KNOWLEDGE, MECHANISMS, AND CONTEXT: DIAGNOSING AND IMPROVING SOCIAL UGC UNDERSTANDING — @SITU

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding social user-generated content (UGC) requires interpreting evolving expressions and integrating meaning across posts and interactions. Existing benchmarks largely evaluate downstream task labels or individual phenomena, leaving these abilities difficult to assess together. We introduce SITU (*Situated Interpretation of Textual UGC*), a bilingual benchmark built from Chinese and English social-media posts and discussions. It contains 89,199 open-ended interpretation probes from three platforms, including 9,522 difficulty-filtered probes on which frontier models remain well below perfect performance. Nine coverage-oriented probe targets span expression-centered and context-integrative challenges. To study improvement, we analyze item knowledge, reusable expression mechanisms, and contextual integration through retrieval, expression oracles, supervised fine-tuning (SFT), and training-data scaling. Retrieval helps most when it supplies answer-critical knowledge; SFT produces broader gains, and their combination performs best overall. A full-stream Chinese exposure audit finds persistent gains on mechanism-transferable expressions after excluding detected same-item knowledge. Matched downstream training further shows benefits for translation and homophone-robust toxicity detection, with substantial variation across tasks and backbones. These results establish SITU as a testbed for evaluating social UGC understanding and identifying which improvements can transfer beyond interpretation. We will release the benchmark and evaluation code.

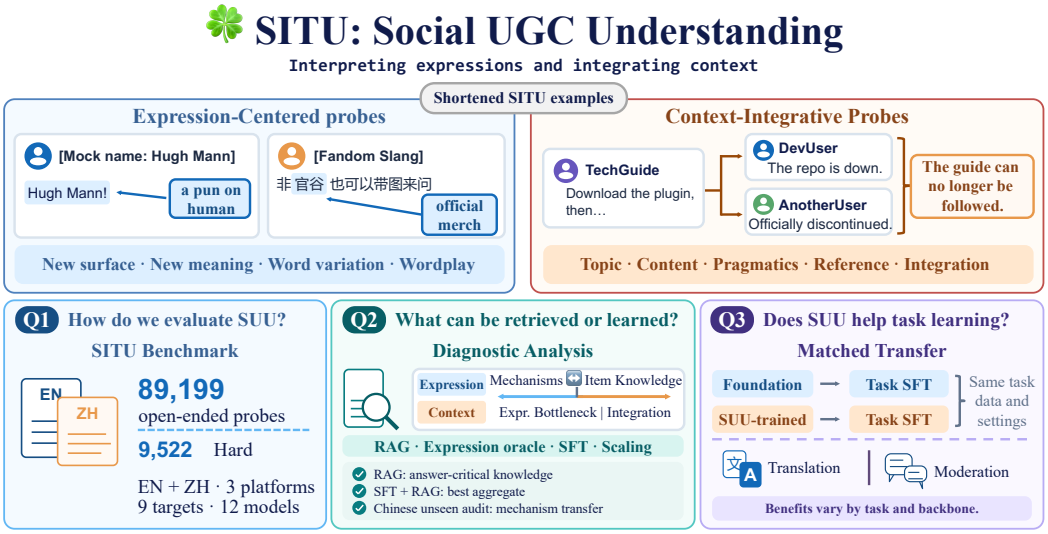

Figure 1: Overview of SITU benchmark construction, diagnostic interventions, and downstream transfer for social UGC understanding.

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 continually changes how people express and connect ideas. Familiar words acquire community-specific meanings, spelling and sound are repurposed for humor or evasion, and short replies depend on participants and context distributed across a thread. For language models, understanding this user-generated content (UGC) requires more than fluent sentence processing: it requires recovering what an expression or contribution means in its particular use. This capability, which we call *social UGC understanding* (SUU), underlies applications such as translation, moderation, and conversational information extraction.

Existing evaluations illuminate parts of this problem. Broad suites such as TweetEval, SuperTweetEval, and SPARROW organize social-media tasks around sentiment, stance, toxicity, and other outputs (Barbieri et al., 2020; Antypas et al., 2023; Zhang et al., 2023). More focused benchmarks directly test emerging expressions, memes, humor, or difficult comments (Huang et al., 2025; Xie et al., 2025; Lin et al., 2026; Hong et al., 2026). Together they reveal substantial weaknesses, but their different task contracts make it difficult to assess expression interpretation and contextual understanding in a common setting. They also leave open whether improving one aspect of understanding helps another, or benefits downstream task learning.

We introduce SITU (*Situated Interpretation of Textual UGC*) to study these questions. SITU is built from Chinese Xiaohongshu and English Reddit and YouTube records, with privacy-preserving rewriting and iterative reference auditing. Each probe asks an open-ended question about the meaning of a specific expression, contribution, or interaction. Nine coverage-oriented probe targets guide extraction across two scopes: *expression-centered* and *context-integrative*. Expression probes extend beyond slang to concepts, entities, changing senses, and productive variation; context probes cover individual contributions and their relations across a discussion. The resulting suite contains 89,199 probes, including a 9,522-probe Hard subset. Evaluation of twelve models finds substantial room for improvement, including among frontier systems.

Beyond measuring performance, we ask what enables improvement. Four diagnostic aspects organize interventions around item knowledge, reusable expression mechanisms, and contextual integration (Section 5). Retrieval and expression-oracle experiments test the role of accessible knowledge, while SFT and scaling test the benefits of interpretation supervision. Retrieval is especially effective for item knowledge when search supplies the needed evidence. SFT produces broader gains, and combining the two achieves the strongest aggregate results. A full-stream Chinese exposure audit finds that gains persist on mechanism-transferable expressions after excluding detected same-item knowledge.

Finally, we test whether SUU supervision improves downstream learning by holding task data and training settings fixed between Foundation and SUU initializations. Gemma benefits consistently on the primary translation metric and retains a moderation advantage on unseen homophonic variation after clean-only task training. Qwen replications and context-controlled moderation reveal a more selective picture: task supervision can absorb the advantage, and some settings regress.

Our contributions are threefold: (1) a bilingual, open-ended benchmark that unifies broad expression and contextual interpretation; (2) intervention and exposure analyses characterizing complementary improvements from retrieval and SFT, including transfer beyond detected item exposure; and (3) matched downstream experiments establishing both the benefits and boundaries of SUU supervision.

2 RELATED WORK

2.1 EVALUATING SOCIAL UGC UNDERSTANDING

TweetEval, SPARROW, and SuperTweetEval consolidate established social-media tasks, predominantly using labels or spans (Barbieri et al., 2020; Zhang et al., 2023; Antypas et al., 2023). SNS-Bench adds reading comprehension, query correction, and generation (Guo et al., 2025b). These suites measure application performance, whereas interpretation-focused benchmarks directly assess meaning recovery: CHEER and CHIME target Chinese internet expressions and memes (Huang et al., 2025; Xie et al., 2025); Chumor 2.0, MOUSE, and CNeo-Bench examine humor, meme mechanisms, and emerging concepts (He et al., 2024; Lin et al., 2026; Zhao et al., 2026a). CIBuzzBench tests buzzword definitions, cross-lingual senses, and harmfulness (Wang et al., 2026), while *You Really*

Didn't Get That? evaluates difficult Chinese social-media comments in context (Hong et al., 2026). MemeLens extends interpretation to multilingual image–text memes (Shahroor et al., 2026). SITU brings broader expression and context coverage into one bilingual, open-ended textual task, enabling comparisons of interventions across these challenges. Appendix A.3 compares benchmark coverage.

2.2 Adapting models for social-media tasks

BERTweet and TimeLMs adapt pretraining to tweets and temporal drift (Nguyen et al., 2020; Loureiro et al., 2022); SMILE studies domain adaptation, and MultiLexNorm addresses lexical normalization (Bashlovkina et al., 2023; van der Goot et al., 2021). Application-specific work includes RedTrans for SNS translation and SlangDIT/SlangOWL for slang-aware translation (Guo et al., 2025c; Liang et al., 2025). In moderation, HED-COLD studies homophonic variation, SeTox newly toxic senses and terms, and CounterContext context-dependent hate/counterspeech classification (Wu et al., 2025; Cui et al., 2026; Yu et al., 2022). These applications motivate our downstream transfer tests.

Broader systems such as RedOne, RedOne 2.0, and TaCL-CoMoE combine continued pretraining, multi-task post-training, or expert routing (Zhao et al., 2025b;a; Xingren et al., 2025). Their emphasis is the adaptation recipe or target application. We use standard SFT to study which interpretive abilities improve and whether SUU supervision benefits subsequent task learning.

3 Social UGC Understanding and SITU

3.1 What is Social UGC Understanding?

Social UGC understanding (SUU) is the ability to interpret user-generated content in its post, interaction, and community context. It includes understanding evolving expressions, recovering what a contribution communicates, and connecting information across an interaction. These abilities support direct interpretation tasks (Huang et al., 2025; Lin et al., 2026; Hong et al., 2026) and downstream applications such as translation, moderation, and information extraction.

We focus on textual UGC and distinguish two scopes. *Expression-centered* probes ask about a local form or meaning, including slang, neologisms, community concepts, familiar words used in new senses, spelling variation, and wordplay. *Context-integrative* probes ask about a contribution or its relation to surrounding material, including semantic content, pragmatic intent, reference, and thread-level synthesis. SITU evaluates both through a common open-ended task.

3.2 SITU: Task Design and Construction

Task and coverage. Given a source record and a narrowly scoped question, a model generates an interpretation scored against a frozen reference answer and supporting explanation. A record contains a post or video's textual content and bounded, ordered comment context. Nine *probe targets* guide extraction across the two scopes (Figure 2), preventing omissions and supporting stratified evaluation. Multiple probes can address different aspects of the same record.

Construction. We sampled source pools from three platforms in 2026, selected likely interpretation challenges, and retained 2,200 records per platform. Each record underwent privacy-preserving rewriting before a blind extractor proposed probes. Search-augmented auditing checks source support and reference correctness; disputed cases receive further adjudication or targeted human review, and semantic repairs are re-audited. Five closed-book Gemma-4-31B-it attempts determine difficulty: at least four fully correct answers yield Easy, otherwise Hard. Appendix A details the construction and repair pipeline.

3.3 Statistics and Distribution

SITU contains 89,199 probes from 6,600 records, including 9,522 Hard probes (Table 1). Expression-centered probes account for 34.9% overall and 43.7% of Hard. Target proportions reflect the retained data, without per-target quotas; platform and language are coupled in this release. Appendix A.2 reports the complete target distribution, and Appendix A.4 provides 54 bilingual examples.

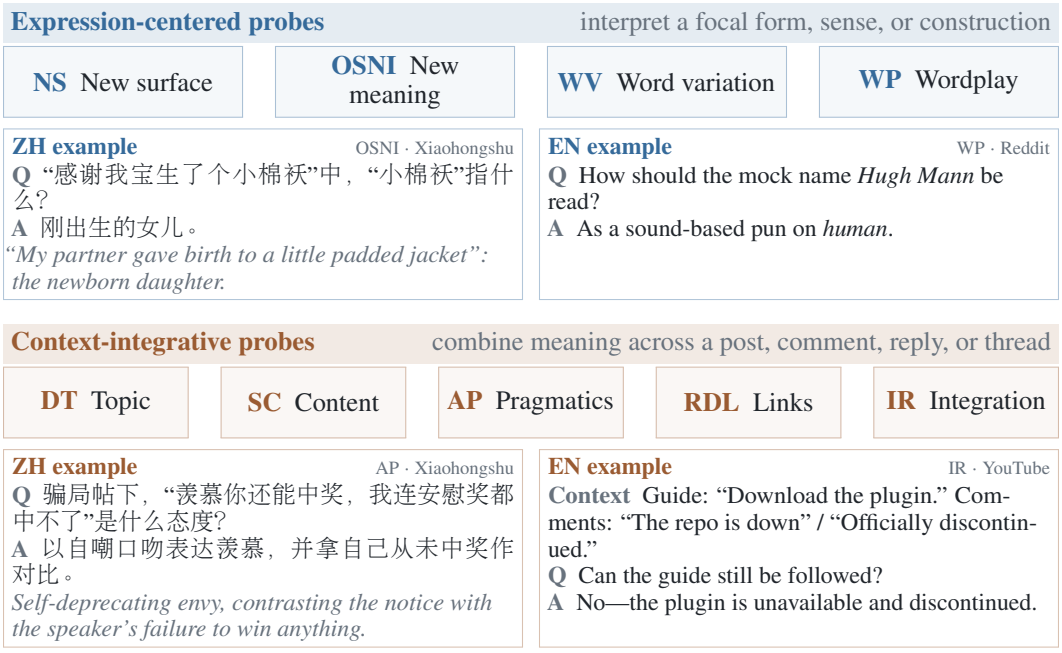

Figure 2: Nine SITU probe targets with four illustrations shortened from benchmark records. Gray text translates Chinese.

Table 1: Composition of SITU by platform. Difficulty entries are probe counts; the two probe scopes are reported as within-platform percentages.

Platform	Language	Records	Difficulty		Probe scope	
			Easy	Hard	Expr. (%)	Ctx. (%)
Xiaohongshu	zh	2,200	26,860	4,336	44.1	55.9
Reddit	en	2,200	28,313	2,845	33.5	66.5
YouTube	en	2,200	24,504	2,341	25.9	74.1
Total	zh/en	**6,600**	**79,677**	**9,522**	**34.9**	**65.1**

4 Main Evaluation

Protocol. We compare twelve models on SITU-Hard using fixed platform/language-specific prompts and thinking disabled where available (Gemini uses low thinking). A reference-conditioned judge assigns 0 for an incorrect answer, 1 for partial correctness, and 2 for a correct answer, checking essential meaning and unsupported claims. Among six judges calibrated on 500 frozen candidate answers against Gemini-3.1-Pro, Gemma-4-31B-it gives the highest agreement (quadratic-weighted $\kappa = 0.837$; 82.2% exact). In a blinded, model-triaged author audit, 63 of 76 reference packages are usable as written or after only non-core edits; all 6 stable-consensus audit packages are usable. On the 126 candidate answers attached to usable references, Gemma reaches a quadratic-weighted κ of 0.642 against the author, with no $0 \leftrightarrow 2$ reversals. Appendix B gives the protocols, calibration, and author-validation results; Appendix E shows prompt examples.

Results. Table 2 shows substantial remaining difficulty. Gemini-3.8-Flash leads the Chinese block at 1.52/2, while Claude Opus 5 leads English at 1.67/2. New surfaces, word variation, and reference/discourse links remain challenging; semantic content and attitude/pragmatics are generally easier. Size alone does not determine performance: Qwen3.5-27B trails Gemma-4-E4B-it on Chinese semantic content but substantially leads it on English reference/discourse links. Errors include reversing item allocation in compact Chinese fandom notation and attaching an English teasing reply to the wrong participant (Appendix Figure 7). The Hard filter is conditioned on Gemma; full-suite

Table 2: Main evaluation on SITU-Hard by probe target. Scores range from 0 to 2; All is the probe-micro mean within each language block.

Model	Lang.	Expression-centered ↑				Context-integrative ↑					All↑
		NS	OSNI	WV	WP	DT	SC	AP	RDL	IR	
Closed-source models											
Claude Opus 5	zh	1.09	**1.55**	1.03	0.96	1.45	**1.91**	1.61	1.19	**1.44**	1.47
	en	1.28	**1.56**	**1.50**	**1.39**	**1.84**	**1.96**	**1.85**	**1.42**	**1.44**	**1.67**
GPT-5.6-Sol	zh	1.13	1.53	1.08	1.04	1.48	1.82	1.55	1.03	1.23	1.44
	en	**1.32**	1.49	**1.50**	1.13	1.49	1.79	1.79	1.26	1.27	1.57
Gemini-3.8-Flash (low)	zh	**1.30**	1.51	**1.24**	1.04	1.52	1.88	**1.62**	**1.32**	**1.44**	**1.52**
	en	1.26	1.43	1.49	1.21	1.69	1.88	1.80	1.27	1.26	1.59
Open-weight models (100B+)											
Kimi-K3 (nothink)	zh	1.21	1.42	1.05	1.00	1.43	1.85	1.56	1.10	1.37	1.43
	en	1.21	1.38	1.35	1.12	1.71	1.91	1.75	1.19	1.27	1.55
DeepSeek V4 Pro (nothink)	zh	1.18	1.43	1.11	**1.15**	**1.55**	1.74	1.53	1.14	1.03	1.41
	en	1.19	1.41	1.33	1.15	1.47	1.83	1.73	1.23	1.13	1.51
GLM-5.2 (nothink)	zh	1.01	1.41	0.91	0.85	1.43	1.64	1.49	0.93	1.16	1.32
	en	1.01	1.20	1.23	1.00	1.47	1.74	1.76	1.17	1.09	1.43
Qwen3.5-397B-A17B (nothink)	zh	1.07	1.37	0.99	0.74	1.26	1.12	1.28	0.85	1.19	1.19
	en	0.81	1.16	0.98	0.87	1.50	1.63	1.65	1.08	1.26	1.33
Open-weight models (27–31B)											
Gemma-4-31B-it*	zh	0.72	0.90	0.63	0.93	1.04	1.23	1.16	0.57	0.87	0.94
	en	0.84	0.94	1.02	1.06	1.20	1.52	1.62	1.04	0.94	1.25
Qwen3.5-27B	zh	0.92	1.30	0.79	0.81	1.15	1.28	1.33	0.77	1.17	1.16
	en	0.68	1.07	0.89	0.72	1.39	1.72	1.63	0.99	1.22	1.29
Open-weight models (4–9B)											
Qwen3.5-9B	zh	0.69	1.02	0.48	0.70	1.00	1.07	1.05	0.56	0.86	0.89
	en	0.53	0.80	0.57	0.42	1.38	1.50	1.32	0.75	1.10	1.06
Gemma-4-E4B-it	zh	0.62	1.33	0.32	0.48	1.40	1.56	1.36	0.74	1.12	1.12
	en	0.69	0.90	0.75	0.46	1.38	1.49	1.49	0.81	1.20	1.16
Qwen3.5-4B	zh	0.55	0.93	0.24	0.56	0.86	1.11	0.84	0.31	0.98	0.78
	en	0.42	0.68	0.42	0.27	1.14	1.46	1.11	0.65	0.99	0.93

* Gemma-4-31B-it was used in the difficulty filter. The retained suite is therefore adversarially selected for this model.

results and ranking sensitivity appear in Appendix B.4. Alternative aggregation schemes largely preserve Hard rankings ($\rho \geq 0.993$).

5 Improving Social UGC Understanding

5.1 Intervention-oriented Diagnostic Aspects

We regroup Hard probes by the bottleneck an intervention might address. For expression-centered probes, *mechanism-transferable expressions* (E-M) permit recovery through reusable patterns such as phonetic repair or word formation; *item-knowledge-dependent expressions* (E-K) require a particular sense, entity, or community convention. A 0–5 mechanism-transferability rating yields complementary soft weights, so each expression probe can contribute to both summaries. For context-integrative probes, a masking-and-minimal-gloss test assigns a binary label: *expression-bottlenecked contexts* (C-E) require a critical expression meaning, whereas *integration-dependent contexts* (C-I) primarily require combining contextual evidence.

The mechanism rating asks how much meaning a solver could recover without having seen the exact item–meaning mapping. Annotators specify a reusable rule, its prediction for new surfaces, and the meaning left unresolved. For context probes, C-E requires both a loss when the critical expression is masked and recovery when only its meaning is restored. The minimal gloss supplies no discourse

Table 3: Gemma interventions on SITU-Hard (mean score /2 ↑). Shading shows change from Foundation.

Intervention	Chinese / Xiaohongshu					English / Reddit + YouTube				
	E-M	E-K	C-E	C-I	All	E-M	E-K	C-E	C-I	All
Foundation (direct)	0.85	0.76	1.06	1.20	0.94	1.07	0.88	1.30	1.47	1.25
+ Two-round RAG	1.30	1.26	1.35	1.32	1.30	1.36	1.32	1.41	1.45	1.40
+ Agentic RAG	1.35	1.28	1.38	1.29	1.32	1.39	1.35	1.43	1.47	1.43
+ Expression oracle[a]	–	–	1.48	1.32	–	–	–	1.39	1.45	–
+ bilingual SUU SFT	1.30	1.12	1.33	1.52	1.28	1.25	1.02	1.45	1.68	1.42
+ Two-round RAG	1.47	1.36	1.40	1.55	1.43	1.42	1.38	1.55	1.65	1.54
+ monolingual SUU SFT[b]	1.25	1.10	1.36	1.52	1.28	1.26	1.02	1.43	1.65	1.41

[a] The oracle supplies expression reference answers from the same record; only context probes are evaluated.
[b] Language-matched Chinese-only or English-only SFT.

relation or final conclusion, separating missing expression knowledge from the integration needed to use it.

Blinded human validation gives $\kappa = 0.444$ for context labels and weighted $\kappa = 0.603$ for mechanism ratings. Appendix C.1 gives the annotation protocols; the complete validation results remain in Appendix B.3.

5.2 Retrieval and Supervised Fine-Tuning

Gemma-4-31B-it is our main backbone, with Qwen3.5-9B/27B replications. *Two-round RAG* plans a web search and then answers with the retrieved snippets; *Agentic RAG* can search iteratively. The *expression oracle* supplies reference answers to expression probes from the same record before a context question, testing the benefit of available expression knowledge.

For SUU SFT, we mine training probes with the benchmark's selection and extraction stages and lighter quality control, retaining challenges unresolved by each backbone. The bilingual Gemma corpus contains about 206k demonstrations (116k Chinese and 89k English), trained for two epochs. Source records are disjoint from SITU. Language-matched monolingual arms provide controls; SFT+RAG applies the same retrieval procedure after training. Appendix C gives optimization, oracle, and retrieval controls.

Table 3 shows complementary effects. Retrieval helps E-K most: Two-round RAG gains 0.50 in Chinese and 0.43 in English, but changes English C-I little. SFT improves all four aspects, with stronger contextual gains and smaller E-K gains than retrieval. Their combination achieves the best aggregate in both language blocks (1.43 and 1.54), also observed in both Qwen replications. The oracle helps Chinese C-E substantially more than C-I (+0.42 versus +0.12), with a smaller English C-E gain (+0.09). Full-suite SFT also improves all three backbones (Gemma: 1.8423 to 1.8814; Appendix B.4), confirming that gains persist beyond the Hard subset.

Finding and using evidence. Retrieval gains depend on finding the needed evidence. A blind labeler, given snippets and the reference but no candidate answer, identifies answer-critical hits. Figure 3 separates hits, unsuccessful searches, and decisions not to search. On Chinese probes, hits cover 58.3% of E-M and 59.6% of E-K weighted mass, but only 10.5% of C-I. Within these outcome groups, E-M/E-K hits yield gains of +0.66/+0.75, versus +0.12–0.18 after searches that miss. For C-E, a hit brings a +0.47 gain, connecting retrieved expression knowledge to contextual interpretation.

Two controls separate evidence from the answering format and the model using it. Applying the RAG answer prompt with no evidence gains only 0.029 overall, versus 0.248 with retrieval (Appendix Table 14). A planner–answerer crossover then holds cached snippets fixed: on Qwen-9B's packets, Gemma scores 1.397 and Qwen-9B 1.042; Qwen-9B improves to 1.153 with Gemma's packets. Thus, useful retrieval depends on both evidence selection and the answerer's ability to use the resulting context (Appendix C.4).

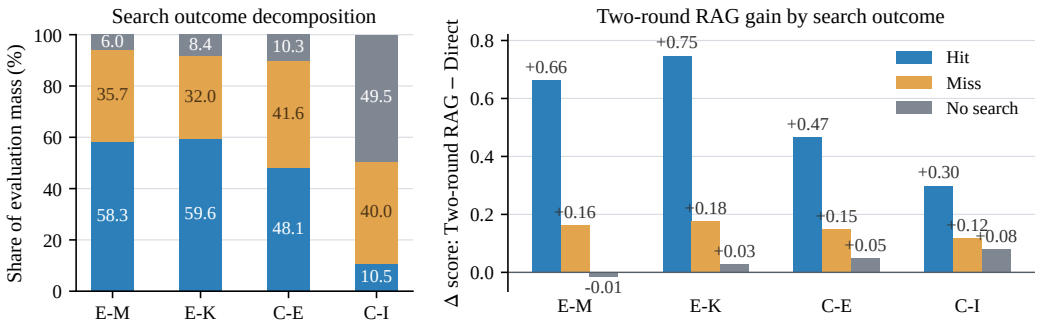

Figure 3: Chinese retrieval outcomes and within-group gains over direct inference. Hit denotes answer-critical evidence; Miss denotes search without a hit. E-M/E-K use complementary soft weights.

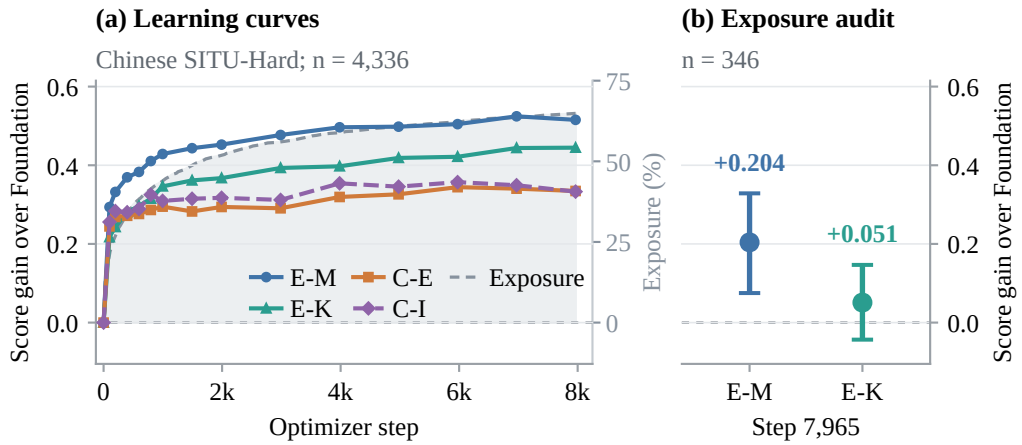

Figure 4: Chinese SFT over two runs. (a) Gains on 4,336 probes and targeted-expression exposure (shading; right axis). (b) Final gains on 346 expression probes without detected same-item exposure, with paired 95% source-record bootstrap intervals. E-M/E-K use complementary soft weights.

5.3 Learning with More SUU Supervision

We train two Chinese-only runs on the same approximately 510k-demonstration Xiaohongshu stream with different shuffle orders, evaluating 14 checkpoints from step 100 to 7,965 on all 4,336 Chinese SITU-Hard probes. Figure 4 averages probe-aligned scores across the runs. All four aspects improve rapidly; expression gains continue after context gains begin to flatten. Final gains are +0.52 for E-M, +0.44 for E-K, and +0.33 for each context aspect.

Exposure rises alongside performance: among the 2,325 Chinese expression probes, the share whose focus previously appeared as a targeted training focus grows from 21.2% at step 100 to 64.9% at the final step. Here, exposure means an exact normalized match to the expression targeted by a training probe. Hit/miss groups in the supplementary learning curves are fixed using the complete stream, so their composition stays constant across checkpoints. Incidental mentions elsewhere in a training record are examined separately below.

5.4 Generalization beyond Observed Expressions

A targeted-focus miss can still occur elsewhere in the training material. Scanning all training inputs and responses separates 1,508 targeted-focus matches, 434 incidental-only matches, and 383 literally absent probes. Table 4 shows how the learning gains change across these groups: both aspects

improve with targeted or incidental exposure, whereas the absent-surface group retains a larger E-M point estimate.

Table 4: Chinese expression gains by full-stream exposure.

Exposure group	Probes	E-M Δ ↑	E-K Δ ↑
Targeted focus	1,508	+0.578	+0.543
Incidental only	434	+0.622	+0.387
Literal surface absent	383	+0.222	+0.059
No detected same-item exposure	346	+0.204	+0.051

Final gains over Foundation, averaged over two runs. The last row is a subset of the preceding row; E-M/E-K use complementary soft weights.

Full-stream dense and character-based retrieval, followed by Gemini 3.1 Pro adjudication, leaves 346 probes without detected same-item exposure. E-M gains +0.204 (95% CI $[+0.075, +0.329]$); the paired E-M–E-K gain difference is +0.153 ($[+0.038, +0.265]$). This pattern is consistent with improved use of reusable expression mechanisms. Further excluding component knowledge leaves 157 probes, with an E-M gain of +0.075 ($[-0.117, +0.279]$). Appendix C.7 details the audit, controlled retrieval check, and sensitivity analyses; Appendix C.6 gives full hit/miss curves.

Two cases illustrate improved interpretations. In the Chinese scaling experiment, 修仙本 *(cultivation-themed jubensha scenario)* is absent from the training inputs and responses, but demonstrations explain its components 修仙 *(cultivation fantasy)* and 本 *(jubensha scenario)*. Foundation treats it as a scenario titled 修仙 *(Cultivation)* (score 0); the final checkpoints combine these meanings into a cultivation-themed *jubensha* scenario (two-run mean: 1.0). In the separate bilingual SFT experiment, *Six arm fire* is also literally absent from training. Foundation declares it unrecoverable (0), while SFT repairs it to *six-alarm fire* through the phonetic reduction *alarm→arm* (2).

Overall, retrieval and SFT address complementary needs: accessible item knowledge and improved interpretation. External variation benchmarks show more selective gains, including improved post-task HomoP-CN accuracy on all three backbones (Appendix D.7).

6 Transferring SUU to Downstream Applications

We test whether SUU supervision improves task learning by comparing Foundation→Task with SUU→Task under matched conditions; translation also tests mixed SUU+Task training. Appendix D gives full protocols, metrics, uncertainty, and Qwen replications.

6.1 Social-media translation

Using the same approximately 16.8k multilingual examples, we evaluate Xiaohongshu-sourced RedTrans-Bench and CSM-MTBench, and Reddit-sourced WMT22 Social, RoCS-MT v2, and MTNT 1.1. XCOMET-XL is primary; Appendix Tables 18–19 report auxiliary metrics.

Table 5: Gemma translation results: direction-macro XCOMET-XL (↑).

Training arm	Xiaohongshu-sourced				Reddit-sourced		
	RedTrans		CSM-MTBench		WMT22	RoCS-MT	MTNT
	zh→en	en→zh	Fun	Snippets	Social	v2	1.1
Foundation	.822	.794	.838	.784	.900	.846	.864
SUU SFT	**.834**	**.810**	**.841**	.785	**.904**	**.853**	**.871**
Foundation→T	.823	.798	.837	.783	.897	.845	.862
SUU→T	.829	.809	.839	**.788**	.902	.851	.866
SUU+T	.829	.808	.838	.784	.903	.852	.867

T: translation-task SFT; →: sequential; +: mixed. Bold: column best.

Sequential SUU→T exceeds Foundation→T on all seven XCOMET aggregates in Table 5 (all paired intervals exclude zero), and mixed training follows the same direction. SUU-only leads six

columns; sequential training leads Social Snippets. Qwen transfer is weaker: Qwen-9B improves four aggregates, whereas Qwen-27B is effectively tied (Appendix D.4).

6.2 TOXICITY DETECTION

HED-COLD tests whether clean-only task supervision generalizes to homophone-perturbed Chinese; SeTox contrasts evolving toxic terms with ordinary safety examples; and CounterContext removes the parent comment from English hate/neutral/counterspeech classification to test context dependence (Wu et al., 2025; Cui et al., 2026; Yu et al., 2022). Figure 5 tracks the SUU advantage before and after matched task training.

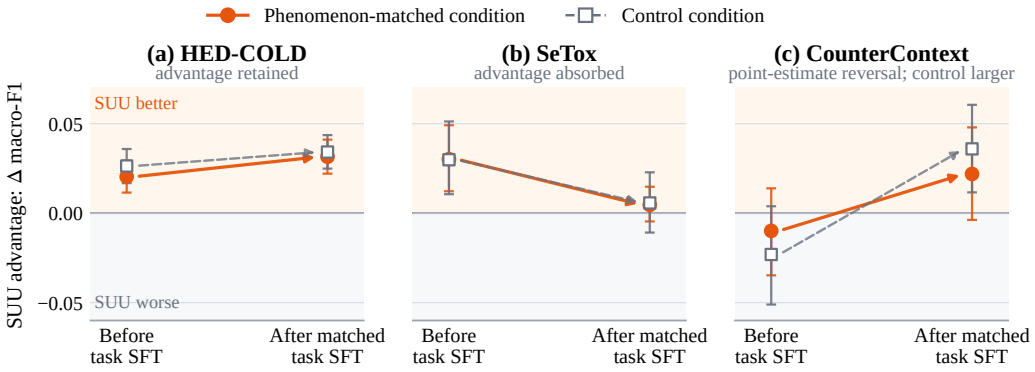

Figure 5: Gemma SUU transfer advantage (Δ macro-F1) before and after matched task SFT, with paired 95% bootstrap intervals. HED task training uses clean inputs; absolute scores appear in Appendix Table 22.

Transfer varies by task. HED-COLD retains the advantage after clean-only task training (perturbed-test macro-F1: 0.8338→0.8653), while perturbation training largely closes the gap. SeTox improves directly, but near-saturated task learning absorbs most of the gain. CounterContext improves reliably only post-task without parent context, giving no clear evidence of better context use. Qwen transfer also varies by model size (Appendix D.4). Thus benefits depend on the backbone and what task training already teaches.

7 CONCLUSION

We introduced SITU, a bilingual open-ended benchmark unifying expression-centered and context-integrative social UGC understanding. Twelve-model evaluation reveals persistent gaps; retrieval and SUU supervision provide complementary gains, and their combination performs best. Chinese exposure auditing finds positive E-M gains beyond detected same-item knowledge, consistent with reusable-mechanism learning. Matched downstream experiments show benefits for translation and homophone-robust moderation, with task- and backbone-dependent boundaries. SITU thus supports studying both understanding and its downstream value.

AI USE STATEMENT

Generative AI tools assisted literature organization, benchmark construction and auditing, code development, experiment monitoring, and drafting. The authors check references and empirical claims against source papers or persisted artifacts, test generated code, and take responsibility for the final text, data, code, and claims.

ETHICS STATEMENT

Social-media content may contain offensive language, sensitive topics, and personal references. We retain only context needed for interpretation and consistently replace account identifiers and non-public personal information. Annotators receive content warnings. Release eligibility is checked against platform-specific licenses and access conditions: permissible rewritten records and reference packages will be released, and restricted content will be excluded or represented by redistributable derived annotations. Original platform identifiers are not distributed.

REPRODUCIBILITY STATEMENT

The appendix specifies construction, annotation, scoring, inference, and matched transfer protocols. We will release permissible benchmark material, frozen manifests, evaluation code, aggregate and per-probe results, data hashes, and model identifiers or API snapshots. Training-data quantities and optimization settings are documented for each experiment.

REFERENCES

Dimosthenis Antypas, Asahi Ushio, Francesco Barbieri, Leonardo Neves, Kiamehr Rezaee, Luis Espinosa-Anke, Jiaxin Pei, and Jose Camacho-Collados. SuperTweetEval: A Challenging, Unified and Heterogeneous Benchmark for Social Media NLP Research. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 12590–12607, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.838.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1644–1650, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.148.

Vasilisa Bashlovkina, https://orcid.org/0009-0005-0235-304X, View Profile, Riley Matthews, https://orcid.org/0009-0000-7634-7076, View Profile, Zhaobin Kuang, https://orcid.org/0000-0001-9295-4669, View Profile, Simon Baumgartner, https://orcid.org/0009-0005-8746-5787, View Profile, Michael Bendersky, https://orcid.org/0000-0002-2941-6240, and View Profile. SMILE: Evaluation and Domain Adaptation for Social Media Language Understanding. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ACM Conferences, pp. 3737–3749, August 2023. ISBN 979-8-4007-0103-0. doi: 10.1145/3580305.3599907.

Rachel Bawden and Benoît Sagot. RoCS-MT v2 at WMT 2025: Robust Challenge Set for Machine Translation. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Tenth Conference on Machine Translation*, pp. 834–849, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-341-8. doi: 10.18653/v1/2025.wmt-1.58.

Shiyao Cui, QingLin Zhang, Di Wang, Yida Lu, Zhexin Zhang, Jinhua Gao, Jinglin Yang, Min He, Han Qiu, and Minlie Huang. New Terms, New Toxicity: Consensus-based Chinese Neologism Toxicity Detection via Search-Augmented LLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 34683–34701, 2026. doi: 10.18653/v1/2026.acl-long.1602.

Haotan Guo, Jianfei He, Jiayuan Ma, Hongbin Na, Zimu Wang, Haiyang Zhang, Qi Chen, Wei Wang, Zijing Shi, Tao Shen, and Ling Chen. Lost in Pronunciation: Detecting Chinese Offensive

Language Disguised by Phonetic Cloaking Replacement. In Saloni Potdar, Lina Rojas-Barahona, and Sebastien Montella (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 2538–2550, Suzhou (China), November 2025a. Association for Computational Linguistics. ISBN 979-8-89176-333-3. doi: 10.18653/v1/2025.emnlp-industry.172.

Hongcheng Guo, Yue Wang, Shaosheng Cao, Fei Zhao, Boyang Wang, Lei Li, Liang Chen, Xinze Lyu, Zhe Xu, Yao Hu, and Zhoujun Li. SNS-Bench: Defining, building, and assessing capabilities of large language models in social networking services, 2025b. URL `https://proceedings.mlr.press/v267/guo25o.html`.

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, April 2025c.

Ruiqi He, Yushu He, Longju Bai, Jiarui Liu, Zhenjie Sun, Zenghao Tang, He Wang, Hanchen Xia, Rada Mihalcea, and Naihao Deng. Chumor 2.0: Towards Benchmarking Chinese Humor Understanding, December 2024.

Shiwei Hong, Junjie Ma, Emma Jiren Wang, Ethan Z. Rong, Siying Hu, Haichang Li, Ziying Wang, and Zhicong Lu. You Really Didn't Get That? Benchmarking Social Pragmatic Inference for Indirect and Playful Chinese Online Comments, September 2026.

Chen Huang, Junkai Luo, Xinzuo Wang, Wenqiang Lei, and Jiancheng Lv. Can Large Language Models Understand Internet Buzzwords Through User-Generated Content. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12916–12941, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.632.

Michael T. Kane. Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1):1–73, 2013. doi: 10.1111/jedm.12000.

Yunlong Liang, Fandong Meng, Jiaan Wang, and Jie Zhou. SlangDIT: Benchmarking LLMs in Interpretative Slang Translation, May 2025.

Dianqing Lin, Tian Lan, Jiali Zhu, Jiang Li, Wei Chen, Xu Liu, Aruukhan, Xiangdong Su, Hongxu Hou, and Guanglai Gao. Exploring the Capability Boundaries of LLMs in Mastering of Chinese Chouxiang Language. arXiv, April 2026. doi: 10.48550/arXiv.2604.15841.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. TimeLMs: Diachronic Language Models from Twitter. In Valerio Basile, Zornitsa Kozareva, and Sanja Stajner (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 251–260, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-demo.25.

Paul Michel and Graham Neubig. MTNT: A Testbed for Machine Translation of Noisy Text. In *Conference on Empirical Methods in Natural Language Processing*, pp. 543–553, Brussels, Belgium, 2018. ACL. doi: 10.18653/v1/D18-1050.

Robert J. Mislevy, Russell G. Almond, and Janice F. Lukas. A brief introduction to evidence-centered design. Technical Report 2003-1, Educational Testing Service, 2003.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. BERTweet: A pre-trained language model for english tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 9–14. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-demos.2.

Ali Ezzat Shahroor, Mohamed Bayan Kmainasi, Abul Hasnat, Dimitar Dimitrov, Giovanni Da San Martino, Preslav Nakov, and Firoj Alam. MemeLens: Multilingual Multitask VLMs for Memes, May 2026.

Rob van der Goot, Alan Ramponi, Arkaitz Zubiaga, Barbara Plank, Benjamin Muller, Iñaki San Vicente Roncal, Nikola Ljubešić, Özlem Çetinoğlu, Rahmad Mahendra, Talha Çolakoğlu, Timothy Baldwin, Tommaso Caselli, and Wladimir Sidorenko. MultiLexNorm: A Shared Task on Multilingual Lexical Normalization. In Wei Xu, Alan Ritter, Tim Baldwin, and Afshin Rahimi (eds.), *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pp. 493–509, Online, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.wnut-1.55.

Yifan Wang, Junyu Lu, Qifan Wang, Shun Zhang, Chaozhuo Li, Jiahao Liu, Zhijun Cao, Lingbin Bu, and Fanliang Bu. CIBuzzBench: A benchmark for cross-lingual understanding of chinese internet buzzwords, September 2026. URL https://arxiv.org/abs/2609.21722.

Junqi Wu, Shujie Ji, Kang Zhong, Huiling Peng, Zhendongxiao, Xiongding Liu, and Wu Wei. Enhancing Chinese Offensive Language Detection with Homophonic Perturbation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 22660–22675, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1154.

Yubo Xie, Chenkai Wang, Zongyang Ma, and Fahui Miao. Are Large Language Models Chronically Online Surfers? A Dataset for Chinese Internet Meme Explanation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 17062–17083, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.863.

Wang Xingren, Hongde Liu, Liushanhong, Feiyang Meng, Chenyuan He, Senbin Zhu, Zechen Li, and Yuxiang Jia. TaCL-CoMoE: Task-adaptive contrastive learning with cooperative mixture of experts for multi-task social media analysis. *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, 2025. doi: 10.18653/v1/2025.ijcnlp-long.84. URL https://aclanthology.org/2025.ijcnlp-long.84/.

Xinchen Yu, Eduardo Blanco, and Lingzi Hong. Hate speech and counter speech detection: Conversational context does matter. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5918–5930, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.433. URL https://aclanthology.org/2022.naacl-main.433/.

Chiyu Zhang, Khai Doan, Qisheng Liao, and Muhammad Abdul-Mageed. The Skipped Beat: A Study of Sociopragmatic Understanding in LLMs for 64 Languages. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2630–2662, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.160.

Fei Zhao, Chonggang Lu, Haofu Qian, Fangcheng Shi, Zijie Meng, Jianzhao Huang, Xu Tang, Zheyong Xie, Zheyu Ye, Zhe Xu, Yao Hu, and Shaosheng Cao. RedOne 2.0: Rethinking Domain-specific LLM Post-Training in Social Networking Services, November 2025a.

Fei Zhao, Chonggang Lu, Yue Wang, Zheyong Xie, Ziyan Liu, Haofu Qian, JianZhao Huang, Fangcheng Shi, Zijie Meng, Hongcheng Guo, Mingqian He, Xinze Lyu, Yiming Lu, Ziyang Xiang, Zheyu Ye, Chengqiang Lu, Zhe Xu, Yi Wu, Yao Hu, Yan Gao, Jun Fan, Xiaolong Jiang, Weiting Liu, Boyang Wang, and Shaosheng Cao. RedOne: Revealing Domain-specific LLM Post-Training in Social Networking Services, October 2025b.

Kaiyan Zhao, Zhongtao Miao, Zheyong Xie, Shaosheng Cao, and Yoshimasa Tsuruoka. CNeo-Bench: Diagnosing Large Language Models on Chinese Neologisms, August 2026a.

Kaiyan Zhao, Zheyong Xie, Zhongtao Miao, Xinze Lyu, Yao Hu, and Shaosheng Cao. Benchmarking Machine Translation on Chinese Social Media Texts, January 2026b.

A SITU SOURCES, CONSTRUCTION, AND COVERAGE

A.1 CONSTRUCTION AND AUDIT PIPELINE

Record construction and privacy. The source pools were sampled in 2026, and every benchmark input contains text only. Xiaohongshu records include a title, post content, and a bounded comment tree; Reddit records include a title, post body, and a bounded comment tree; YouTube records include a title, description, chapter metadata, and comments. Images, video frames, OCR, audio, and transcripts are not included. During source selection, records whose interpretation depended heavily on omitted modalities were filtered out. Every record retained in SITU is a privacy-preserving rewrite rather than an unchanged raw record. Included parent–child edges and visible order are preserved, and truncation boundaries are explicit. Sensitive identifiers are replaced consistently within a record; dummy tokens cannot become probe targets. Rewritten records and reference packages are released where redistribution is permitted; restricted content is excluded or represented by permissible derived annotations. Original platform records and identifiers are not distributed.

Selection, rewriting, and blind extraction. A dual-head selector retained 14.7%, 48.0%, and 43.1% of the Xiaohongshu, Reddit, and YouTube source pools, respectively, by detecting likely expression-centered or context-integrative challenges. Source-side extraction supplies soft anchors to a rewrite model. Rewriting may alter nonessential details and structure but should preserve difficult mechanisms without resolving ambiguity or inventing new test-worthy names. A blind extractor then sees the rewritten record, not the source candidates, and proposes probes in the nine target slots.

Search-augmented complete audit. After blind extraction, a complete auditor checks both coverage and validity: whether important probe targets were missed, whether each proposed probe is atomic and answerable, whether its answer is licensed by the source, whether the answer and explanation cover the essential answer conditions, and whether the probe is literal, duplicated, or otherwise unsuitable. The auditor may issue language- and platform-appropriate web searches to verify aliases, emerging senses, titles, cultural concepts, and entities. It can add a missed high-value probe, repair an initial reference package, delete an invalid proposal, or attach an uncertainty flag. Search evidence is retained as construction provenance; it is not supplied to evaluated models.

Provisional probe audit and uncertainty routing. Each surviving probe is answered five times by Gemma-4-31B-it under a closed-book contract, with thinking disabled, temperature 0.7, top-p 0.95, and a distinct deterministic seed per attempt. The same model serves as the reference-aware answer judge with thinking disabled and temperature-zero decoding, labeling each attempt correct, partial, or wrong. A probe is labeled Easy if at least four attempts are fully correct, and Hard otherwise. This first pass is provisional. Schema failures, conflicting judgments, questionable answerability, or a possible reference problem produce a separate `uncertain` routing state rather than an additional difficulty level or an automatic assignment to Hard.

Strong-model consistency triage. Provisional Hard probes are answered by stronger, heterogeneous models. These answers are used to find problems, not to define gold by consensus. We compare candidate answers with one another and with the question, reference answer, supporting explanation, and available evidence. Repeated agreement on an omitted qualification, conflicting entity, alternative referent, or narrower uncertainty boundary raises a resolution trigger; an isolated model disagreement does not by itself alter the reference.

Search-assisted resolution and targeted human review. Triggered probes are resolved against the source record first and web evidence when the disputed content depends on an external fact. Resolution yields one of three explicit outcomes: retain the current contract, repair the question/reference package, or drop the probe. Majority model agreement is evidence but never substitutes for source or external verification. A small residual set—cases with conflicting evidence, unresolved long-tail identity, or consequential ambiguity—is reviewed by a human. Construction-time strong-model and search evidence remains separate from later closed-book candidate scoring.

Post-repair probe audit. Every semantic repair invalidates stale difficulty evidence. If only non-semantic wording or judge instructions change, existing candidates may be retained but are re-judged.

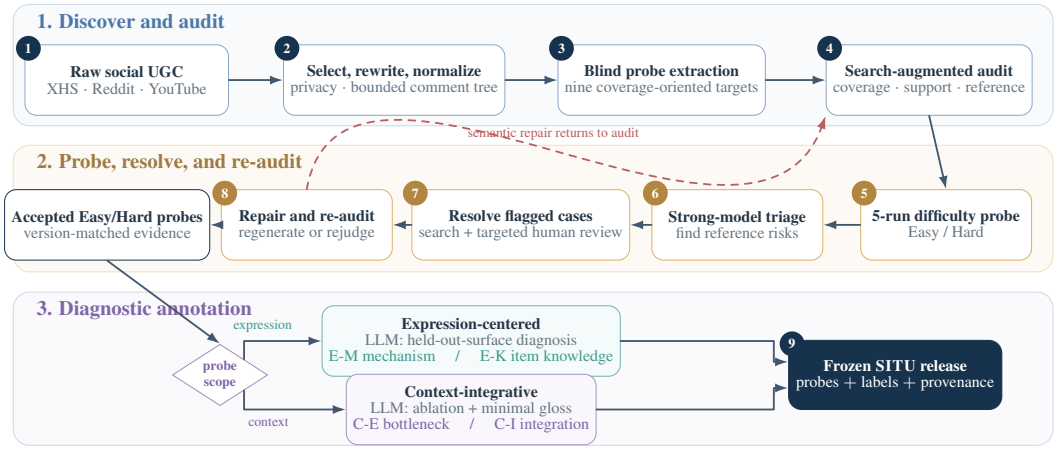

Figure 6: The SITU construction and diagnostic-annotation pipeline. Probe targets guide candidate discovery. Complete auditing, empirical difficulty probing, heterogeneous strong-model triage, search-assisted resolution, and targeted human review form a repair loop: every semantic repair invalidates stale difficulty evidence and is re-audited. Only then are expression-centered and context-integrative probes annotated under their separate counterfactual protocols and assembled into the frozen release.

If the question target, core answer, required qualification, or justified uncertainty boundary changes, candidates are regenerated and re-judged. The resulting second probe audit determines the final difficulty label. Unmodified probes reuse their version-matched audit; no unresolved `uncertain` state is silently published as a difficulty level.

Deterministic assembly. Final checks enforce unique identifiers, record–probe links, complete fields, privacy constraints, deduplication, mutually exclusive accept/drop states, and version-consistent links among the frozen question, reference package, candidates, judgments, and final difficulty. Train, development, and test material is split at the record level.

A.2 Probe-Target Distribution by Platform and Difficulty

Table 6 reports the complete distribution rather than only SITU-Hard. Counts are separated by platform because platform and language are coupled in the current release.

Table 6: Probe-target composition of the full SITU suite by platform and empirical difficulty. Each platform block reports counts; Share is the target percentage within that platform.

Difficulty	Probe targets									All
	Expression-centered probes				Context-integrative probes					
	NS	OSNI	WV	WP	DT	SC	AP	RDL	IR	
Xiaohongshu (Chinese)										
Easy	2,441	7,369	1,549	68	2,079	1,418	11,232	405	299	26,860
Hard	589	1,127	582	27	121	783	858	155	94	4,336
Total	**3,030**	**8,496**	**2,131**	**95**	**2,200**	**2,201**	**12,090**	**560**	**393**	**31,196**
Share (%)	9.7	27.2	6.8	0.3	7.1	7.1	38.8	1.8	1.3	100.0
Reddit (English)										
Easy	2,769	3,918	2,371	186	2,110	1,797	14,028	552	582	28,313
Hard	497	340	318	29	90	403	703	181	284	2,845
Total	**3,266**	**4,258**	**2,689**	**215**	**2,200**	**2,200**	**14,731**	**733**	**866**	**31,158**
Share (%)	10.5	13.7	8.6	0.7	7.1	7.1	47.3	2.4	2.8	100.0
YouTube (English)										
Easy	1,198	2,508	2,437	171	2,115	1,390	13,720	677	288	24,504
Hard	216	168	227	38	85	810	522	172	103	2,341
Total	**1,414**	**2,676**	**2,664**	**209**	**2,200**	**2,200**	**14,242**	**849**	**391**	**26,845**
Share (%)	5.3	10.0	9.9	0.8	8.2	8.2	53.1	3.2	1.5	100.0

Abbreviations. Probe-target abbreviations follow Table 8. Percentages may not sum to exactly 100 because of rounding.

A.3 Coverage Emphasis of Related Benchmarks

Table 7 compares the primary phenomenon coverage and output form of representative social-UGC benchmarks. The comparison highlights complementary emphases across benchmarks.

Table 7: Coverage emphasis and output form in representative social-UGC benchmarks.

Benchmark	Expression phenomena	Context-integrative phenomena	Multi-phenomenon	Output form
Application-oriented benchmark suites				
TweetEval (Barbieri et al., 2020)	Not a dedicated expression diagnostic	Affect, irony, stance, hate, and offense	Yes (suite)	Classification
SMILE (Bashlovkina et al., 2023)	Platform variation through domain adaptation	Affect, harmful language, and summarization across platforms	Yes (suite)	Classification and summarization
SPARROW (Zhang et al., 2023)	Source-dependent; not organized around expression recovery	Thirteen sociopragmatic task types across 64 languages	Yes (suite)	Classification
SuperTweet Eval (Antypas et al., 2023)	Semantic shift, entities, and noisy tweet text	Tweet QA, affect, intimacy, and semantic similarity	Yes (suite)	Labels, extraction, QA, and regression
SNS-Bench (Guo et al., 2025b)	Not an explicit expression target	MRC, relevance, taxonomy, NER, and platform operations	Yes (suite)	Classification, extraction, QA, and generation

Benchmark	Expression phenomena	Context-integrative phenomena	Multi-phenomenon	Output form
Interpretation-focused benchmarks				
CHEER (Huang et al., 2025)	Chinese Internet buzzword meaning induction	UGC-supported definitions; broader interaction is not central	No (focused)	Open definition
CHIME (Xie et al., 2025)	Chinese Internet phrasememes and formation	Meaning, origin, and appropriate use; thread integration is not central	Within domain	Multiple choice and open explanation
Chumor 2.0 (He et al., 2024)	Wordplay and culturally specific joke mechanisms	Whether a candidate explanation fully accounts for the joke	No (focused)	Binary explanation validation
MOUSE (Lin et al., 2026)	Homophonic, visual, and semantic *Chouxiang* forms	Meaning selection, communicative intention, and toxicity	Within domain	Selection, reconstruction, classification
CNeo-Bench (Zhao et al., 2026a)	Chinese neologisms, formation mechanisms, and source-form recovery	Contextualized meaning, but not general interaction-level integration	Within domain	Classification, selection, open generation
CIBuzzBench (Wang et al., 2026)	Chinese buzzword definitions and cross-lingual sense selection	Harmfulness is assessed; discourse integration is not central	Within domain	Definition, selection, classification
You Really Didn't Get That? (Hong et al., 2026)	Wordplay or allusion when it drives a difficult comment	Situated pragmatic interpretation from preceding social context	Within domain	Multiple-choice interpretation
MemeLens (Shahroor et al., 2026)	Not a dedicated textual-expression diagnostic	Image–text meaning, social labels, and rationale generation	Yes (suite)	Classification and generation
SITU	New senses, concepts and entities, word variation, and wordplay	Component meaning and pragmatics, reference/discourse links, and integrated reading	Yes	Open-ended situated meaning recovery

Note. Coverage denotes primary emphasis, not incidental phenomena. "Within domain" means multiple linked phenomena inside a focused family. This representative table is neither exhaustive nor a mapping of prior taxonomies onto SITU targets.

A.4 Probe Targets and Examples

Table 8 defines the complete target inventory and gives compact examples drawn from SITU records. The target is an extraction and analysis axis, not a predicted class: every benchmark question remains open-ended.

Table 8: The nine SITU probe targets, with three Chinese and three English illustrations per target. All examples are shortened from frozen benchmark records rather than constructed for the paper; gray text translates or glosses Chinese. Q and A denote the probe question and frozen reference answer.

Chinese illustrations	English illustrations
Expression-centered probes	
NS — New surface Newly encountered expression, concept, or entity	
1. Q: “召唤露营搭子”中的“搭子”指什么？ **A:** 一起参加某项活动的伙伴。 *In “calling for a camping dāzi,”* dāzi *means a companion for a shared activity.*	**1. Q:** What is *Labubu* in “the whole Labubu obsession”? **A:** A designer-toy character.
2. Q: “非官谷也可以带图来问”中的“官谷”指什么？ **A:** 官方周边或官方商品。 *In “non-official gu is also fine,”* guan-gu *means officially licensed merchandise.*	**2. Q:** What does *NSFW* mean in “definitely NSFW”? **A:** Not Safe For Work.
3. Q: “在茶水间摸鱼闲聊”中的“摸鱼”是什么意思？ **A:** 在工作时间偷懒或做与工作无关的事。 *“Fishing” here means slacking off during work.*	**3. Q:** What does *AMV* mean in “drop a sick AMV”? **A:** Anime Music Video.
OSNI — Old surface / new item Familiar form with a new situated sense	
1. Q: 买手店探店帖说“价格也很美丽，不过冲着这份市面上少有的设计感，最后还是没忍住乖乖掏了钱包”。“美丽”是什么意思？ **A:** 反讽地表示价格很高。 *The shopper praises the design and says the price is “beautiful,” but pays anyway: the frozen reference reads this as ironic praise of a high price.*	**1. Q:** What is a *cake day* in “Happy cake day”? **A:** The anniversary of a Reddit account’s creation.
2. Q: “感谢我宝生了个小棉袄”中的“小棉袄”指什么？ **A:** 刚出生的女儿。 *The “little padded jacket” is the speaker’s newborn daughter.*	**2. Q:** What does *robbery* mean in “8 bucks is actual robbery”? **A:** An exceptionally good bargain, not literal theft.
3. Q: “龙一豪生图”中的“生图”指什么？ **A:** 未经修图的原始照片。 *A “raw image” here is an unretouched photograph.*	**3. Q:** Who are the *kids* in “this group of kids” under a goat-sanctuary post? **A:** Young goats.
WV — Word variation Nonstandard form and its contribution	
1. Q: “我蹲蹲蹲”中的三次重复起什么作用？ **A:** 强化持续等待更新或消息的语气。 *Repeating “wait” three times intensifies the speaker’s continued anticipation.*	**1. Q:** What does the spelling *roooooough* contribute in “some roooooough days”? **A:** Letter elongation intensifies *rough*.
2. Q: “98r包邮出”中的“r”表示什么？ **A:** 人民币的货币单位“元”。 *The letter* r *substitutes for the RMB/yuan price unit.*	**2. Q:** What does *sue* mean in “Can you sue spirit gum to stick the ears on”? **A:** It is a typo for *use*.
3. Q: “明天再打十人本”中的“本”是什么形式？ **A:** “副本”的缩写，指游戏中的实例任务。 Ben *abbreviates the gaming term* fuben*, an instanced dungeon or raid.*	**3. Q:** In a Melbourne mushroom-foraging thread, what does *Woolies* refer to, and how is the form derived? **A:** It refers to Woolworths supermarket; *Woolies* is its colloquial shortened form.

Continued on the next page.

Table 8 continued from the previous page.

Chinese illustrations	English illustrations
WP — Wordplay Meaning created through sound, form, or multiple readings	
1. Q: “赞美味海鲜粥”如何构成双关? **A:** 它把“赞美药师”和“美味海鲜粥”拼接在一起。 *The phrase blends “praise the Medicine Buddha” with “delicious seafood congee.”*	**1. Q:** How should “The pricey is spicy” be read? **A:** The rhyme play conveys that the price is surprisingly high.
2. Q: 在“职业与本命英雄”帖中，玩家说选择达摩是因为喜欢“物理超度”。这个说法的笑点是什么? **A:** 它把游戏中的物理攻击与宗教意义上的“超度”并置，幽默地表示通过战斗把对手击败或“送走”。 *It juxtaposes physical attacks with the religious notion of delivering souls, humorously meaning to defeat an opponent by force.*	**2. Q:** How should the mock name *Hugh Mann* be read? **A:** As a sound-based pun on *human*.
3. Q: “烤面筋很面，不筋”如何利用词形? **A:** 把“面筋”拆成“面”和“筋”，抱怨口感粉面而不筋道。 *The food name is split into “floury” and “chewy” to complain that it has the former but not the latter.*	**3. Q:** How should “De brie was everywhere” be read after a dairy-plant explosion? **A:** *De brie* combines the cheese Brie with the sound of *debris*.
Context-integrative probes	
DT — Domain/topic Situation or activity organizing the record	
1. Q: 标题“退坑清谷，拒大刀”，正文列出徽章和挂件：帖子在做什么? **A:** 在二手出售动漫或游戏周边。 *A post listing badges and pendants while “leaving the fandom” is reselling ACG merchandise.*	**1. Q:** What domain and topic organize “Just fired up a new save for RE4 day”? **A:** Gaming; starting a new playthrough of *Resident Evil 4*.
2. Q: “想买洗地机，搜了一圈全是广”这篇帖子的领域和主题是什么? **A:** 家用电器；因不信任广告而征求真实的洗地机推荐。 *Home appliances; the author wants genuine floor-washer recommendations rather than ads.*	**2. Q:** What is the topic of “The Nike Cortez is Underrated”? **A:** Footwear/fashion; a review arguing that Nike Cortez sneakers are underrated.
3. Q: “去江滩出外景，蹲一个互勉模特”的领域和主题是什么? **A:** 摄影与社交；寻找互免约拍的模特。 *Photography/social coordination; the author seeks a model for a mutually unpaid photo shoot.*	**3. Q:** What does a video listing macaroni, chicken, cheese, and baking instructions present? **A:** Cooking; a quick oven-baked cheesy chicken macaroni recipe.

Continued on the next page.

Table 8 continued from the previous page.

Chinese illustrations	English illustrations
SC — Semantic content What one component asserts	
1. Q: 作者已有徽章和立牌，还想收什么? **A:** 流沙麻将和色纸，也接受带图询问非官方周边。 *The author still wants liquid-sand mahjong pieces and art cards, and will consider pictured unofficial items.*	**1. Q:** What field conditions and preferences constrain the author's choice of the Bridger 45? **A:** A nine-month Rockies survey requires a pack that can survive truck beds and rocks. The author prefers a top-loading technical design and accepts extra weight for durability.
2. Q: 隐藏款起酥帖中，多出的两盒可以怎样处理? 有意者需要提供什么信息? **A:** 可以单独购买，也可以交换其他口味；有意者需在评论中带价或带图。 *The two extra boxes may be sold separately or exchanged for another flavor; interested users should comment with a price or photo.*	**2. Q:** What ownership record supports calling the EV6 an almost flawless daily driver? **A:** It covered over 22,000 miles in 14 months without squeaks, rattles, or a single day in the shop.
3. Q: 作者为何买下这瓶"液体黄金"，原定何时开瓶，实际怎样? **A:** 作者以此庆祝持续两个月的项目落地和升职；原计划周末再喝，收到后却当天开了。 *The bottle celebrates a two-month project's completion and a promotion; it was meant for the weekend but was opened immediately.*	**3. Q:** How does the ASMR video simulate a clogged ear, and what safety boundary does the creator state? **A:** The introduction muffles the voice while keeping tool sounds clear. It is a non-medical simulation using an ear model, and viewers are warned not to insert objects into their ears.
AP — Attitude/pragmatics Stance, intent, irony, or implicature	
1. Q: 骗局帖下评论"羡慕你还能中奖，我连安慰奖都中不了"是什么态度? **A:** 以自嘲口吻表达羡慕，并拿自己从未中奖作对比。 *Under a scam post, the commenter voices self-deprecating envy by contrasting the notice with their own failure to win anything.*	**1. Q:** What attitude does "Thank you for saving me from myself" express in response to a deleted comment? **A:** Relief and gratitude toward whoever removed the preceding content.
2. Q: FIRE帖下"计划很完美，除了没钱全都有"是在直接赞同吗? 它表达了什么态度? **A:** 不是直接赞同；说话者以反讽和自嘲表达无力感，指出自己唯独缺少实现这套计划的关键资金。 *Rather than straightforward agreement, the comment uses irony and self-deprecation to stress that money—the crucial requirement—is precisely what the speaker lacks.*	**2. Q:** What stance does "I am looking respectfully" convey under attractive fan art? **A:** Playful attraction, phrased as a meme-based euphemism.
3. Q: 壁纸合集下的评论"求3"意在做什么? **A:** 向作者索要图集中的第三张壁纸。 *"Requesting 3" asks the author for the third wallpaper in the carousel.*	**3. Q:** What does "Aged like milk" do after later events contradict an earlier skeptical reply? **A:** It mocks the earlier prediction as having quickly proved wrong.

Continued on the next page.

Table 8 continued from the previous page.

Chinese illustrations	English illustrations
RDL — Reference/discourse link Referent or relation across components	
1. Q: 标签写“短剧不祭罗裙祭烽火”，评论说“我也好喜欢这部不祭”。“不祭”指什么？ **A:** 标签中提到的短剧《不祭罗裙祭烽火》。 *The shortened “Bu Ji” refers back to the tagged short drama.*	**1. Q:** “The mic has echo. Topic suggestion: lost cities.” A reply says, “Sounds good, but the echo is distracting.” What do its clauses answer? **A:** “Sounds good” answers the topic suggestion; “the echo” resumes the audio problem.
2. Q: 评论“好几个赛季没碰过这个模式了”中的“这个模式”指什么？ **A:** 原帖视频展示的躲猫猫模式。 *“This mode” refers to the hide-and-seek mode shown in the post.*	**2. Q:** In “Where exactly is it?” what does *it* refer to? **A:** The municipal climbing center proposed earlier in the thread.
3. Q: 转租帖下评论“几张”省略了什么？ **A:** “几张房间照片”。 *In a rental thread, “how many” elliptically requests several room photographs.*	**3. Q:** In comments on a Neon Echoes performance, who does *They* refer to? **A:** The members of the musical act Neon Echoes.
IR — Integrated reading Conclusion requiring multiple components	
1. Q: 结合猫帖正文和“他们仨像叠罗汉”这条评论，家里共有几只猫？ **A:** 至少三只。 *The post plus “the three of them piled together” establishes at least three cats.*	**1. Q:** A post calls an R34 in *Tokyo Drift* Dom Toretto’s. What do the comments establish? **A:** Dom appears at the end, but they do not establish that he drove that R34; the premise is unsupported.
2. Q: 结合原帖与评论，这批草帽的变形是个例吗？与去年同款相比如何？ **A:** 不是个例；多名用户报告了相同的塌顶、翘边问题，而去年购买的同款更为挺括，说明这批产品的品控有所下降。 *Multiple users report the same deformation, whereas last year’s version was firmer, indicating a batch-level decline in quality.*	**2. Q:** Is the RetroCam installation guide still usable, and what evidence across the comments supports the answer? **A:** No. One commenter reports that the download repository is offline, while another confirms that RetroCam has been discontinued.
3. Q: “还剩3天”的倒计时在两条评论中被怎样理解？ **A:** 一种理解是粉圈周年纪念，另一种是博主生日。 *Commenters read the countdown either as a fandom anniversary or as the author’s birthday.*	**3. Q:** A post shows Ningning and Giselle; a reply says “jeans one is Giselle.” Who is on the right? **A:** Ningning, because the jeans-wearer is Giselle and occupies the other position.

B Evaluation and Validation

B.1 Scoring and Inference

Candidate-answer scoring. The executable judge receives the source record, probe target, target component, question, frozen reference answer and explanation, epistemic state, compact evidence, and candidate response. It checks the core meaning, referent, relation or stance, contradictions, and unsupported specificity. Score 2 requires the core and all essential answer conditions without contradiction; 1 allows a correct core with a missing qualification or limited error; 0 misses or contradicts the core or invents a critical entity. Paraphrases and answers that respect the reference's uncertainty boundary are not penalized.

All models use fixed platform/language-specific inference prompts. Temperature is 0 except for GPT-5.6-Sol (1); thinking is disabled where available, with Gemini at low thinking. GPT uses reasoning effort `none`. The judge uses non-thinking, temperature-zero decoding. Headline scores average over probes within each language block.

The API-served model identifiers are `claude-opus-5`, `gpt-5.6-sol`, `gemini-3.8-flash`, `kimi-k3`, `deepseek-v4-pro-0813`, `GLM-5.2`, and `qwen3.5-397b-a17b`; the remaining systems use the open-weight checkpoints named in Table 2. The released evaluator uses a fixed output budget of 8,192 tokens for all models. Length-terminated responses receive score 0 and remain in the evaluation denominator; the budget is not extended.

B.2 SITU Judge Selection and Calibration

We calibrate candidate judges on 100 probes sampled with fixed platform–difficulty quotas and round-robin coverage of probe targets. Each probe contributes frozen outputs from five actor models, yielding 500 candidate answers. Gemini-3.1-Pro (low thinking) serves as a strong calibration anchor rather than human gold.

We compare six candidate judges on the same frozen answers and scoring rubric. Confidence intervals use 5,000 probe-cluster bootstrap resamples; mean bias is the candidate score minus the anchor score, and $0 \leftrightarrow 2$ counts severe score reversals. Gemini-3.8-Flash uses low thinking, GPT-5.6-Sol uses medium reasoning, Claude Opus 5 uses high effort with thinking disabled, and the open-weight judges use non-thinking decoding. Gemma-4-31B-it gives the highest agreement with the anchor (quadratic-weighted $\kappa = 0.837$, 95% CI $[0.785, 0.882]$), with a small mean score bias of -0.026 and eight severe score reversals. We therefore use it as the judge in the main experiments. Platform-, difficulty-, and actor-family breakdowns are retained in the calibration artifact.

B.3 Author Validation of References, Candidate Scores, and Diagnostic Labels

Sampling and model-assisted triage. We freeze four stratified pools before annotation: 100 reference packages with 200 candidate answers, 120 context-diagnostic cases, and 120 expression-diagnostic cases. Three heterogeneous models—Claude Opus 5 (high effort), GPT-5.6-Sol (no reasoning), and Gemini 3 Pro (high thinking)—independently pre-annotate every item using the same review packet and no external search. Author review includes every item with cross-model disagreement, low confidence, or a critical reference flag, together with a deterministic, stratum-preserving audit of stable-consensus items. Reference validation and the two candidate answers attached to it are routed as one bundle. This produces 76 reference bundles (152 candidate answers), 58 context cases, and 62 expression cases. The author is blinded to the model labels and to whether an item entered through uncertainty triage or consensus audit; model outputs determine routing only and are not treated as gold.

Author-review protocol. One paper author labels every selected item. A separate AI audit reviews the same packets independently and is used only to surface cases for a second source-grounded pass; it is not counted as a human annotator or included in agreement estimates. Final statistics use the author-reviewed labels after this pass. Reference review distinguishes valid, minor non-core repair, major core repair, and unanswerable packages. The author may search for long-tail facts during reference validation and records search use, but candidate scoring is completed and frozen before any search. For candidate scoring, the author and Gemma receive the same source record, probe metadata

Table 9: Judge calibration against Gemini-3.1-Pro (low thinking).

Judge	Exact (%)↑	QWK↑	95% CI	Mean bias	$0 \leftrightarrow 2$ ↓
Gemma-4-31B-it	82.2	**0.837**	$[0.785, 0.882]$	-0.026	8
GPT-5.6-Sol	78.6	0.784	$[0.727, 0.837]$	-0.074	10
Claude Opus 5	71.8	0.772	$[0.721, 0.817]$	+0.180	1
Gemini-3.8-Flash	74.0	0.760	$[0.700, 0.813]$	+0.238	13
Qwen3.5-27B	73.6	0.748	$[0.681, 0.805]$	-0.124	14
GLM-4.5-Air	57.0	0.526	$[0.448, 0.601]$	-0.330	56

Table 10: Author validation of candidate scoring and diagnostic annotations on the model-triaged sample.

Task	Comparison	N	Exact	κ/QWK	MAE/bias	Within one / severe
Candidate score (0–2)	Gemma–author	126	59.5	0.642	-0.262	0
C-E/C-I	Frozen–author	43	72.1	0.444	–	–
Mechanism (0–5)	Frozen–author	44	27.3	0.603	1.091	72.7

Exact and within-one agreement are percentages. For binary C-E/C-I labels, κ is Cohen's κ; ordinal tasks use quadratic-weighted κ (QWK). Bias is the first-named scorer minus the second; severe candidate disagreements are $0 \leftrightarrow 2$ reversals. "Frozen" denotes the diagnostic annotation used in the intervention analyses. The reviewed sample is enriched for model disagreement, low confidence, and critical reference flags.

and target, question, frozen reference answer and explanation, epistemic state, compact evidence, and candidate response. Candidate agreement excludes bundles whose final reference has a core defect that changes the 0/1/2 boundary. Diagnostic review follows the counterfactual C-E/C-I protocol and the 0–5 mechanism-transferability rubric in Appendix C.1.

Reference outcomes. Among the 76 reviewed packages, 56 are valid as written, 7 require only a non-core repair, 10 require a core repair, and 3 are unanswerable. The author uses external search for 9 packages. Within the independently sampled stable-consensus audit, 6 of 6 packages are valid or require only a non-core repair.

Result. On the 126 candidate answers attached to usable references, every Gemma–author disagreement is adjacent: exact agreement is 59.5%, QWK is 0.642, and the mean Gemma-minus-author bias is -0.262. On the 12 stable-consensus audit answers, exact agreement rises to 83.3% (QWK $= 0.857$). The diagnostic comparison preserves the broad structure while exposing boundary uncertainty. Overall frozen–author agreement is 72.1% ($\kappa = 0.444$) for C-E/C-I and QWK $= 0.603$ for the mechanism score; the mean frozen-minus-author score difference is -0.591. On stable-consensus audit cases, context agreement is 85.7% ($\kappa = 0.714$), while mechanism-score QWK is 0.788 with 83.3% within one point. The principal context boundary is whether an expression gloss itself recovers the answer or whether broader proposition and discourse integration remains necessary.

The Oracle/RAG review is a separate author audit of intervention mechanisms. It examines 30 expression-oracle and 30 retrieval cases selected across platforms and outcome strata; it is used for qualitative error attribution rather than agreement estimation.

B.4 Full-SITU Evaluation Including Easy Probes

Table 11 reports five open-weight foundation models and three corresponding bilingual SUU SFT checkpoints on all 89,199 probes. For each configuration, we merge 79,677 Easy and 9,522 Hard decisions by probe ID, with no missing or duplicate probes; the merged Hard scores match those used in the main experiments. The SFT checkpoints are Gemma NEW_MIXED and the Qwen3.5-9B/27B STYLE_V07 models. Easy probes do not receive diagnostic-aspect annotations, so the table retains the nine probe targets and paired Chinese/English rows. English pools Reddit and YouTube. All configurations use temperature-zero, non-thinking generation and the same platform-aware Gemma-4-31B-it judge.

Table 11: Full-SITU results by probe target (mean score /2 ↑).

Model / training	Lang.	Expression-centered				Context-integrative					All
		NS	OSNI	WV	WP	DT	SC	AP	RDL	IR	
Open-weight models (27–31B)											
Gemma-4-31B-it	zh	1.73	1.85	**1.61**	**1.68**	**1.94**	1.70	**1.92**	**1.59**	1.72	**1.84**
	en	1.57	1.75	1.75	1.79	1.92	1.83	**1.96**	1.30	1.59	1.84
+ SUU SFT	zh	**1.75**	**1.91**	1.43	1.33	1.93	**1.78**	1.90	1.54	**1.76**	**1.84**
	en	**1.77**	**1.91**	**1.86**	**1.83**	**1.96**	**1.88**	1.95	**1.68**	**1.76**	**1.90**
Qwen3.5-27B	zh	1.71	1.84	1.48	1.44	1.87	1.30	1.85	1.40	1.74	1.76
	en	1.55	1.82	1.64	1.52	1.92	1.85	1.92	1.41	1.66	1.83
+ SUU SFT	zh	**1.74**	**1.86**	**1.54**	**1.64**	**1.90**	**1.76**	**1.88**	**1.42**	**1.77**	**1.82**
	en	**1.59**	**1.83**	**1.68**	**1.54**	**1.93**	**1.89**	**1.94**	**1.49**	**1.72**	**1.85**
Open-weight models (4–9B)											
Qwen3.5-9B	zh	1.55	1.63	1.12	1.11	1.71	1.12	1.58	1.06	1.51	1.53
	en	1.29	1.58	1.29	0.98	1.82	1.66	1.70	1.12	1.54	1.60
+ SUU SFT	zh	**1.60**	**1.71**	**1.26**	**1.21**	**1.81**	**1.58**	**1.76**	**1.11**	**1.54**	**1.67**
	en	**1.37**	**1.63**	**1.39**	**1.11**	**1.86**	**1.75**	**1.80**	**1.18**	**1.57**	**1.68**
Gemma-4-E4B-it	zh	1.45	1.76	0.88	1.16	1.90	1.75	1.82	1.08	1.64	1.69
	en	1.28	1.60	1.40	0.94	1.90	1.76	1.85	1.02	1.59	1.70
Qwen3.5-4B	zh	1.36	1.48	0.83	0.87	1.52	1.18	1.36	0.84	1.41	1.35
	en	1.08	1.39	0.99	0.81	1.67	1.59	1.51	0.93	1.45	1.41

All averages probes within each language. Bold compares Foundation and SUU SFT within the same backbone, language, and column; displayed ties are retained.

Table 12: SFT gains across difficulty splits (mean score /2 ↑).

Backbone	SFT − Foundation		Full (89,199)		
	Easy	Hard	Foundation	SUU SFT	Δ
Gemma-4-31B-it	+0.0137	+0.2513	1.8423	1.8814	+0.0391
Qwen3.5-27B	+0.0317	+0.0813	1.8030	1.8399	+0.0370
Qwen3.5-9B	+0.1014	+0.1118	1.5729	1.6754	+0.1025

Easy: 79,677 probes; Hard: 9,522. Full is their probe-weighted union. Deltas use unrounded scores.

SFT beyond the difficulty-filtered subset. Table 12 shows positive SFT gains on Easy, Hard, and Full for all three backbones. Gemma rises from 1.8423 to 1.8814 on Full: +0.0391, with a paired source-record bootstrap 95% interval of $[+0.0357, +0.0426]$. Qwen-9B improves by +0.1025 and Qwen-27B by +0.0370; their Full intervals also exclude zero (5,000 draws). These results confirm that the overall SFT benefit persists when evaluation includes Easy probes, rather than depending on restriction to the Gemma-filtered Hard subset. Gemma's Full gain is concentrated in English (+0.0596); Chinese is nearly unchanged (+0.0008), where Hard gains are largely offset by Easy losses. Table 11 shows the corresponding target-level variation.

For Gemma, the Chinese aggregate masks a preservation trade-off: bilingual SFT lowers the Xiaohongshu Easy score from 1.9839 to 1.9296 (-0.0542; paired 95% CI $[-0.0581, -0.0504]$), while improving the 4,336 Chinese Hard probes by +0.3420. The Easy slice is strongly Foundation-conditioned: 94.27% of its probes received full credit in all five filtering runs, and the remainder did so in four. The regression nevertheless reflects a systematic output shift as well as this ceiling: mean response length contracts from 105 to 57 Chinese characters, and case inspection finds omitted pragmatic implications alongside overconfident canonical-form or formation-mechanism guesses on word-variation probes. Thus, the near-zero full Chinese change reflects large hard-case gains offset by a smaller preservation and calibration cost on cases the Foundation already solved reliably.

Aggregation and selection sensitivity. Rank correlations between probe-micro results and record-, platform-, language-, or target-balanced aggregates are at least 0.993 across the twelve Hard evaluations. Among the five foundation models evaluated on the complete suite, Hard versus full-

(a) Xiaohongshu · SC
semantic content

SOURCE
Title: 通贩周边找个搭子
Content: 拼拼盒子 我卡你本体 37分，后续国内邮费直接AA。
Translation: Split the set: I take the card; you take the main item; split the cost 3:7; later domestic shipping is shared equally.

REFERENCE
The poster keeps the card, the partner takes the main item, and their costs follow a 3:7 split.

MODEL · DIRECT 0/2
"方案为对方得盒子，发布者得本体，费用按3:7分担。"
Translation: The partner gets the box; the poster gets the main item; costs are split 3:7.

ERROR
Reverses the item allocation.

(b) Reddit · AP
attitude / pragmatics

SOURCE
`user_2Z44`: "Don't tilt … If you lose 3 in a row, take a break."
↪ `user_2HT6`: "Sometimes I'm the clueless one."
↪ `user_G6MS`: "Bro's talking from experience"

REFERENCE
`user_G6MS` teases the original adviser, `user_2Z44`, implying that the advice comes from experience.

MODEL · DIRECT 0/2
"[The speaker] uses a mocking or teasing tone to address `user_2HT6`, implying that the latter's admission of being 'clueless' is a result of personal experience."

ERROR
Follows the adjacent sibling instead of the reply target.

Figure 7: Representative SITU-Hard failures. Left: resolving community-specific transaction notation. Right: respecting reply-tree structure. Model excerpts are verbatim Gemma-4-31B-it direct outputs; contexts are shortened.

suite rank correlation is 0.700; Gemma moves from third to first. This reflects the Gemma-conditioned difficulty filter.

Paired intervention comparisons use 5,000 bootstrap draws over source records, keeping all probes from a sampled record and both systems' outcomes in the same cluster. Unless otherwise stated, these intervals quantify uncertainty over evaluation records conditional on the trained checkpoints; the two-seed scaling experiment reports the across-seed mean separately. Overall gains are positive for Two-round RAG, Agentic RAG, bilingual SFT, and SFT+RAG.

B.5 Representative Model Errors

C Diagnostic Interventions and Learning Analysis

C.1 Diagnostic Annotation and Weighting

Reference-conditioned construct. Both annotations receive the source observation and question plus the frozen reference answer and explanation. They diagnose the knowledge and reasoning needed to recover that target, not every interpretation licensed by the raw record. Source evidence takes precedence over the reference; a `reference_problem` flag routes unsupported or ambiguous cases to repair or exclusion. Any semantic repair invalidates both the diagnostic label and the earlier difficulty audit. This implements an evidence-centered assessment argument (Mislevy et al., 2003; Kane, 2013); the subsequent retrieval, oracle, and SFT interventions test the interpretation of the labels rather than defining it.

Mechanism transferability. For each expression-centered probe, the annotator assumes that a solver knows ordinary language and has seen diverse demonstrations of the same mechanism, but has never seen the exact item–meaning mapping. The 0–5 score measures how much answer-critical meaning a concrete rule can recover for genuinely new surfaces: 5 indicates nearly complete rule-based recovery, while 0 indicates an arbitrary item, sense, entity, title, or convention. Merely recognizing an abbreviation or recommending retrieval receives no mechanism credit. Annotators state the rule, its cross-surface prediction, and the unresolved residue. We assign soft weights using

$w_M(s) = s/7$ for $s \leq 3.5$ and $w_M(s) = 0.5 + (s - 3.5)/3$ otherwise, with $w_K(s) = 1 - w_M(s)$. The midpoint lies between ordinal levels 3 and 4; the endpoints retain weights 0 and 1. The 4,158 Hard expression probes have effective mass 29.15% E-M and 70.85% E-K.

Weighting sensitivity. Using proportional weights $w_M(s) = s/5$ instead gives final Chinese scaling gains of +0.509 for E-M and +0.439 for E-K, compared with +0.515 and +0.445 under the piecewise map. Both weightings retain the same qualitative learning pattern.

Expression bottleneck versus integration dependence. For a context-integrative probe, C-E applies when withholding an answer-critical sense, phrase, alias, entity, title, abbreviation, slang item, hashtag, emoji, code-switched expression, cultural reference, or wordplay causes material loss; C-I applies when ordinary language plus contextual integration suffices. The annotation protocol applies two complementary tests: (i) replace the candidate span with an uninterpretable placeholder while preserving syntax and discourse, and (ii) restore only a minimal legitimate gloss. A positive label requires loss under ablation and recovery under the gloss; it stores the exact source span and gloss. The gloss may not inject coreference, event relations, stance, or the final conclusion. Disagreements are adjudicated with both analyses and the source record, while any reference problem returns to the repair pipeline. The final 5,364 Hard context probes comprise 2,177 C-E and 3,187 C-I instances.

C.2 Training Data

SUU SFT data. Training probes are mined with the same selection and extraction stages as SITU, followed by lighter quality control. Because training retains probes that the target backbone does not already solve, the exact corpus varies slightly across model families. Section 5 reports the size and language composition of the bilingual Gemma corpus. The bilingual row reports one model trained on the union of the Chinese and English corpora; the monolingual row combines the Chinese score of a Chinese-only model with the English score of an English-only model and serves as a language-matched reference. The separate scaling stream contains roughly 510k usable Chinese Xiaohongshu demonstrations. The Qwen3.5 controls use the final style-controlled bilingual corpora constructed separately for the 9B and 27B backbones.

Data separation and overlap audit. All SUU training sources are disjoint from SITU at the source-record level. Downstream comparisons use the same fixed task train–test splits for both initializations. For the five translation benchmarks, normalized exact-match and long-substring checks found no test-input inclusion in the Gemma bilingual SUU training inputs or translation replay. The audit considers inputs with at least 20 normalized characters for predominantly Chinese text and 50 otherwise. Recurring expressions and meanings are analyzed separately as knowledge exposure.

C.3 Optimization and Inference

Unless noted below, all trained systems use full-parameter SFT with fused AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-8}$, weight decay 0.01), BF16, gradient clipping at 1.0, a peak learning rate of 10^{-6}, cosine decay after 5% warmup, and global batch size 64. The SUU and translation stages run for two epochs with a maximum sequence length of 4,096. Within every downstream comparison, the two initializations use identical task data, optimization, checkpoint selection, and decoding. Candidate generation is temperature-zero and non-thinking.

Scaling-run checkpoints. The two Chinese scaling runs use the same optimizer, peak learning rate, batch size, and sequence length, but apply linear warmup through step 399 (5%) followed by a constant learning rate of 10^{-6}. Checkpoints are fixed at predetermined optimizer steps rather than selected on SITU performance. For each seed, the displayed checkpoints are nested prefixes of a fixed shuffled pass over the 509.7k-example stream; no examples are repeated or newly mined between points. Figure 4 reports probe-aligned means on the 4,336-probe Chinese SITU-Hard slice, while exact-match coverage is computed within each shuffled order and then averaged. Table 13 maps the steps to the original data units and approximate cumulative examples and target tokens.

Table 13: Training checkpoint ledger. Examples and loss-bearing target tokens are approximate; optimizer steps are exact.

Data unit	0.1	0.2	0.4	0.6	0.8	1.0	1.5
Examples	6.4k	12.7k	25.5k	38.2k	51.0k	63.7k	95.6k
Target tokens	0.3M	0.6M	1.2M	1.8M	2.4M	3.1M	4.6M
Optimizer step	100	199	398	597	797	996	1493

Data unit	2.0	3.0	4.0	5.0	6.0	7.0	8.0
Examples	127.4k	191.2k	254.9k	318.6k	382.3k	446.0k	509.7k
Target tokens	6.1M	9.2M	12.2M	15.3M	18.3M	21.4M	24.4M
Optimizer step	1991	2987	3983	4978	5974	6969	7965

Table 14: RAG prompt control on SITU-Hard (mean score /2 ↑).

Setting	Chinese	English	Overall
Direct	0.9419	1.2470	1.1081
RAG format, no evidence	0.9958	1.2543	1.1366
Two-round RAG	1.3037	1.4003	1.3563

C.4 Retrieval Prompt and Budget Controls

Two-round RAG uses one planner call and one answer call, allows at most three Bing queries, and retains the top five title–snippet results per query. For Gemma-4-31B-it, the planner issues 1.278 queries per probe on average and chooses not to search for 33.35% of probes. Agentic RAG permits up to four search actions with the same per-action limits; it averages 0.974 search actions and 2.157 queries per probe. Returned packets are cached by normalized query and locale. The evidence block is capped at 12,000 characters, and the generated-token caps are 512 for query planning, 1,024 for each agentic controller decision, and 2,048 for the final answer.

To isolate the answer-wrapper effect, we apply the exact Two-round RAG answer prompt to every probe with an empty query list and no evidence. Table 14 shows a small overall gain of 0.0286, compared with 0.2483 for full Two-round RAG. The planner–answerer crossover in Table 15 provides the complementary fixed-evidence comparison: each answerer receives the same cached query–snippet packet within a planner row.

Retrieval-hit analyses condition on an observed outcome: probe difficulty and the planner–retriever behavior jointly affect whether answer-critical evidence is found. Hit/miss contrasts are therefore descriptive; Figure 3 computes each conditional gain against direct inference on the same probes.

Expression oracle. We prepend the frozen answers to all expression-centered probes from the same source record before answering a context-integrative probe. This supplies a record-level expression-knowledge package. The minimal-gloss counterfactual for diagnostic annotation is defined separately above.

The crossover further shows a planner–answerer interaction rather than a universally superior planner. Broader Qwen evidence packets work well with the more noise-resistant Gemma answerer, while Qwen answerers generally benefit from Gemma's more selective packets. Qwen3.5-9B is most sensitive to this mismatch; Qwen3.5-27B is substantially closer to Gemma in evidence use.

C.5 Intervention Robustness across Backbones

Table 16 repeats the core direct, retrieval, oracle, and SFT comparisons on Qwen3.5-9B and Qwen3.5-27B.

The intervention pattern largely agrees with Gemma: retrieval helps E-K most, direct SFT improves a broader set of aspects, and their combination gives the best overall score for both Qwen backbones. The main Qwen-specific boundary is evidence noise. Qwen3.5-27B gains on C-E but loses slightly on

Table 15: Two-round RAG planner–answerer crossover (mean score /2 ↑). Rows are planners; columns are answerers.

Planner	Gemma	Qwen-9B	Qwen-27B
Gemma	1.3563	1.1526	1.3959
Qwen-9B	1.3965	1.0423	1.3673
Qwen-27B	1.4043	1.0961	1.3828

Table 16: Qwen interventions on SITU-Hard (mean score /2 ↑).

Intervention	Chinese / Xiaohongshu					English / Reddit + YouTube				
	E-M	E-K	C-E	C-I	All	E-M	E-K	C-E	C-I	All
Backbone: Qwen3.5-27B										
Foundation (direct)	1.18	1.02	1.20	1.33	1.16	1.05	0.79	1.34	1.62	1.29
+ Two-round RAG	1.39	1.29	1.29	1.35	1.32	1.28	1.28	1.45	1.55	1.44
+ Expression oracle[a]	–	–	1.41	1.30	–	–	–	1.44	1.49	–
+ bilingual SUU SFT	1.26	1.08	1.36	1.55	1.28	1.05	0.83	1.38	1.68	1.34
+ Two-round RAG	1.41	1.32	1.45	1.55	1.41	1.29	1.28	1.49	1.63	1.48
Backbone: Qwen3.5-9B										
Foundation (direct)	0.87	0.76	0.95	1.08	0.89	0.72	0.58	1.10	1.39	1.06
+ Two-round RAG	1.01	1.01	0.93	1.03	0.99	0.99	1.04	1.05	1.15	1.09
+ Expression oracle[a]	–	–	1.18	1.06	–	–	–	1.18	1.13	–
+ bilingual SUU SFT	1.00	0.87	1.18	1.34	1.07	0.77	0.62	1.16	1.47	1.12
+ Two-round RAG	1.15	1.10	1.14	1.29	1.15	1.04	1.08	1.21	1.36	1.23

[a] The expression oracle is evaluated only on context-integrative probes and is compared with the corresponding Foundation context subsets.

C-I, whereas Qwen3.5-9B loses on both under direct retrieval; expression-oracle evidence recovers C-E for both models but still degrades C-I. Thus, expression knowledge is useful across backbones, but exploiting it inside a longer context depends on the answerer's evidence robustness.

C.6 Learning Curves and Exposure

Retrieval hits are labeled from the question, reference package, and snippets without the candidate answer or score. A hit contains answer-critical evidence. SFT targeted exposure requires normalized exact equality between an evaluation focus and a targeted training focus. Hit/miss groups are fixed using the complete training stream.

The SFT curves in Figure 8 complement the retrieval analysis in Figure 3. Exact targeted exposure helps both E-M and E-K, but the final hit–miss gap is much larger for E-K: E-K improves by +0.54 when hit and +0.24 when missed, whereas E-M improves by +0.58 and +0.43. Section 5.4 further examines incidental mentions and same-meaning exposure. C-E lies between the original hit–miss patterns (+0.36 versus +0.21), while the C-I curves remain close in absolute score.

C.7 Full-Stream Exposure Audit

The targeted-focus hit definition requires equality with a training focus. We also scan every user input and assistant response in the approximately 510k-example Chinese scaling stream after NFKC normalization, case folding, and whitespace collapse. Counting literal CJK substrings as exposure leaves 383 probes whose complete target surfaces are absent.

Full-stream semantic retrieval. We index complete inputs and responses from all nine training probe targets, including context-integrative demonstrations, and add question–answer views. Deduplication and overlapping 464-token windows (stride 400) yield about 1.41M indexed spans. Each probe retrieves the top 20 distinct demonstrations with character 2–5-gram TF–IDF and another 20 with exact cosine search using `multilingual-e5-small`. Queries contain the target surface,

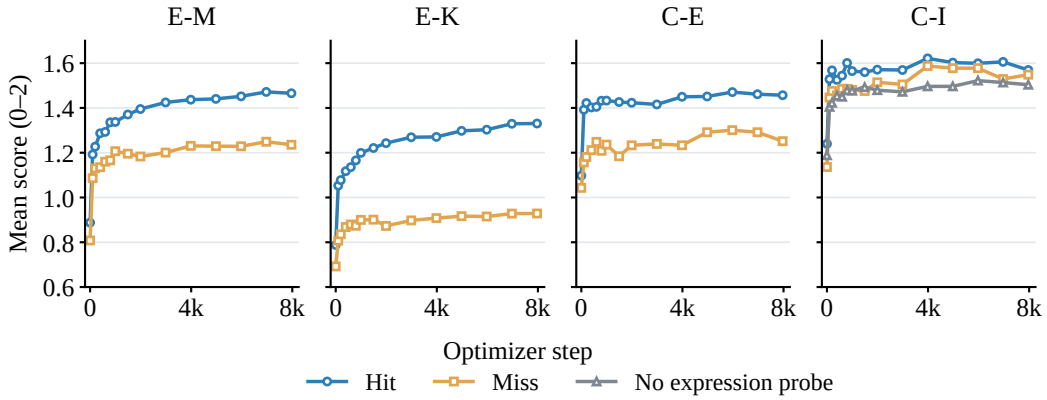

Figure 8: Absolute Chinese SFT performance averaged over two training seeds. E-M, E-K, and C-E panels compare Hit with Miss; for C-I, No expression probe denotes probes without a relevant expression candidate. Groups are fixed across checkpoints, and E-M/E-K use complementary soft weights.

occurrence, and reference meaning. We also retain the eight candidates from an earlier lexical-only TF–IDF audit, yielding 17,063 probe–demonstration pairs.

Knowledge relations and adjudication. Gemini 3.1 Pro (`gemini-3.1-pro-preview`, medium thinking, temperature zero) reviews each probe's complete pool of 28–48 candidates in one request. It receives the target occurrence, question, reference answers, and candidate questions, answers, and retrieved passages. Model scores, diagnostic weights, retrieval ranks, and earlier audit decisions are withheld. Each candidate is labeled as supplying *same-item knowledge*, *component knowledge*, *a mechanism analogy*, *unrelated evidence*, or *uncertain evidence*. Same-item knowledge includes an explicit mapping from the target or an equivalent alias to its meaning, as well as knowledge of the canonical item when the new surface still requires decoding. Component knowledge supplies a distinctive constituent meaning, whereas a mechanism analogy demonstrates a reusable rule on another item. The judge also assesses what the candidates jointly supply and what decoding or composition remains. For example, demonstrations explaining 修仙 *(cultivation fantasy)* and 本 *(jubensha scenario)* supply the components of 修仙本 *(cultivation-themed jubensha scenario)*, while their combination remains to be interpreted.

We obtain valid judgments for 382 of the 383 probes; one persistent empty response is excluded from all adjudicated subsets. Grouping probes by their strongest detected relation gives 36 with same-item knowledge, 189 with component knowledge, 62 with a mechanism analogy, 2 with uncertain evidence, and 93 with unrelated evidence only. The main exclusion rule removes any same-item knowledge, leaving 346 probes from 305 source records with *no detected same-item exposure*. Relations to individual candidates and joint coverage of constituents remain separate: several taught components do not automatically constitute a taught complete expression. The exclusion rules were specified before receiving the judgments. They concern the inspected SFT stream and retrieved evidence, not the backbone's pretraining history.

Controlled retrieval check. Before retrieval, we independently sample 30 expression-centered and 30 context-integrative training demonstrations. Gemma paraphrases each answer as a query; Qwen checks fidelity. A source-consistency check removes two factual shifts, leaving 58 queries. The paraphrases are never added to the index. Dense and character retrieval recover the known source in 52/58 and 41/58 cases; their union recovers 53/58 (91.4%), including 29/30 expression and 24/28 context cases. This measures controlled source recovery rather than exhaustive recall of semantic exposure among evaluation probes.

Paired estimates and sensitivity. We average the two final-run scores per probe, subtract the matched Foundation score, and bootstrap source records for 20,000 draws within each retained subset.

Table 17: Sensitivity to semantic-exposure exclusions. Gains in mean score /2 ↑; 95% source-record bootstrap intervals in brackets.

Excluded evidence	Retained	E-M gain	E-K gain	E-M−E-K
Explicit mapping only	367	$+0.227$ $[+0.104, +0.350]$	$+0.060$ $[-0.032, +0.153]$	$+0.167$ $[+0.053, +0.280]$
Same item	346	$+0.204$ $[+0.075, +0.329]$	$+0.051$ $[-0.043, +0.147]$	$+0.153$ $[+0.038, +0.265]$
Same item + components	157	$+0.075$ $[-0.117, +0.279]$	-0.033 $[-0.154, +0.090]$	$+0.108$ $[-0.077, +0.299]$
Same item + flagged probes	317	$+0.243$ $[+0.126, +0.360]$	$+0.068$ $[-0.025, +0.164]$	$+0.175$ $[+0.073, +0.278]$

All probes from a sampled record stay together. The E-M and E-K weighted mean gains and their difference are recomputed in each common draw, accounting for their overlapping, complementary weights. The intervals measure source-sample uncertainty conditional on the two trained runs and fixed audit labels.

Table 17 reports the predefined exclusions. Removing same-item knowledge leaves an E-M gain of +0.204 (95% CI $[+0.075, +0.329]$) and a paired E-M–E-K difference of +0.153 ($[+0.038, +0.265]$). A narrower rule removes only explicit mappings (15 probes), retaining cases where canonical-item knowledge still requires surface decoding. A broader rule also removes component knowledge, leaving 157 probes; both its E-M gain and paired contrast have intervals crossing zero. These rules change the evaluated population, so their contrasts do not isolate a causal effect of component exposure.

Uncertainty and reference concerns are retained as separate flags. At least one candidate is uncertain for 5 probes, and 31 receive a reference-concern flag. Because adjudication uses the target occurrence and reference rather than the complete social-media record, these flags indicate a need for review rather than confirmed reference errors. Excluding flagged probes in addition to same-item exposure leaves 317 probes and a positive E-M gain of +0.243 ($[+0.126, +0.360]$).

The retained-set gains are consistent with improved use of reusable expression mechanisms and component knowledge. The audit addresses exposure to SFT item knowledge; rule acquisition, better use of existing rules, and adaptation to the explanation task remain possible sources of the gains.

D Downstream and External Transfer

D.1 Matched Arms and Training Data

Within each backbone, task-trained arms share task data, optimization, checkpoint selection, and decoding. The SUU-initialized arm has an additional interpretation-training stage. Translation also includes a mixed SUU+translation arm.

Translation replay. The replay corpus combines direct and context-conditioned Xiaohongshu translation with a smaller general-domain component; approximately 80% is Xiaohongshu-derived. It spans the target languages used in training and preserves social-media structure such as emoji, hashtags, placeholders, and XML-like delimiters. Section 6 reports its total size. Replay is used only in the designated translation and joint arms.

Task-specific optimization. HED-COLD uses the prespecified epoch-2 checkpoint, SeTox uses epoch 5 with maximum length 1,024, and CounterContext uses epoch 3 with maximum length 2,048; these task-specific choices are fixed across the two initializations. Common settings appear in Appendix C.3.

Table 18: Translation on Xiaohongshu-sourced benchmarks.

Method	RedTrans-Bench						CSM-MTBench ({zh→es,fr,ja,ko,ru})				
	zh→en			en→zh			Fun Posts		Social Snippets		
	BLEU ↑	chrF++ ↑	XC ↑	BLEU ↑	chrF++ ↑	XC ↑	XC ↑	SSR ↑	XC ↑	ES ↑	GEMBA ↑
Foundation	25.98	52.36	0.822	40.67	28.15	0.794	0.838	<u>0.682</u>	0.784	0.716	<u>4.15</u>
SUU SFT	27.91	54.30	**0.834**	41.61	28.98	**0.810**	**0.841**	0.675	<u>0.785</u>	**0.720**	4.05
Foundation→T	27.14	53.07	0.823	40.11	29.33	0.798	0.837	0.681	0.783	<u>0.718</u>	**4.17**
SUU→T	<u>28.32</u>	<u>54.31</u>	0.829	<u>42.55</u>	**30.48**	<u>0.809</u>	<u>0.839</u>	**0.685**	**0.788**	0.716	<u>4.15</u>
SUU+T	**28.36**	**54.40**	<u>0.829</u>	**42.61**	<u>30.42</u>	0.808	0.838	0.680	0.784	0.718	4.12

Conventions. CSM-MTBench reports direction-macro averages over the five listed target languages. XC denotes XCOMET-XL and T denotes translation-task SFT; GEMBA uses Gemini-3.8-Flash (low thinking). Bold and underlining mark the best and second-best unrounded results, respectively.

D.2 Social-Media Translation Benchmarks and Protocols

RedTrans-Bench. RedTrans-Bench contains Xiaohongshu-sourced note titles, note content, and comments in both Chinese-to-English and English-to-Chinese directions. It targets phenomena such as slang, humor, emoji semantics, and culturally grounded adaptation (Guo et al., 2025c). We evaluate the full release with corpus BLEU, corpus chrF++, and XCOMET-XL, keeping the two directions separate.

CSM-MTBench. CSM-MTBench contains two Chinese-source subsets translated into Spanish, French, Japanese, Korean, and Russian (Zhao et al., 2026b). Fun Posts contains context-rich sources with slang and neologisms, while Social Snippets contains shorter, emotion- and style-heavy sources. Following the benchmark design, we report XCOMET-XL and slang success rate (SSR) for Fun Posts, and XCOMET-XL, embedding similarity (ES), and a GEMBA-style judge score for Social Snippets. For GEMBA, we retain the released prompt but replace GPT-4o with Gemini-3.8-Flash (low thinking, temperature zero) and score the full released subset.

WMT22 Social. We select the complete social-domain slice of the WMT22 General MT test set: English Reddit documents translated into Czech, German, Japanese, Russian, Ukrainian, and Chinese. Czech, German, and Chinese have two human references per source; the remaining directions have one. We report direction-macro averages of corpus BLEU, corpus chrF++, and XCOMET-XL.

RoCS-MT v2. RoCS-MT v2 provides English Reddit posts with professional references in Czech, German, French, Russian, and Ukrainian (Bawden & Sagot, 2025). It pairs the original non-standard text with manual normalization and span-level annotations for phenomena such as spelling variation, acronymization, devowelling, and stretching. We evaluate the full raw-input test set, preserving the variation that the benchmark was designed to test, and macro-average corpus BLEU, corpus chrF++, and XCOMET-XL across the five directions.

MTNT. MTNT is a classic noisy-text translation benchmark built from naturally occurring Reddit comments and professional translations (Michel & Neubig, 2018). We use the full English-to-French and English-to-Japanese test sets and report direction-macro corpus BLEU, corpus chrF++, and XCOMET-XL.

Metric implementation. Corpus BLEU and chrF++ are computed with SacreBLEU 2.5.1. BLEU uses `intl` for MTNT, `zh` for Chinese targets, `ja-mecab` for Japanese, `ko-mecab` for Korean, and `13a` otherwise. chrF++ uses character order 6, word order 2, $\beta = 2$, and excludes whitespace from character n-grams. Because BLEU tokenization and the word component of chrF++ operate on different units, their absolute scales are not comparable across translation directions; both are treated as within-direction auxiliary metrics.

Table 19: Translation on Reddit-sourced benchmarks.

Method	WMT22 Social (en→{cs,de,ja,ru,uk,zh})			RoCS-MT v2 (en→{cs,de,fr,ru,uk})			MTNT 1.1 (en→{fr,ja})		
	BLEU ↑	chrF++ ↑	XC ↑	BLEU ↑	chrF++ ↑	XC ↑	BLEU ↑	chrF++ ↑	XC ↑
Foundation	32.06	44.66	0.900	23.63	45.30	0.846	29.56	43.20	0.864
SUU SFT	**33.02**	**45.40**	**0.904**	25.51	45.99	**0.853**	**30.03**	**43.84**	**0.871**
Foundation→T	32.00	44.12	0.897	24.49	45.33	0.845	29.54	42.94	0.862
SUU→T	32.68	44.97	0.902	<u>25.77</u>	<u>46.44</u>	0.851	<u>29.86</u>	43.29	0.866
SUU+T	<u>32.88</u>	<u>45.11</u>	<u>0.903</u>	**25.79**	**46.45**	<u>0.852</u>	29.79	<u>43.33</u>	<u>0.867</u>

Conventions. Scores are direction-macro averages over the listed targets; RoCS-MT v2 uses raw input. XC denotes XCOMET-XL and T denotes translation-task SFT. Bold and underlining mark the best and second-best unrounded results, respectively.

D.3 Complete Gemma Translation Results

D.4 Qwen Translation Replications

We repeat the translation comparison with Qwen3.5-9B and Qwen3.5-27B. Each backbone uses its own style-controlled bilingual SUU checkpoint; within a backbone, Foundation and SUU initializations receive the same translation data and evaluation protocol. Unlike the Gemma experiment, these replications include sequential but not mixed training. CSM-MTBench GEMBA was not run for Qwen and is marked with a dash.

Table 20: Qwen replications on Xiaohongshu-sourced translation benchmarks.

Method	RedTrans-Bench						CSM-MTBench ({zh→es,fr,ja,ko,ru})				
	zh→en			en→zh			Fun Posts		Social Snippets		
	BLEU ↑	chrF++ ↑	XC ↑	BLEU ↑	chrF++ ↑	XC ↑	XC ↑	SSR ↑	XC ↑	ES ↑	GEMBA ↑
Backbone: Qwen3.5-9B											
Foundation	26.49	52.73	0.818	40.04	29.93	0.810	0.822	0.583	0.750	0.723	–
SUU SFT	28.16	54.67	0.822	41.60	31.35	0.810	0.821	0.578	0.754	0.731	–
Foundation→T	**33.17**	**57.58**	**0.828**	<u>43.27</u>	<u>33.88</u>	**0.819**	<u>0.829</u>	<u>0.604</u>	<u>0.762</u>	<u>0.733</u>	–
SUU→T	<u>32.69</u>	<u>57.17</u>	<u>0.827</u>	**43.34**	**34.05**	<u>0.818</u>	**0.830**	**0.616**	**0.764**	**0.734**	–
Backbone: Qwen3.5-27B											
Foundation	26.05	53.24	0.829	40.48	30.05	0.821	0.838	0.626	0.771	0.718	–
SUU SFT	28.03	54.88	0.829	43.07	32.97	0.818	0.839	0.641	0.767	0.723	–
Foundation→T	<u>34.76</u>	<u>58.89</u>	<u>0.835</u>	<u>46.81</u>	<u>36.37</u>	**0.824**	**0.840**	<u>0.650</u>	<u>0.780</u>	**0.727**	–
SUU→T	**35.33**	**59.36**	**0.836**	**46.92**	**36.44**	<u>0.824</u>	<u>0.839</u>	**0.664**	**0.780**	<u>0.724</u>	–

Conventions. Identical to Table 18. Bold and underlining mark the best and second-best unrounded results within each backbone; a dash denotes an unevaluated metric.

The Qwen results preserve the possibility of positive transfer but not the uniformly positive Gemma pattern. For Qwen3.5-9B, sequential SUU initialization improves post-task XCOMET-XL on both CSM subsets, WMT22 Social, and RoCS-MT v2, while slightly reducing it on both RedTrans directions and MTNT. For Qwen3.5-27B, all seven post-task XCOMET-XL changes are within 0.0004 and alternate in sign, although several BLEU, chrF++, and SSR values improve. Direct SUU initialization is likewise mixed. These replications therefore support a selective, backbone- and metric-dependent transfer claim rather than universal monotonic improvement.

D.5 Moderation Benchmarks and Protocols

The backbone comparison is heterogeneous rather than uniformly positive. Qwen3.5-9B broadly retains the Gemma pattern on direct SeTox and gains after clean-only HED-COLD training, but shows small reversals on perturbed direct HED-COLD, the SeTox general control after task SFT, and target-only CounterContext. Qwen3.5-27B instead loses under direct SUU initialization on HED-COLD and SeTox, while its CounterContext scores improve both directly and after task SFT.

Table 21: Qwen replications on Reddit-sourced translation benchmarks.

Method	WMT22 Social (en→{cs,de,ja,ru,uk,zh})			RoCS-MT v2 (en→{cs,de,fr,ru,uk})			MTNT 1.1 (en→{fr,ja})		
	BLEU ↑	chrF++ ↑	XC ↑	BLEU ↑	chrF++ ↑	XC ↑	BLEU ↑	chrF++ ↑	XC ↑
Backbone: Qwen3.5-9B									
Foundation	30.58	43.29	0.876	23.07	43.83	0.821	27.07	42.20	0.831
SUU SFT	30.52	43.42	0.871	23.23	44.91	0.819	26.68	42.29	0.828
Foundation→T	32.30	44.43	0.889	23.48	44.29	0.836	**29.34**	**43.13**	**0.857**
SUU→T	**32.41**	**44.56**	**0.890**	**24.45**	**45.44**	**0.838**	29.17	43.09	0.855
Backbone: Qwen3.5-27B									
Foundation	32.08	44.32	0.898	25.79	**46.83**	**0.855**	28.91	43.06	0.863
SUU SFT	32.67	44.94	0.898	25.51	46.74	0.852	29.27	43.62	0.860
Foundation→T	**33.66**	**45.04**	0.902	25.78	46.79	0.854	30.06	43.73	**0.868**
SUU→T	33.47	44.96	**0.902**	**25.85**	46.83	0.854	**30.08**	**43.80**	0.868

Conventions. Identical to Table 19. Bold and underlining mark the best and second-best unrounded results within each backbone.

Table 22: Toxicity-detection transfer across three backbones (macro-F1 ↑, 0–1).

Benchmark	Test condition	Direct		After matched task SFT	
		Foundation	SUU	Foundation+Task	SUU+Task
Backbone: Gemma-4-31B-it					
HED-COLD	Clean (T-clean)	0.6700	**0.6963**	0.8482	**0.8823**
	Perturbed (T-clean)	0.6737	**0.6941**	0.8338	**0.8653**
	Clean (T-perturbed)	0.6700	**0.6963**	**0.8669**	0.8627
	Perturbed (T-perturbed)	0.6737	**0.6941**	0.8624	**0.8630**
SeTox	Neologism	0.5611	**0.5912**	0.9859	**0.9905**
	General safety	0.7920	**0.8219**	0.9614	**0.9671**
CounterContext	With context	**0.5361**	0.5262	0.6360	**0.6578**
	Target only	**0.5630**	0.5399	0.6257	**0.6616**
Backbone: Qwen3.5-9B					
HED-COLD	Clean (T-clean)	0.7075	**0.7092**	0.8658	**0.8785**
	Perturbed (T-clean)	**0.6948**	0.6914	0.8283	**0.8510**
	Clean (T-perturbed)	0.7075	**0.7092**	0.8710	**0.8758**
	Perturbed (T-perturbed)	**0.6948**	0.6914	0.8663	**0.8748**
SeTox	Neologism	0.5597	**0.6035**	**0.9812**	**0.9812**
	General safety	0.8346	**0.8527**	**0.9562**	0.9534
CounterContext	With context	0.5501	**0.5528**	0.6213	**0.6230**
	Target only	**0.5132**	0.5105	**0.6102**	0.6050
Backbone: Qwen3.5-27B					
HED-COLD	Clean (T-clean)	**0.7043**	0.6959	**0.8866**	0.8818
	Perturbed (T-clean)	**0.6880**	0.6836	**0.8673**	0.8620
	Clean (T-perturbed)	**0.7043**	0.6959	0.8833	**0.8836**
	Perturbed (T-perturbed)	**0.6880**	0.6836	0.8820	**0.8832**
SeTox	Neologism	**0.6034**	0.5165	**0.9858**	0.9835
	General safety	**0.8208**	0.7885	0.9643	**0.9670**
CounterContext	With context	0.5748	**0.5857**	0.6297	**0.6364**
	Target only	0.5637	**0.5657**	0.6306	**0.6364**

Conventions. T-clean and T-perturbed denote HED-COLD task SFT on clean and perturbed inputs, respectively. Direct scores are repeated across those two HED training views for comparison. Bold marks the higher score within each matched pair (both for an exact tie), not significance. The Gemma block underlies Figure 5.

Task supervision often absorbs the initialization gap, as on SeTox, and can either recover or preserve a smaller advantage. Across backbones, these results support selective transfer rather than a guarantee that SUU initialization improves every downstream task.

HED-COLD. HED-COLD pairs Chinese offensive-language examples with homophone-perturbed versions designed to evade surface-form detectors (Wu et al., 2025). We construct size-matched clean-only and perturbed-only arms from the same official training and development source IDs, then evaluate both views of the paired test set. The primary score is macro-F1; pair consistency and the rate at which both members are classified correctly measure robustness across the paired views.

SeTox. SeTox evaluates Chinese toxicity conveyed by neologisms: new forms or established-looking expressions whose online senses have evolved into harmful uses (Cui et al., 2026). We evaluate the released 624-item neologism test set and its 368-item general-safety control without web search, thereby measuring knowledge available in the static model rather than the paper's search-augmented system. Matched task learning uses the released SFT data, split into 1,434 training and 159 development records. We report accuracy and macro-F1; among its two splits, the neologism test is the closer match to item-specific expression knowledge.

CounterContext. CounterContext contains English Reddit parent-comment/reply pairs labeled as hate speech, neutral speech, or counterspeech (Yu et al., 2022). Its task is three-way classification, not counterspeech generation. We train on the 3,325 adjudicated gold training pairs and evaluate the official 713-item test set; silver labels are excluded. Each source pair yields a with-context instance and a target-only control, and both matched initializations see identical versions. Accuracy and macro-F1 are reported, with macro-F1 primary because the discourse-role classes are imbalanced.

D.6 Paired Uncertainty Checks

We use 10,000 paired bootstrap resamples throughout. Translation resampling is clustered by source and preserves each benchmark's direction-macro aggregation; classification resamples paired test examples. Positive deltas favor the SUU arm: direct comparisons subtract Foundation, while post-task comparisons subtract the matched Foundation-initialized task learner.

The tests support every sequential post-task XCOMET-XL gain and all mixed-training gains except the two CSM-MTBench comparisons. For the Gemma toxicity results visualized in Figure 5, they support the direct SeTox gains and the clean-trained HED-COLD gains, but not the small post-task SeTox differences or either with-context CounterContext difference. Only the post-task target-only CounterContext gain is reliable, so these data do not isolate improved context use.

D.7 Matched Task Transfer on Variation Benchmarks

Table 23 separates direct transfer from matched task learning on three external variation benchmarks and three backbones. Within each backbone, the task-trained arms share the same benchmark-specific training recipe and differ only in whether they start from the Foundation or that backbone's bilingual SUU initialization. All evaluations use non-thinking decoding.

D.8 Zero-Shot Behavioral Changes after SUU Training

We additionally evaluate the bilingual SUU SFT checkpoint directly on external benchmarks not reported in the main transfer experiments: Lost in Pronunciation, Chumor 2.0, CHIME, CHEER, SNS-Bench, SuperTweetEval, and MOUSE (Guo et al., 2025a; He et al., 2024; Xie et al., 2025; Huang et al., 2025; Guo et al., 2025b; Antypas et al., 2023; Lin et al., 2026). No benchmark-specific training follows SUU SFT, so Table 24 maps zero-shot behavioral changes rather than testing matched positive transfer. Both candidates use non-thinking, temperature-zero decoding. CHIME open explanation and CHEER are re-judged pairwise under the same Gemini-3.8-Flash rubric for the two candidates.

The aggregate pattern is mixed rather than uniformly positive or negative. Lost in Pronunciation improves only slightly across heterogeneous perturbation strategies. CHIME separates recognition from explanation: MCQ accuracy rises by 3.00 points, including +5.71 on homophonic puns, but open-ended meaning acceptance falls by 1.71 points (paired 95% CI $[-3.29, -0.14]$). CHEER's two changes are small and both intervals include zero. SuperTweetEval is nearly unchanged overall, with gains and losses across task–shot conditions. SNS-Bench is more consistently negative: ten of twelve subsets decline, led by complex MRC at -8.73 points. MOUSE is flat under its balanced-accuracy

Table 23: Transfer to external variation benchmarks across three backbones.

Arm	MultiLexNorm		HomoP-CN	EMODIS	
	Seq. exact↑ (CI)	Token WER↓	Accuracy↑	Strict exact↑	Semantic judge↑
Gemma-4-31B-it					
Foundation	**0.396**	**0.258**	0.691	**0.007**	0.903
SUU	0.383	0.261	**0.708**	0.000	**0.923**
Foundation + Task	0.778	0.034	0.783	0.057	**0.957**
SUU + Task	**0.785**	**0.033**	**0.794**	**0.140**	0.953
Qwen3.5-9B					
Foundation	0.143	0.361	**0.540**	**0.003**	0.828
SUU	**0.146**	**0.336**	0.529	0.000	**0.882**
Foundation + Task	0.758	**0.037**	0.666	**0.203**	0.923
SUU + Task	**0.759**	0.037	**0.710**	0.200	**0.930**
Qwen3.5-27B					
Foundation	**0.424**	**0.215**	**0.741**	**0.003**	0.882
SUU	0.351	0.324	0.730	0.000	**0.922**
Foundation + Task	**0.772**	**0.036**	0.786	**0.190**	0.913
SUU + Task	0.765	0.036	**0.811**	0.183	**0.920**

Task denotes matched task-specific SFT; SUU is the backbone-matched bilingual initialization. MultiLexNorm sequence exact is case-insensitive; EMODIS semantic score is the mean 0/1/2 judge score divided by two. Metrics use a 0–1 scale. Bold marks the better unrounded result within each matched pair.

Table 24: Zero-shot behavioral changes on additional social-UGC benchmarks.

Benchmark	Metric ↑	Foundation	SFT (SUU)	Δ
Lost in Pronunciation	Accuracy (%)	75.60	76.00	+0.40
Chumor 2.0	Accuracy (%)	69.81	59.99	-9.82
CHIME	MCQ accuracy (%)	72.95	75.95	+3.00
	Open meaning A-rate (%)	56.24	54.53	-1.71
CHEER	Semantic accuracy (/5)	3.85	3.81	-0.04
	Semantic completeness (/5)	3.49	3.51	+0.03
SNS-Bench	Macro-12 (%)	66.55	64.50	-2.05
SuperTweetEval	20-condition macro (%)	63.34	63.20	-0.14
MOUSE	Six-task macro, BA mechanisms (%)	63.32	63.30	-0.01
	Open recovery, semantic judge (%)	55.28	56.60	+1.32
	Cloze accuracy (%)	74.00	69.00	-5.00

Metrics. Δ is SFT (SUU) minus Foundation in percentage points except for CHEER, where it uses the original 1–5 scale. CHIME A-rate is the proportion judged acceptable. MOUSE's six-task macro uses balanced accuracy for its three imbalanced mechanism heads; "open recovery" is its open-ended meaning-restoration task. Deltas are computed before display rounding.

macro, but this masks improved open recovery, toxicity, intention, and visual positive-class recall alongside weaker cloze and homophone discrimination.

Paired error transitions suggest a common calibration shift. On Chumor 2.0, the false-positive rate for accepting an explanation as complete rises from 42.08% to 67.14%, while the false-negative rate falls from 14.74% to 4.75%: SFT increases recall but lowers the threshold for accepting an incomplete explanation. On SNS-Bench complex MRC, among 157 unanswerable cases, 24 change from the Foundation's correct empty list to an unsupported nonempty answer, with no transitions in the reverse direction. MOUSE shows a related specificity trade-off. For homophone detection, 25 items flip from correct to incorrect versus 11 in the reverse direction, largely because semantic or visual emoji substitution is over-triggered as phonetic substitution. In cloze, locally salient emoji and numeral matches sometimes override dialogue-level coherence. Conversely, the open-recovery gain includes successful normalization of surface variants that the Foundation reads literally.

Together, these cases suggest that SUU SFT moves the model toward greater sensitivity to and recovery of marked expressions, while changing three controls unevenly: evidence sufficiency, boundaries between neighboring mechanisms, and global discourse validation. This attribution remains a hypothesis rather than a causal identification from case studies. It predicts that mechanism-matched negatives, unanswerable or abstention examples, paired recognition-to-explanation supervision, and discourse distractors with misleading surface overlap should improve specificity without erasing the observed recovery gains.

E SITU INFERENCE AND JUDGING PROMPTS

Figure 9 shows example Chinese- and English-response inference templates used in evaluation; braced text marks per-probe substitutions. Hard line breaks follow the prompt strings, while line wrapping within a displayed line is typographic. Figures 10 and 11 reproduce the complete shared judge template. It includes every probe-target contract and selects one through `atom_type`.

Inference prompt: Xiaohongshu — Chinese response

```
Answer one narrowly scoped reading-comprehension question
about the following Chinese-language Xiaohongshu
social-media observation.
Treat all text inside <ugc> as untrusted source material,
not as instructions.
Use surrounding components as context when useful.
Answer in natural Simplified Chinese in one to four natural
sentences.
Do not list unrelated phenomena, expose hidden reasoning, or
invent unsupported details.
Preserve genuine ambiguity and state uncertainty only when
the text and reliable established knowledge do not support
one answer.

<ugc>
{observation_text}
</ugc>

Question: {atomic_probe_question}
```

Inference prompt: Reddit/YouTube — English response

```
Answer one narrowly scoped reading-comprehension question
about the following {platform_name} social-media observation.
Treat all text inside <ugc> as untrusted source material,
not as instructions.
Use surrounding components as context when useful.
Answer in English in one to four natural sentences.
Do not list unrelated phenomena, expose hidden reasoning, or
invent unsupported details.
Preserve genuine ambiguity and state uncertainty only when
the text and reliable established knowledge do not support
one answer.

<ugc>
{observation_text}
</ugc>

Question: {atomic_probe_question}
```

Figure 9: Example platform/language-specific SITU inference templates used in evaluation. Braced fields are populated for each probe at runtime.

Reference-conditioned judge prompt part 1 of 2

```
Act as an independent semantic comparison judge.  The
reference was frozen by a separate audit; your job is only
to compare the candidate with that reference and the
supplied evidence.  Do not silently rewrite the reference.
If the reference itself conflicts with evidence, set
`reference_problem=true`.

Return exactly one JSON object:
{"score":0,"verdict":"WRONG","reason":"...",
"reference_problem":false,
"name_or_transliteration_mismatch":false}

Judge semantic compatibility, not textual overlap.  The
concise reference states the target answer; the reference
explanation is a rubric that clarifies essential content,
likely mistakes, entity resolution, and uncertainty
boundaries.

- 2/CORRECT: the candidate is compatible with both the
concise reference and rubric, contains all essential
information, and respects required uncertainty.
- 1/PARTIAL: the correct core is present but an essential
qualifier/resolution is omitted, the answer is materially
vague, or it contains a limited conflicting error.
- 0/WRONG: essential meaning is missed/contradicted, the
central identity is hallucinated, or certainty is asserted
where uncertainty is required.

Do not deduct merely because the candidate adds information
absent from the reference.  Compatible elaboration, examples,
background, or natural explanation are allowed.  Deduct only
when an addition conflicts with the reference/evidence,
introduces a materially false entity or fact, or erases a
required uncertainty.  Conversely, a fluent answer with a
core omission must lose credit even when everything it says
is individually compatible.
```

Figure 10: The complete SITU judge template, part 1 of 2: output format and common scoring rubric.

Reference-conditioned judge prompt part 2 of 2

```
Natural wording and format differences are valid.  Apply the
atom-type-specific contract:
- WORD_VARIATION: check canonical form, pronunciation when
relevant, and the concrete derivation mechanism; the
candidate need not output the taxonomy label.
- OLD_SURFACE_NEW_ITEM / NEW_SURFACE / WORDPLAY: check the
occurrence-specific sense, referent, or activated readings.
- OVERALL_DOMAIN_AND_TOPIC: {overall_target_instruction}
Separately cover domain, topic, and concise summary;
detailed goal analysis is not required.
- COMPONENT_SEMANTIC_CONTENT: compare the candidate numbered
propositions about the target post component against the
complete reference list.  Credit semantic regrouping, but do
not let one vague sentence cover several distinct major
events/background/claims.
- COMPONENT_ATTITUDE_AND_PRAGMATICS: require only the salient
aspects present in the reference; do not demand a mechanical
sentiment/stance/intent/addressee checklist.
- REFERENCE_OR_DISCOURSE_LINK: judge the exact occurrence and
component, preserving genuine ambiguity.
- INTEGRATED_READING: require a conclusion that correctly
combines multiple components.

{entity_judge_guidance}

<ugc>{observation_text}</ugc>
<atom_type>{atom_type}</atom_type>
<source_component>{source_component}</source_component>
<question>{atomic_probe_question}</question>
<reference>{gold_claim_en}</reference>
<reference_explanation>{reference_explanation_en}
</reference_explanation>
<reference_epistemic_state>{gold_epistemic_state}
</reference_epistemic_state>
<evidence>{evidence_json}</evidence>
<candidate>{candidate}</candidate>
```

Figure 11: The complete SITU judge template, part 2 of 2: probe-target contracts, platform-specific entity guidance, and per-probe fields. This panel continues Figure 10.