/, 2026. Accessed: 2026-09-13.

Qwen Team. Qwen3.8-max: A new bar for coding and cowork, August 2026. URL https://qwen.ai/blog?id=qwen3.8.

Relax Contributors. Relax: An asynchronous reinforcement learning engine for omni-modal post-training at scale. https://github.com/redai-studio/Relax, 2026. Software repository. Accessed: 2026-09-26.

Zhewei Sun, Qian Hu, Rahul Gupta, Richard Zemel, and Yang Xu. Toward informal language processing: Knowledge of slang in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2024. URL https://arxiv.org/abs/2404.02323.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, Clément Farabet, Johan Ferret, Filippo Galgani, Sertan Girgin, Jean-Bastien Grill, Maarten Grootendorst, Jiaxian Guo, Cassidy Hardin, Yanzhang He, Steven M. Hernandez, Omri Homburger, Léonard Hussenot, Juyeong Ji, Armand Joulin, Aishwarya Kamath, Parnian Kassraie, Olivier Lacombe, Preethi Lahoti, Gaël Liu, Gus Martins, Luciano Martins, Tatiana Matejovicova, Ramona Merhej, Nikola Momchev, Sneha Mondal, Ryan Mullins, Sindhu Raghuram Panyam, Shreya Pathak, Sarah Perrin, André Susano Pinto, Etienne Pot, Angéline Pouget, Alexandre Ramé, Sabela Ramos, Douglas Reid, David Rim, Morgane Rivière, Karsten Roth, Louis Rouillard, Omar Sanseviero, Pier Giuseppe Sessa, Shane Settle, Danila Sinopalnikov, Sara Smoot, Piotr Stanczyk, Andreas Steiner, Lawrence Stewart, Ilya Tolstikhin, Michael Tschannen, Anton Tsitsulin, Nino Vieillard, Renjie Wu, Pingmei Xu, Haichuan Yang, Edouard Yvinec, Biao Zhang, Li Zhang, Joe Zou, Nicolas Aagnes, Abdelrahman Abdelhamed, Jakub Adamek, Shivani Agrawal, Shubham Agrawal, Ibrahim Alabdulmohsin, Jean Baptiste Alayrac, Uri Alon, Chandramouli Amarnath, Ankesh Anand, Chrysovalantis Anastasiou, Setareh Ariafar, François-Xavier

Aubet, Kyriakos Axiotis, Federico Barbero, Joelle Barral, Alexei Bendebury, Urs Bergmann, Stanley Bileschi, Kat Black, Mathieu Blondel, Sebastian Borgeaud, Arthur Bražinskas, Ryan Burnell, Robert Busa-Fekete, Mu Cai, Daniele Calandriello, Glenn Cameron, Charlotte Caucheteux, Rahma Chaabouni, Garima Chadha, Jetha Chan, Blake Jianhang Chen, Jesse Chen, Lin Chen, Xu Chen, Derek Cheng, Tzu hsiang Chien, Nikolai Chinaev, Yi Chou, Zhaohui Chu, Benjamin Coleman, Pooja Consul, Sam Conway-Rahman, Scott Crowell, Dylan Cutler, Vivek Dani, Samira Daruki, Anil Das, Daniel Deutsch, Nishanth Dikkala, Li Ding, Qiuhan Ding, Shenil Dodhia, Konstantin Donhauser, Tulsee Doshi, Anca Dragan, Alex Druinsky, Sahil Dua, Zoltan Egyed, Danielle Eisenbud, Daniel Eppens, Cindy Fan, Bahare Fatemi, Yassir Fathullah, Vlad Feinberg, Milen Ferev, Sebastian Flennerhag, Takumi Fujimoto, João Gabriel Oliveira, Isaac Galatzer-Levy, João Gante, Simon Geisler, Soham Ghosal, Antonious M. Girgis, Tamara von Glehn, Alec Go, Alhaad Gokhale, Alex Grills, Yiming Gu, Mayank Gupta, Pramod Gupta, Guru Guruganesh, Raia Hadsell, Hamza Harkous, Jitendra Harlalka, Demis Hassabis, Anja Hauth, Joe Heyward, Arian Hosseini, Chih-Yang Hsia, I-Hung Hsu, Xiaopeng Huang, Yangsibo Huang, Kevin Hui, Adrian Hutter, Te I, Fotis Iliopoulos, Advait Jain, Ganesh Jawahar, Ziwei Ji, Qilin Jin, Melvin Johnson, Kandarp Joshi, Arun Kandoor, Wang-Cheng Kang, Koray Kavukcuoglu, Mehran Kazemi, Kathleen Kenealy, Amr Khalifa, Phoebe Kirk, Ivan Korotkov, Suraj Kothawade, Vitaly Kovalev, Neel Kovelamudi, Adam Kraft, Ravin Kumar, Vivek Kumar, Harish Kuppam, Justin Lannin, Chen-Yu Lee, Seungji Lee, Dmitry Lepikhin, Alon Levkovitch, Dongdong Li, Qiujia Li, Valentin Liévin, Ethan Lin, Ziqian Lin, Casper Liu, Tianlin Liu, Tianqi Liu, Xin Liu, Ivan Lobov, Mayank Lunayach, Min Ma, Gagan Madan, Andrii Maksai, Eric Malmi, Michal Matuszak, Daniel McDuff, Gaurav Menghani, Maciej Mikuła, Daniil Mirylenka, Karolis Misiunas, Vedant Misra, Andreea Mitran, Kareem Mohamed, Maksim Mukha, Eric Noland, James O'Donnell, Brendan O'Donoghue, Kate Olszewska, Bernett Orlando, Wanqiong Pan, Rina Panigrahy, Unnati Parekh, Nicolas Perez-Nieves, Chunjong Park, Eric Paskie, Liqian Peng, Bryce Petrini, Slav Petrov, Jonas Pfeiffer, Bilal Piot, Martyna Plomecka, Siim Poder, Octavio Ponce, Arijit Pramanik, David Racz, Anish Rajan, Michelle Ramanovich, Anand Rao, Marvin Ritter, Vitor Rodrigues, Evan Rosen, Mikołaj Rybiński, Noveen Sachdeva, Michaël E. Sander, Rohit Sathyanarayana, Sagar Savla, Samuel Schmidgall, Tal Schuster, George Scrivener, Benoit Seguin, Andrew Sellergren, Aliaksei Severyn, Izhak Shafran, Dhruv Shah, Bobak Shahriari, Yuan Shangguan, Ashish Shenoy, Pradeep Shenoy, Rakesh Shivanna, Pauline Sho, Lucas Spangher, Wojciech Stokowiec, Tim Strother, Yao Su, Yinghao Sun, Mukund Sundararajan, Andrea Tacchetti, Mor Hazan Taege, Pouya Tafti, Jean Tarbouriech, Chetan Tekur, Shantanu Thakoor, Rahul Thapa, Madeleine Traverse, Lenart Treven, Tao Tu, Chien Te Tung, Çağlar Ünlü, Petar Veličković, Malini Pooni Venkat, Sagar Gubbi Venkatesh, Vidya Venkiteswaran, Francesco Visin, Alex Vitvitskyi, Kiran Vodrahalli, Weiyi Wang, Xin Wang, Tris Warkentin, Jan Wassenberg, John Wieting, Cindy Wu, Lechao Xiao, Hao Xu, Yuhui Xu, Fuzhao Xue, Arun Yadav, Jun Yan, Antoine Yang, Lin Yang, Ming-Hsuan Yang, Ziyu Ying, Jae Hyeon Yoo, Morteza Zadimoghaddam, Sajjad Zafar, Fred Zhang, Jiageng Zhang, Jianyi Zhang, Xiaofan Zhang, Chao Zhao, David Zhou, and Chen Zou. Gemma 4 technical report, 2026a. URL https://arxiv.org/abs/2607.02770.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, M. C., Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y. Charles, H. S. Che, Guanduo Chen, Guangyu Chen, Guanzheng Chen, Huarong Chen, Jia Chen, Jianlong Chen, Jun Chen, Kexin Chen, Peng Chen, Ruijue Chen, Wentao Chen, Xin Chen, Yang Chen, Yanru Chen, Yifei Chen, Yingjiang Chen, Yuankun Chen, Yujie Chen, Yutian Chen, Zhirong Chen, Dazhi Cheng, Yean Cheng, Jialei Cui, Jingbing Cui, Anqi Dai, Jiaqi Deng, Hao Ding, Rui Ding, Shaofeng Ding, Mengfan Dong, Mengnan Dong, Yuhao Dong, Yuxin Dong, Angang Du, Chenzhuang Du, Dikang Du, Jusen Du, Yulun Du, Yu Fan, Jing Feng, Qiulin Feng, Yichen Feng, Kelin Fu, Qiang Fu, Fuxuan Gao, Hongcheng Gao, Jingyue Gao, Tong Gao, Weijia Gao, Shangyi Geng, Jie Gong, Linhu Gong, Shengao Gong, Xiaochen Gong, Qizheng Gu, Yicheng Gu, Shuhao Guan, Haiqing Guo, Shiqi Guo, Xiang Guo, Zhengyan Guo, Beixi Hao, Wenxin Hao, Xiaoru Hao, Dailan He, Haotian He, Lehan He, Qi He, Weiran He, Xinran He, Xinyi He, Yibo He, Yunjia He, Chao Hong, Tiange Hong, Hao Hu, Jiaxi Hu, Ruikun Hu, Weiming Hu, Yangyang Hu, Zhenxing Hu, Liang Hua, Jinbin Huang, Ke Huang, Ruiyuan Huang, Siying Huang, Weixiao Huang, Yan Huang, Zhengjie Huang, Zhiqi Huang, Yulong Hui, Chaobo Jia, Yutong Jiang, Zhejun Jiang, Zuoyou Jiang, Wenyi Jin, Xinyi Jin, Yu Jing, Huanjun Kong, Guokun Lai, Aidi Li, Cheng Li, Chengyuan Li, Cong Li, Fang Li, Guanyu Li, Haoyang Li, Jia Li, Junxiong Li, Lei Li, Letian Li, Lincan Li, Weihong Li, Wentao Li, Xintong Li, Yang Li, Yishen Li, Yiwei Li, Yuxiao Li, Zhaowei Li, Zhaoxi Li, Zheming Li, Zhengxiao Li, Zhiyuan

Li, Jiawei Lin, Xiaohan Lin, Yibo Lin, Zichao Lin, Ziyan Lin, Bill Liu, Boxiao Liu, Chuan Liu, Liang Liu, Shaowei Liu, Shudong Liu, Shuran Liu, Tianwei Liu, Weizhou Liu, Yangyang Liu, Yanming Liu, Yibo Liu, Yipeng Liu, Zhengying Liu, Zhiheng Liu, Enzhe Lu, Haoyu Lu, Linqiang Lu, Tingzhan Lu, Zhiyuan Lu, Aotian Luo, G. Luo, Junyu Luo, Yifan Luo, B. Lyu, Wenzhou Lyu, Shaoguang Mao, Yuan Mei, Xin Men, Minqing Ni, Yixuan Niu, Siyuan Pan, Shujun Peng, Zhangyang Qi, Ruoyu Qin, ZeChao Qin, Zeyu Qin, Haiquan Qiu, Jianxin Qiu, Jiezhong Qiu, Bowen Qu, Yuhao Qu, Zeyu Shang, Youbo Shao, Han Shen, Jincheng Shi, Juanfeng Shi, Lidong Shi, Shengyuan Shi, Wingchun Siu, Pengwei Song, Xiaoxi Song, Jianlin Su, Yunfeng Su, Zhaochen Su, Lin Sui, Jingsong Sun, Junyao Sun, Shaoning Sun, Shuzhe Sun, Tongyu Sun, Yujun Sun, Yunpeng Tai, Chuning Tang, Heyi Tang, Sirui Tang, Zecheng Tang, Chaoran Tian, Rongpeng Tian, Yu Tian, Wei Tu, Chensi Wang, Chuang Wang, Chunjie Wang, Dinglu Wang, Feng Wang, Hailong Wang, Haiming Wang, Hao Wang, Hao Wang, Huaqing Wang, Hui Wang, Jiayi Wang, Jinglong Wang, Jinhong Wang, Jiuzheng Wang, Linian Wang, Shaobo Wang, Shenzhi Wang, Shuyi Wang, Si Wang, Siyuan Wang, Tianfu Wang, Wenjue Wang, Xingran Wang, Xinmei Wang, Xinyuan Wang, Xusheng Wang, Yalin Wang, Yangkun Wang, Yao Wang, Yaoyu Wang, Yejie Wang, Yiqin Wang, Yucheng Wang, Yuzhi Wang, Zhaoji Wang, Zhaowei Wang, Zhengtao Wang, Zhenhao Wang, Zhongsheng Wang, Zifan Wang, Chu Wei, Ming Wei, Shouxin Wei, Zichen Wen, Fan Wu, Haoning Wu, Rucong Wu, Wenhao Wu, Xiaoxue Wu, Yingcong Wu, Yongqi Wu, Yuxin Wu, Zijian Wu, Xinglang Xian, Chenxuan Xiang, Yuye Xiang, Bocheng Xiao, Chenjun Xiao, Xin Xiao, Jin Xie, Xiaotong Xie, Yifeng Xie, Zhe Xie, Bowei Xing, Yiming Xiong, Baosheng Xu, Boyu Xu, Jiale Xu, Jianfan Xu, Jing Xu, Jinjing Xu, L. H. Xu, Qingtao Xu, Shuyao Xu, Suting Xu, Tiantian Xu, Tianxiang Xu, Weixin Xu, Xinran Xu, Yangchuan Xu, Ye Xu, Yueni Xu, Ziyao Xu, Haonan Xue, Junjie Yan, Yaoyao Yan, Fan Yang, Guangyao Yang, Hao Yang, Junwei Yang, Ruoyu Yang, Wenjie Yang, Xiaofei Yang, Xinyu Yang, Yi Yang, Yiling Yang, Ying Yang, Yuchen Yang, Zhen Yang, Zhilin Yang, Zian Yang, Zuhao Yang, Haotian Yao, Dan Ye, Haoran Ye, Wenjie Ye, Zhanbo Ye, Bohong Yin, Haoxiang Yin, Xietong Yin, Chengzhen Yu, Haozhen Yu, Longhui Yu, Shengnan Yu, Shuying Yu, Tianxiang Yu, Enming Yuan, Mengjie Yuan, Tongtian Yue, Wei Yue, Yang Yue, Dunyuan Zha, Haobing Zhan, B. H. Zhang, Dehao Zhang, Fei Zhang, Hao Zhang, Haoyuan Zhang, Huanyu Zhang, Jiapei Zhang, Jiaxuan Zhang, Jin Zhang, Kaiyi Zhang, Miaozhen Zhang, Puqi Zhang, Qinglei Zhang, Rong Zhang, Rui Zhang, Shaoshuai Zhang, Shiyi Zhang, Xiaobin Zhang, Xiaoyun Zhang, Y. Zhang, Yangkun Zhang, Ye Zhang, Yichi Zhang, Yikun Zhang, Yizhi Zhang, Yongting Zhang, Yu Zhang, Yutao Zhang, Yutong Zhang, Zheng Zhang, Zijing Zhang, Bin Zhao, Chenguang Zhao, Feifan Zhao, Jinglun Zhao, Jinxiang Zhao, Shuai Zhao, Wenshuo Zhao, Xiangyu Zhao, Xuanle Zhao, Yikai Zhao, Zijia Zhao, Haozhi Zheng, Huabin Zheng, Ruihan Zheng, Shaojie Zheng, Tengyang Zheng, Haofeng Zhong, Lei Zhong, Longguang Zhong, M. Zhou, Qiankang Zhou, Runjie Zhou, Ruozhang Zhou, Xinyu Zhou, Yiqiao Zhou, Zaida Zhou, Jinguo Zhu, Liya Zhu, Xinhao Zhu, Yangjunfeng Zhu, Yuxuan Zhu, Zhen Zhu, Chen Zhuang, Weiyu Zhuang, and Xinxing Zu. Kimi k3: Open frontier intelligence, 2026b. URL https://arxiv.org/abs/2607.24653.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models, 2023. URL https://arxiv.org/abs/2203.11171.

Linjuan Wu, Ruiqi Zhang, Xinze Lyu, Ye Guo, Daoxin Zhang, Zhe Xu, Yao Hu, Yixin Cao, Yongliang Shen, and Weiming Lu. Beyond literal translation: Evaluating cultural effectiveness in social media ugc, 2026. URL https://arxiv.org/abs/2605.25626.

Ifeoluwa Wuraola, Nina Dethlefs, and Daniel Marciniak. Understanding slang with LLMs: Modelling cross-cultural nuances through paraphrasing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA, 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.emnlp-main.869/.

Chiyu Zhang, Muhammad Abdul-Mageed, and Ganesh Jawahar. Contrastive Learning of Sociopragmatic Meaning in Social Media. In *Findings of the Association for Computational Linguistics: ACL 2023*, 2023. URL https://aclanthology.org/2023.findings-acl.152/.

Kaiyan Zhao, Zheyong Xie, Zhongtao Miao, Xinze Lyu, Yao Hu, and Shaosheng Cao. Benchmarking machine translation on chinese social media texts. *CoRR*, abs/2601.22931, 2026. doi: 10.48550/ARXIV.2601.22931. URL https://doi.org/10.48550/arXiv.2601.22931.

Mao Zheng, Zheng Li, Tao Chen, Bo Lv, Mingrui Sun, Mingyang Song, Jinlong Song, Hong Huang, Decheng Wu, Hai Wang, Yifan Song, Yanfeng Chen, and Guanwei Zhang. Hy-mt2: A family of fast, efficient and powerful multilingual translation models in the wild, 2026. URL https://arxiv.org/abs/2605.22064.

A Related Work

A.1 Social Media and Culturally Grounded Translation

Machine translation for social-media text has long been recognized as challenging because of its noisy, informal, and rapidly evolving nature. One challenge is robustness to linguistic noise, which MTNT (Michel & Neubig, 2018) studies through naturally occurring Reddit comments. A complementary challenge is that translations must preserve not only understandable content, but also community-specific meanings, tone, and expressive style. RedTrans (Guo et al., 2025) addresses this through specialized training for social-media translation, while CULTURE-MT (Wu et al., 2026) evaluates cultural effectiveness through expression accuracy and cultural adaptability. CSM-MTBench (Zhao et al., 2026) further distinguishes longer posts from short, reactive comments, highlighting difficulties in preserving slang, tone, and style. Related studies examine slang interpretation and cross-cultural nuances more directly (Sun et al., 2024; Wuraola et al., 2024). Studies of emerging slang and internet buzzwords further examine how contextual usage helps interpret unfamiliar expressions (Mei et al., 2024; Huang et al., 2025).

CULTURE-CMT complements these efforts by evaluating translation within the original interaction in which a comment occurs. It retains the corresponding post and parent comment where available, preserving the contextual information naturally accessible to readers. The **Slang**, **Context**, and **Both** subsets additionally support separate analyses of linguistic-cultural difficulty, contextual dependence, and their combination.

A.2 Context-aware and Discourse-level Translation

DiP diagnoses anaphora, lexical consistency, coherence/readability, and discourse connectives in machine translation (Jwalapuram et al., 2020). MuDA uses a data-driven procedure to identify translation phenomena requiring context and supports multilingual discourse evaluation (Fernandes et al., 2023). CULTURE-CMT complements this line of work by retaining explicit post and parent-comment fields and selecting culturally situated comment-translation challenges. Its contribution is the combination of interaction structure, difficulty-based selection, three diagnostic subsets, and cultural-effectiveness evaluation, rather than the introduction of context into MT evaluation.

A.3 Scale, Reasoning, and Adaptation for Translation

Large-scale multilingual training has advanced machine translation (NLLB Team et al., 2022), while specialized systems demonstrate the potential of more compact models. Seed-X (Cheng et al., 2025) develops 7B multilingual translation models, and Hy-MT2 (Zheng et al., 2026) similarly targets efficient translation in real-world settings. RedTrans (Guo et al., 2025) explores domain-specific training for social-media translation, while SlangDIT (Liang et al., 2025) investigates reasoning-based interpretative slang translation.

We bring these directions into a shared evaluation setting to examine **Scale**, **Reasoning**, and **Adaptation**. Specifically, we compare models across sizes, evaluate thinking and non-thinking modes within the same model where available, and include a social-media translation model fine-tuned on benchmark-disjoint data. The contribution is a descriptive assessment under a common cultural-effectiveness criterion, rather than a controlled decomposition of the three strategies. Our model coverage does not include Seed-X or Hy-MT2. Prior CULTURE-MT results place the evaluated Seed-X variants below Qwen3-32B on cultural effectiveness (Wu et al., 2026), motivating our fo-

cus on Qwen-family baselines; this earlier result does not establish their relative performance on CULTURE-CMT. The omission of specialized MT systems limits the scope of the comparison.

B Prompts for Benchmark Construction

This appendix provides the prompts used in the construction of CULTURE-CMT. All prompts operate on the original Chinese social-media content. For clarity, we present the prompts in English while preserving the exact task definitions and output formats used in our pipeline.

B.1 Model-assisted Candidate Annotation

After rule-based candidate mining, we use an LLM to determine whether each candidate contains a genuine translation challenge and to assign its difficulty type. The model is given the target comment together with its original interaction context.

Candidate Annotation Prompt

System Prompt
You are an expert reviewer for a challenging social-media machine translation benchmark. Evaluate each example strictly according to the criteria below and return only valid JSON.

Input
Target comment: {comment}
Post title: {post_title}
Post content: {post_content}
Parent comment: {parent_comment, if available}

Task
Determine whether the target comment is suitable for a challenging Chinese–English social-media translation benchmark.
An example should be marked as `pass` only if it contains a genuine translation challenge. Ordinary conversational expressions that can be translated directly and reliably should be marked as `reject`.
Assign one of the following difficulty types:

- **Context**: The intended meaning of the target comment cannot be reliably recovered from the comment alone and requires the surrounding post or conversation.
- **Slang**: The target comment contains genuine internet slang, abbreviations, community-specific expressions, culturally situated usages, or other non-standard language that requires non-literal translation or cultural adaptation.
- **Both**: The example contains both substantial contextual dependence and slang/cultural-language difficulty.

Do not label an example as **Context** merely because the context is related to the comment. The context must be important for determining the intended meaning or an appropriate translation.
Do not label ordinary colloquial expressions as **Slang**. The expression should involve genuine internet, community-specific, or culturally situated usage.
Also identify:

- the topical domain of the target comment;
- the main translation challenge(s).

Output Format

```
{
  "quality": "pass" or "reject",
  "difficulty_type": "Context" or "Slang" or "Both",
  "domain": "...",
  "challenges": ["...", "..."]
}
```

B.2 Independent Translation Generation

For candidates that pass the annotation stage, we generate independent translations for difficulty verification. The translation model receives the target comment and, when required by the experimental condition, its interaction context.

Translation Generation Prompt

System Prompt
Translate the following Chinese social-media comment into natural English. Preserve its intended meaning, tone, and social-media style. Use the provided context only to understand the target comment. Do not translate or summarize the context itself. Output only the English translation.

Input
Post title: {post_title}
Post content: {post_content}
Parent comment: {parent_comment, if available}
Target comment: {comment}

For the comment-only condition used in context-dependence verification, the same instruction is used but the post and parent-comment fields are omitted.

B.3 Translation Difficulty Verification

Model-assisted annotation may retain examples that appear unusual but do not produce substantial translation difficulty. We therefore generate two translations independently and ask a verifier to determine whether their differences reflect a genuine translation challenge or only stylistic variation.

Translation Difficulty Verification Prompt

System Prompt
You are evaluating whether a social-media comment presents a genuine machine translation challenge. Return only valid JSON.

Input
Target comment: {comment}
Interaction context: {context}
Translation A: {translation_a}
Translation B: {translation_b}

Task
Compare Translation A and Translation B.
Determine whether their difference reflects a substantive translation challenge in the source comment.
Return `hard` if the difference is caused by at least one of the following:

- different interpretations of the source meaning;
- slang, internet language, or community-specific expressions;
- culturally grounded or metaphorical language;
- implicit or omitted information whose recovery affects the translation;
- interaction context that changes the intended interpretation.

Return `style` if both translations convey essentially the same meaning and differ mainly in:

- word choice;
- sentence structure;
- fluency;

- degree of formality;
- other stylistic preferences.

Do not mark an example as `hard` solely because the two translations use different wording.

Output Format

```
{
  "reason": "hard" or "style",
  "explanation": "brief reason"
}
```

B.4 Context-dependence Verification

For candidates labeled as **Context** or **Both**, we perform an additional verification step to ensure that the interaction context is genuinely necessary for recovering the intended meaning.

For the same target comment, we generate one translation using the target comment alone and another using the full interaction context. A verifier then determines whether access to context leads to a substantive improvement or change in interpretation.

Context-dependence Verification Prompt

System Prompt
You are evaluating whether a social-media comment genuinely depends on its interaction context for correct translation. Return only valid JSON.

Input
Target comment: {comment}
Post title: {post_title}
Post content: {post_content}
Parent comment: {parent_comment, if available}
Translation without context: {translation_no_context}
Translation with context: {translation_with_context}

Task
Determine whether the interaction context is necessary to recover the intended meaning of the target comment.
Mark the example as `context-dependent` if the translation without context:

- misinterprets what the comment refers to;
- assigns an incorrect meaning to an expression or metaphor;
- misses important implicit information required by the target language;
- expresses a substantially different stance or communicative intent;
- or otherwise fails to recover the intended meaning that becomes clear when the context is provided.

Mark the example as `context-independent` if the comment can be translated reliably without the surrounding context, even if the context makes the interpretation slightly clearer or more specific.
The existence of topical relevance alone is not sufficient for a `context-dependent` label. The context must materially affect the interpretation or translation.

Output Format

```
{
  "label": "context-dependent"
           or "context-independent",
  "explanation": "brief reason"
}
```

B.5 Domain Labels

We use 21 topical domains for annotation and balancing: Games; Anime / ACG; Film and Television; Celebrities / Fandom; Music; Relationships / Dating; Psychology / Personality Tests; Food; Fashion / Beauty; Fitness / Exercise; Sports; Travel; Housing / Home; Careers / Job Seeking; Education / Learning; Parenting / Family; Pets; Drawing / Handcraft; Technology / Automotive; Finance / Investment; and Daily Life / Other.

C Agent-enhanced Judger Prompts

This appendix provides the prompts used by the retrieval component of the Agent-enhanced Judger described in Section 3.2. The retrieval module supplies auxiliary cultural knowledge when necessary, while the trained cultural-effectiveness Judger retains final scoring authority. All prompts are presented in English for readability.

C.1 Search Planning Prompt

The search planner determines whether external knowledge is required to evaluate a candidate translation and generates a small number of searchable queries when necessary.

Search Planning Prompt

System Prompt

You are a pre-analysis assistant for social-media translation evaluation.
Given a Chinese source text, optional interaction context, and its English translation, determine whether external web search is required to correctly evaluate the translation.
Return JSON only. Do not use Markdown.

```
{
  "needs_search": ["query1", "query2"],
  "reason": "brief explanation"
}
```

Follow these rules:

1. Search only when the input contains knowledge-sensitive content, such as:
   - internet memes or emerging slang;
   - fandom, anime, gaming, or community-specific jargon;
   - platform-specific expressions;
   - names of brands, works, people, places, or events;
   - culturally loaded expressions that may be mistranslated literally;
   - expressions whose meaning may have changed over time.
2. Do not search for ordinary words, standard grammar, or expressions whose meaning is already clear from the source and context.
3. Each query should be short and independently searchable.
4. Preserve Chinese source expressions in Chinese when they are the object of the search.
5. If an official English alias is required, use both the Chinese expression and the English alias in the query.
6. Generate at most four queries.
7. Return an empty list when no external knowledge is required.

User Template

```
[Source]
{source}
```

```
[Context]
{context}

[Translation]
{translation}
```

C.2 Hint Generation Prompt

The hint generator converts retrieved information into concise, translation-specific judging points. It does not assign the final cultural-effectiveness score.

Hint Generation Prompt

System Prompt
You are a rigorous but restrained assistant for social-media translation evaluation.
Your task is not to assign a final score. Instead, provide concise and referenceable judging points for a downstream cultural-effectiveness Judger.
Given the source text, interaction context, candidate translation, search queries, and retrieved evidence, identify only the information that is directly relevant to evaluating the translation.
Return JSON only. Do not use Markdown.

```
{
  "judging_points": [
    "point 1",
    "point 2"
  ],
  "risk_flags": [
    "risk label 1",
    "risk label 2"
  ]
}
```

Follow these requirements:

1. Judging points should explain how the candidate translation should be evaluated.
2. Relevant aspects may include:
   - missing or mistranslated source meaning;
   - slang, meme, or community-specific meaning;
   - culturally grounded metaphors or references;
   - correctness of named entities and official aliases;
   - preservation of affect, stance, humor, or pragmatic intent;
   - whether a literal translation loses the intended cultural meaning;
   - whether the register is appropriate for the original social-media context.
3. Retrieved evidence is auxiliary information only. Ignore evidence that is irrelevant, contradictory, or unreliable.
4. Do not fabricate cultural knowledge that is not supported by either the source, context, or retrieved evidence.
5. You may identify potential translation problems, but you must not assign a final score of 0, 1, 2, or 3.
6. For emerging slang, fandom expressions, special forms of address, memes, or titles of works, explain the intended meaning and the potential risk of literal translation.
7. Generate at most six judging points.
8. Keep each judging point concise and directly useful for final evaluation.

User Template

```
[Source]
{source}

[Context]
{context}

[Translation]
{translation}

[Search Queries]
{queries}

[Search Results]
{search_results}
```

C.3 Hint Injection Template

The structured hint object is serialized into an auxiliary block and inserted into the Judger input. The prompt explicitly instructs the Judger to treat the hints as supplementary evidence rather than authoritative labels.

Hint Injection Template

```
Referenceable judging points:

These points are auxiliary information only.
Rely primarily on the source text, interaction context,
and candidate translation.
Do not adopt the hints blindly.

1. {judging_point_1}
2. {judging_point_2}
...
Possible risk flags:
{risk_flag_1}, {risk_flag_2}, ...

Please evaluate the candidate translation using the
four-level cultural-effectiveness criterion.

[Source]
{source}

[Context]
{context}

[Translation]
{translation}
```

D Social-Media Translation Fine-Tuning

Data Collection and Refinement. We collect 1.5M Chinese comments from public social-media platforms between April 12 and May 12, 2026, together with their available post and parent-comment context. Following the data-distillation and refinement approach of CULTURE-MT (Wu et al., 2026), we use Gemini to generate English translations and a static cultural-effectiveness Judger to assess their quality. Translations requiring improvement are refined by Gemini using Judger feed-

back. After filtering and refinement, we retain 1.2M contextualized source–translation examples, strictly disjoint from CULTURE-CMT. The translation-generation prompts are provided below.

Prompt for Contextualized Comment Translation

```
You are an expert in social media translation. Please translate the
following content into English. Follow these requirements:
1. Placeholders in the form ``#placeholder-x#'' must not be
translated. Preserve each placeholder at the corresponding position
in the translated text.
2. Output only the translated result. Do not provide explanations or
any additional content.
3. Preserve structural tags such as <title></title>,
<content></content>, and <comment></comment>. The translated result
must follow the same tag structure as the input.
When the sentence to be translated contains an unclear reference or an
ambiguous meaning, use the following post or parent-comment context to
determine its intended meaning:
{context}
Return only the final translation and no other content.
Sentence to be translated:
{input}
Please provide the translated result:
```

Thinking and Non-thinking Supervision. For each retained sample, GLM-5 generates a concise reasoning trace consisting of *context analysis* and *translation points*. The context analysis summarizes the information relevant to interpreting the comment, while the translation points identify key expressions and explain their translation considerations. This task-specific supervision encourages attention to contextual and cultural challenges without unnecessarily long reasoning. The reasoning-generation prompt is provided below.

Prompt for Reasoning Trace Generation

```
You are a multilingual translation expert for Xiaohongshu-style social
media content.
Given the context, source text, and reference translation, identify
the key translation points that are necessary to produce a translation
like the reference.
First, identify important words or phrases directly from the
source text, and then briefly explain their meanings or the main
considerations for translating them.
Keep the analysis concise and direct.
Example:
Context: None
Source: 真的是自嗨的剧情，不好看。这不是高智商，这是走到哪里都有神助的玛丽苏
Translation: The plot is truly self-indulgent and unengaging; this
isn't a display of high intelligence but a Mary Sue story where the
protagonist is divinely favored and effortlessly succeeds wherever she
goes.
Translation Points:
Pay attention to the following points when translating:
1. ``自 嗨'' refers to a plot that is immersed in the creator's own
self-satisfaction while ignoring the audience's perspective; it can be
rendered as ``self-indulgent plot.''
2. ``高 智 商'' here refers to either a plot that requires logical
reasoning from the audience or characters whose actions demonstrate
intelligence.
3. ``神 助'' refers to the protagonist repeatedly receiving convenient
external help whenever difficulties arise.
4. ``玛 丽 苏'' comes from the literary term ``Mary Sue'' and refers to
an unrealistically favored protagonist who seems to have a built-in
advantage or halo.
Context: {context}
```

```
Source: {content_zh}
Translation: {trans_en}
Translation Points:
```

We organize the retained data into an equal mixture of thinking and non-thinking supervision, with 600K examples in each subset. Thinking examples include the generated reasoning trace followed by the final English translation, whereas non-thinking examples retain only the final translation as the supervision target. Both subsets are combined to train a single model that supports evaluation in either inference mode.

Training Configuration. We train **Qwen3.6-27B-SFT** on the combined 1.2M-example dataset using the `Relax` framework (Relax Contributors, 2026). Training runs for one epoch with a global batch size of 256 and a learning rate of 5×10^{-6}. Development validation examples are separated from the training collection, as described in Appendix E.1.

E Data Separation, Annotation, and Judger Configuration

E.1 Data provenance and separation

All source collections in this study were gathered between April 12 and May 12, 2026. The 40K benchmark source comments and the 200K Judger source–translation pairs are disjoint collections gathered during the same period; they do not constitute a temporal train–test split. The 1.2M retained translation-training examples are disjoint from the final CULTURE-CMT benchmark. Development validation sets contain 1,000–2,000 examples held out from the corresponding training collections. These development sets are separate from the 432-pair final human evaluation of Judger outputs. The latter consists of multi-model translations of final benchmark examples and is not used for training, validation during training, checkpoint selection, or evaluator selection.

Table 2: Roles of the data collections. Collection sizes describe the data pools; development validation examples are held out from training.

Collection	Size	Role and stated separation
Benchmark source pool	40K comments	Source of the 2,205 candidates and final 572 examples; disjoint from the Judger collection.
Judger collection	200K pairs	160K Gemini-labeled and 40K human-labeled pairs for staged training and held-out development validation.
Judger GRPO subset	20K pairs	Subset of the human-labeled Judger collection, not an additional independent collection.
Translation collection	1.5M collected; 1.2M retained	Translation adaptation; retained examples are disjoint from CULTURE-CMT.
Development validation	1K–2K per set	Held out from training collections; used separately from final evaluation.
Final Judger evaluation	825 candidates; 432 retained	Multi-model benchmark translations, human-labeled and subsampled by score; evaluation only.

The static Judger used for translation-data refinement and the agent-enhanced evaluator follow the same cultural-effectiveness criterion. Their shared criterion is part of the training/evaluation design, not evidence of independent human confirmation of the adapted model's ranking. The reported collection-level separation does not by itself establish that all related posts, threads, or recurring expressions are independent across collections.

E.2 Human annotation and scoring rubric

Nine Chinese–English bilingual experts annotate translation quality in both the training and evaluation collections. Each batch is annotated twice. In each round, its examples are divided equally among three experts, so each example receives one label per round, not three votes. The two rounds use six different experts; assignments are randomized across batches rather than fixed throughout the project. After both rounds, the remaining three experts adjudicate disagreements and determine the

final labels. The mean binary agreement between the two rounds is 82.46%, with labels 0–1 mapped to ineffective and labels 2–3 to effective. This is an agreement statistic for the two annotation rounds, not nine-way agreement or four-class exact agreement.

Benchmark candidate review is a separate task: three additional graduate students divide the 2,205 candidates equally, review difficulty and context relevance, and cross-check their work. The resulting 572 examples form the benchmark. For final Judger evaluation, 25 pairs are sampled from each of 33 model configurations (5 Slang, 10 Context, and 10 Both), giving 825 candidate pairs. Human annotation precedes subsampling by adjudicated score to obtain the approximately balanced 432-pair set.

We reuse the four-level CULTURE-MT rubric without changing its definitions (Wu et al., 2026). Its two dimensions concern fidelity to meaning and affect, and appropriate cultural interpretation and natural target-language expression. The following is a concise summary, not a new scoring scale:

- **0:** Major failure to convey the intended meaning or cultural intent.
- **1:** Some meaning is understandable, but cultural communication remains ineffective.
- **2:** Communication is generally effective, with residual weaknesses in style or cultural expression.
- **3:** Meaning, affect, and culturally situated intent are conveyed effectively and naturally.

The effective/ineffective boundary remains between scores 1 and 2. The Gemini judging setup follows CULTURE-MT's protocol, with Gemini 3.1 Pro used in this study and the context input matched to the evaluation condition.

E.3 JUDGER TRAINING AND INFERENCE

All trained Judgers use Qwen3.5-27B. GRPO starts from the Judger checkpoint obtained after the two SFT stages, and uses the Relax framework (Relax Contributors, 2026) with a Megatron training backend and SGLang rollouts. The Judger's GRPO training reward is

$$r = r_{\text{accuracy}} + r_{\text{format}}.$$

The accuracy component rewards agreement with the annotated score: 1 for a correct ineffective/effective decision and 0.5 for a correct four-level score. The format component rewards an analysis followed by the final score. The configuration below records the supplied GRPO launch settings; it does not describe the separate translation SFT run.

Table 3: Judger GRPO launch configuration. Training sampling and deterministic evaluation use different temperatures.

Setting	Value
Initialization / reference	Qwen3.5-27B Judger SFT checkpoint
Rollout iterations / prompts per batch	120 / 64
Samples per prompt / global batch size	4 / 64
Training rollout temperature	0.9
Maximum generated response length	2,048 tokens (training and evaluation)
Evaluation samples / temperature	1 per prompt / 0 (greedy)
Evaluation / saving interval settings	10 / 10
Initial evaluation	Skipped by the launch configuration
Qwen chat-template thinking flag	Disabled (`enable_thinking=false`)
Advantage estimator	GRPO
Policy clipping (lower / upper)	0.20 / 0.28
KL coefficient / entropy coefficient	0.00 / 0.00
Optimizer / learning rate schedule	Adam / constant 10^{-6}
Weight decay / Adam β_1, β_2	0.1 / 0.9, 0.98
Training and rollout allocation	Colocated, 32 GPUs
Tensor / pipeline / context parallelism	4 / 1 / 1
Tokens per GPU / rollout engine size	8,192 / 4 GPUs
SGLang static memory fraction	0.4
Attention / hidden dropout	0 / 0

Disabling the chat template's built-in thinking mode is distinct from the task's required analysis-before-score output format. Final Judger scoring uses temperature 0. The training rollout temperature of 0.9 applies only to generating GRPO samples. Contextualized translations use the with-context Judger prompt; comment-only translations use the without-context prompt. The agent planning and hint templates are given in Appendix C.