Psychologically Potent, Computationally Invisible: LLMs Generate Social-Comparison-Eliciting Posts They Fail to Detect

Anonymous ACL submission

Abstract

We introduce **X**iao**h**ongshu **S**ocial **Co**mparison **R**eader **E**licitation (XHS-SCoRE), a reader-grounded benchmark for detecting whether text-only Xiaohongshu (RedNote) posts elicit UPWARD, DOWNWARD, or NEUTRAL/no clear social comparison from a first-person reader perspective. The task targets a socially meaningful relational, behaviorally real signal not reducible to sentiment. Across prompted LLM classifiers and supervised Chinese encoders, we find a consistent generation–detection mismatch: the signal is textually learnable in-domain, but not robustly accessible to prompt-based classification. Prompted LLM classifiers show stable failures, especially neutralization of comparison-eliciting posts and model-specific directional skew. A controlled pilot shows that LLM-generated Xiaohongshu-style posts can shift perceived standing and comparison-related affect even when prompt-based detection of the same construct remains fragile. XHS-SCoRE contributes a benchmark for reader-grounded comparison detection and a diagnostic framework for studying when socially meaningful relational cues remain only partially visible to prompt-based inference.

1 Introduction

Social media is an always-on setting for social comparison: users infer standing by comparing with others (Festinger, 1954). Attention-optimizing feeds make ordinary posts implicit benchmarks (Wu et al., 2024). On visual, lifestyle-oriented platforms, achievements, bodies, consumption, and family narratives become "rankable" (Jabłońska and Zajdel, 2020; Valkenburg and Peter, 2011), including Xiaohongshu (Gu et al., 2026). Such cues can shift self-evaluations and emotions, contributing to anxiety, dissatisfaction, envy, rumination and negative affect (McComb et al., 2023; Xu and Li, 2024; Buunk et al., 1990; Collins, 1996). In contemporary China, "involution" and "lying flat" reframe comparison as participation in performance norms (Wang et al., 2024; Deng et al., 2025).

This creates a distinct NLP problem: can a model recover reader-perceived social-comparison direction in naturally occurring posts? On Xiaohongshu, relevant cues are often relational, implied, and culturally grounded, not lexically explicit. The task is first-person comparison elicitation, not sentiment or overt-evaluation detection. The question is urgent because generative AI can produce persuasive, affect-calibrated social-media text at scale (Salvi et al., 2025). Recent NLP work also ties platform discourse to model development, including domain-specific post-training and LLMs as instruments in computational social science (Zhao et al., 2025; Ziems et al., 2024). Together, these trends create a risk: models may generate comparison-eliciting content for target readers yet fail to recover the same relational meaning in everyday posts, erasing or distorting socially grounded meaning in auditing, moderation, and measurement (Hovy and Spruit, 2016).

This paper makes four contributions. We 1) introduce Xiaohongshu Social Comparison Reader Elicitation (XHS-SCoRE), a large reader-grounded benchmark for comparison direction in text-only Xiaohongshu posts; 2) show that this signal is textually learnable for strong in-domain encoders but unreliably accessible to prompted LLM classification, revealing a generation–detection gap; 3) identify structured failures–neutralization and directional skew–across prompting regimes; and 4) pair corpus-constrained generation with a controlled human pilot showing that LLM-generated texts can instantiate comparison-relevant signals in readers even when prompt-based detection remains fragile. XHS-SCoRE contributes a benchmark and reusable error-analysis template for cases where generation fluency does not imply reliable detection of reader-grounded relational meaning.

2 Related Work

2.1 Social Comparison and Reader-Perceived Direction

Social comparison theory treats self-evaluation as relational (Festinger, 1954), and platform studies show that social media intensifies comparison by making targets abundant, salient, and decontextualized (Appel et al., 2016; Fardouly and Vartanian, 2015). On lifestyle-oriented platforms, comparison cues are often embedded in ordinary narration rather than explicit comparatives, so direction must be inferred from stance, outcomes, and what appears "normal" in the feed. This is especially salient on Chinese social media (incl. Xiaohongshu), where family life, consumption, schooling, and peer success are narrated in socially rankable ways without overt markers (Wang et al., 2024; Gu et al., 2026). The NLP gap is modeling reader-perceived direction when relational positioning is implied.

2.2 NLP Measurement and Reader-Centered Meaning

NLP has long inferred psychological and social variables from text, from lexicon-based proxies to supervised models for mental-health signals, temporal dynamics, and socially grounded labels such as toxicity or personal attack (Tausczik and Pennebaker, 2010; De Choudhury et al., 2013; Coppersmith et al., 2015; Wulczyn et al., 2017). Open-vocabulary studies associate demographic and psychological variables with recurring word, phrase, and topic patterns (Schwartz et al., 2013), while work on implicit social implications models pragmatic frames projected from social-media language (Sap et al., 2020). This work establishes text-based social measurement, but often treats labels as author- or text-centered targets with stable correlates. Subjective-annotation work likewise cautions that disagreement can encode annotator perspective rather than majority-vote noise (Mostafazadeh Davani et al., 2022). Social comparison direction departs from these assumptions: it is reader-centered and often realized through pragmatic roles, agency framing, evaluative listing, reported dialog, and platform-native registers of aspiration or hardship rather than direct lexical anchors. The gap is measurement form: existing NLP targets provide limited tools for recovering comparison cues as reader-grounded direction in the wild.

2.3 LLM Detection Reliability Versus Generative Potency

LLMs are increasingly used as classifiers for socially grounded labels, yet reliability is uneven when labels depend on pragmatics, implicit norms, or culturally embedded meaning (Ziems et al., 2024; Sravanthi et al., 2024). This difficulty is sharper for reader-centered constructs such as comparison direction. At the same time, LLMs generate socially plausible, emotionally calibrated text that can be persuasive (Salvi et al., 2025), hard to distinguish from platform-native writing (Dugan et al., 2024), and tied to platform discourse through domain-specific post-training and evaluation (Zhao et al., 2025). Underexplored is whether models that realize a socially consequential construct in generation can also recover it through prompted classification. XHS-SCoRE addresses this gap for reader-grounded comparison direction in natural posts.

3 Task Definition

Reader-perceived social comparison direction is a three-way task: given a text-only Xiaohongshu post x and reader profile r, predict one label $y \in \{\textbf{UPWARD}, \textbf{NEUTRAL}, \textbf{DOWNWARD}\}$, corresponding to the immediate first-person comparison response. XHS-SCoRE operationalizes r through a relatively homogeneous reader group and standardized collection protocol. The benchmark is therefore a reader-group-conditioned elicitation task, not author-intent detection or population-invariant semantic labeling. A label is valid in this setup because it records the original collector's immediate first-person response under text-only browsing; stability and in-group agreement calibrate this reader-grounded target via within-reader consistency and partial convergence among similar readers, rather than replacing it with a consensus semantic label.

Operationally, UP denotes posts that position the poster as better off, more successful, more resourced, or otherwise advantaged relative to the target reader, inviting upward comparison; typical cases are idealized travel, consumption, or high-achievement posts that make the poster's situation aspirational. DOWN denotes posts that position the poster as worse off, more distressed, less resourced, lower-agency, or otherwise disadvantaged, inviting downward comparison; typical cases are family-conflict or hardship narratives that make the reader feel relatively better situated. NEU denotes

posts that do not clearly invite reader-poster comparison, such as weather updates, product rankings, tutorials, or advertisements without clear self–other positioning. NEU is a substantive non-comparison label, not abstention or model uncertainty. A post may be affectively positive or negative while labeled NEU, and conversely may elicit comparison without overt evaluative language. Because XHS-SCoRE is nearly class-balanced, we use Accuracy and Macro-F1 as primary metrics and analyze recall, confusion structure, and predicted-label distributions to diagnose direction-specific failures. Section 4.4 provides full examples.

4 Dataset

4.1 Collection protocol and labeling principle

XHS-SCoRE (**X**iao**h**ongs**h**u **S**ocial **Co**mparison **R**eader **E**licitation) consists of text-only Xiaohongshu posts collected by young adult users under a standardized browsing protocol. Inclusion and labeling required that 1) all items come from Xiaohongshu; 2) meaning is recoverable from text alone (excluding posts requiring images or video); and 3) labels reflect the collector's immediate reader reaction during browsing. Participants judged whether the post elicited no comparison or made the poster seem better or worse off than themselves. Instructions gave rubric-aligned examples: a weather post for NEU/no clear comparison, a high-quality travel or lifestyle post for UP, and a family-hardship post for DOWN. Participants were told reactions may differ and should reflect their own immediate feeling. The protocol treats elicitation as partly reader-dependent within a homogeneous group, targeting reader-perceived comparison elicitation rather than author intent.

4.2 Annotation regime and reliability

XHS-SCoRE uses a single-collector immediate-response protocol to preserve the first-person, reader-based construct. Each participant browsed Xiaohongshu on their own device for 7 days and identified 10 posts per label (UP, DOWN, and NEU/no clear comparison) each day, yielding a target of 210 posts. They copied text-only content into structured spreadsheets for daily submission. Participants were students ($N = 67$; $M_{\text{age}} = 21.75$; 85% female; native Chinese) from 8 publicly funded HK universities and active Xiaohongshu users. This homogeneous sampling frame constrains reader heterogeneity by partly controlling platform familiarity and shared background; individual moderators such as aspiration orientation or personal definitions of success are not modeled here. We adopt the single-collector design because the task targets reader-perceived comparison elicitation under naturalistic browsing, rather than author intent or a population-invariant label.

To calibrate this reader-grounded operationalization, we report three complementary reliability signals under the same text-only constraint as modeling (Table 1). On a randomized set of 210 posts, two additional raters ($N = 2$; both female; $M_{\text{age}} = 21.50$; native Chinese young-adult active Xiaohongshu users) from the same target reader population applied the same immediate-response scheme; their 63.2%-68.4% raw agreement with original immediate-response labels conservatively estimates between-reader consistency without multimodal cues. Repeated entries at least 24 hours apart showed ∼82% within-person stability. The raters also re-labeled 20.4% of originally UP and 17.0% of originally DOWN posts as NEU, giving a human ambiguity baseline for model-side NEU collapse. These calibration results contextualize the single-reader target; they do not overwrite original immediate-response labels.

Calibration Signal	Value
In-group agreement (text-only)	63.2%–68.4%
Within-person stability (24h+)	∼82%
Human UP→NEU	20.4%
Human DOWN→NEU	17.0%

Table 1: Reliability calibration for XHS-SCoRE.

4.3 Benchmark size and splits

XHS-SCoRE is label-balanced, with 2,452,665 Chinese characters and 13,916 posts: 4,632 UP, 4,631 NEU, and 4,653 DOWN. Posts are randomized into fixed splits for all comparisons. The TRAIN split has 8,350 posts (2,780 UP; 2,779 NEU; 2,791 DOWN; 1,487,712 characters). The VAL split has 2,783 posts (926 UP; 926 NEU; 931 DOWN; 496,996 characters). The TEST split has 2,783 posts (926 UP; 926 NEU; 931 DOWN; 467,957 characters).

Class balance makes high performance less likely to reflect majority-class behavior and makes Macro-F1 interpretable as directional reliability rather than label-frequency exploitation. See Appendix A for non-raw, policy-compliant artifacts (label schema, prompts, scripts, aggregated results, and AI-generated examples).

4.4 Corpus analysis

To characterize directional realizations, we use Wmatrix 7 keyness analysis over word, part-of-speech, and semantic tags with log-likelihood statistics, followed by concordance inspection and inductive frame analysis (Rayson, 2008). In brief, UP posts more often realize aspirational lifestyles through consumption, mobility, and positive evaluation, whereas DOWN posts more often realize low-agency, conflict-centered narratives with heavier negation, pronouns, reported speech, and passive constructions. These corpus-derived cues inform stimulus construction and error interpretation.

Illustrative examples (anonymized, shortened). UP. "亚庇｜拍到了人生照片...日落、...、烟花，一切美好的元素都在这一刻集齐了。" Idealized, high-aesthetic peak experience → the poster's "perfect moment" as benchmark.

DOWN. "...大到人生选择，生活方式...小到今天穿什么衣服，做什么事...在我父母那里得到的永远是反对，贬低与道德绑架..." Low-agency, emotionally constrained family narrative → readers' relatively better domestic autonomy.

NEU. "AI最新排行榜｜夸克增速狂飙...夸克这波出息了，appstore排名超越元..." Third-party products/rankings, not poster status/agency/experience → no clear comparison.

These examples show that the task targets reader-perceived comparison direction, not sentiment polarity alone: the criterion is whether the post positions the poster relative to the reader in a way that invites upward, downward, or no clear comparison. Full texts, translations, and linguistic interpretations are in Appendix D.

5 Models and Experimental Setup

5.1 Prompted LLM classification settings

We evaluate LLMs as prompted three-way classifiers. The primary zero-shot, first-person prompt aligns with the benchmark's reader profile, uses only post text, and requires a single JSON label. Prompts are in Simplified Chinese. We set temperature=.1 for all models to reduce sampling variability while preserving API executability; for gpt-5-2025-08-07, reasoning.effort=minimal limits unnecessary reasoning tokens. Full prompt templates appear in Appendix B.

Beyond zero-shot, we evaluate persona-primed, few-shot, and cue-explicit prompts as robustness conditions. These variants preserve the text-only design and label inventory, testing whether error structure depends on one prompt formulation. Cue-explicit prompting is a diagnostic robustness probe rather than a routine classifier prompt: it incorporates a manually abstracted cue inventory from prior corpus-linguistic analysis (aspiration/abundance for UP, conflict/low-agency for DOWN, and informational/no-positioning for NEU). Gains test whether comparison direction becomes more recoverable with construct-specific scaffolding, not whether default prompted classification is reliable. Unless otherwise noted, the main comparison in Sections 6.1–6.3 uses zero-shot; Section 6.4 summarizes robustness across regimes.

We test four LLMs across capability and cost tiers to assess robustness of failure modes:

1. **GPT-5** (closed, frontier) with explicit reasoning-control parameters (OpenAI, 2025a).
2. **Qwen3-235B-A22B-Instruct** (open, large MoE): SOTA open comparison (Yang et al., 2025).
3. **Qwen3-30B-A3B-Instruct** (open, smaller MoE) to test scale sensitivity (Yang et al., 2025).
4. **GPT-4.1 nano** (low-cost) as a deployment-oriented model and stimulus-generation family, enabling direct generation–classification comparison (OpenAI, 2025b).

5.2 Supervised encoder baselines

To contrast prompted and supervised inference, we fine-tune three Chinese encoder-only models: hfl/chinese-bert-wwm-ext, hfl/chinese-roberta-wwm-ext, and hfl/chinese-macbert-base. These represent the BERT encoder paradigm (Devlin et al., 2019), a RoBERTa variant (Liu et al., 2019), and MacBERT's masking-as-correction adaptation for Chinese (Cui et al., 2020). Each model is fine-tuned on TRAIN with a classification head for up to 15 epochs from pre-trained weights of Cui et al. (2020); we select the best checkpoint by VAL Macro-F1 and report on TEST. Full hyperparameters, environment, and grid-search details are in Appendix E.

6 Classification Results

6.1 Main results on the test split

Table 2 establishes that the text-only directional signal is learnable in-domain for supervised encoders but unreliable under prompted LLM inference. Figure 1 shows a structural gap: several models absorb comparison-eliciting posts into NEU, whereas Qwen3-235B exhibits a marked UP-skew profile.

Model	Type	Acc	Macro-F1	Rec UP	Rec NEU	Rec DOWN	Pred (as) NEU
GPT-5	LLM	0.521	0.518	0.410	0.752	0.402	0.601
Qwen3-235B	LLM	0.491	0.480	0.670	0.522	0.282	0.425
GPT-4.1-nano	LLM	0.469	0.469	0.379	0.630	0.397	0.558
Qwen3-30B	LLM	0.430	0.400	0.364	0.748	0.179	0.659
CN-BERT WWM	Encoder	0.670	0.671	0.666	0.636	0.708	0.360
CN-RoBERTa WWM	Encoder	0.680	0.679	0.695	0.585	0.759	0.307
CN-MacBERT Base	Encoder	0.665	0.665	0.633	0.631	0.730	0.349

Table 2: Main test performance of the primary zero-shot prompted LLM classifiers and supervised encoder baselines.

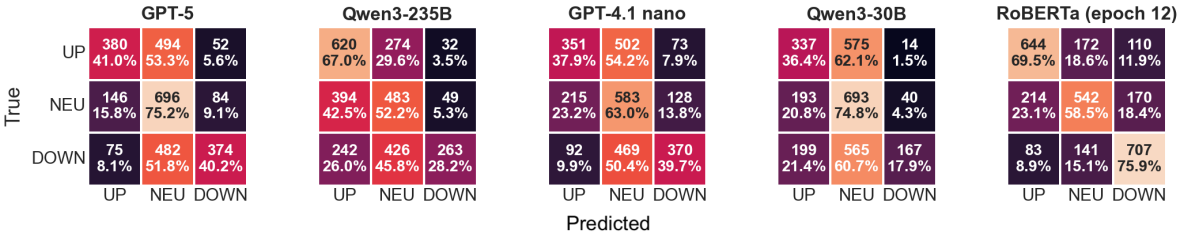

Figure 1: Confusion matrices (test): primary zero-shot LLM vs best encoder baseline. % row-normalized.

The core result is interpretable failure tied to the task's relational structure, not just weaker prompted classification. Section 6.4 shows broad stability under stronger prompting variants.

6.2 Class-level behavior and the DOWN sensitivity gap

DOWN is consistently harder for prompted LLM classifiers to recover than UP. Models often collapse hardship- or disadvantage-framed posts into NEU, or misread directionally, even when encoder baselines recover them better. This identifies systematic weakness in one side of the comparison space and would make monitoring or auditing systems miss important downward-comparison cues.

6.3 When comparison becomes invisible: error patterns beyond aggregate scores

6.3.1 Neutralization of comparison-eliciting posts

The pervasive failure mode is neutralization: posts that invite comparison are repeatedly assigned to NEU. This pattern appears across several prompted LLM settings and exceeds the human ambiguity baseline from Section 4.2. The error is severe because NEU is an explicit non-comparison judgment: it asserts no clear self–other ranking, rendering comparison cues computationally absent.

6.3.2 Directionality asymmetry and DOWN sensitivity

Neutralization is not directionally uniform. Some models wash out original DOWN items more strongly than original UP items, so the risk is directionally selective under-detection, not only under-detection of comparison. Since direction is part of the construct, this is a substantive failure of construct recovery, not a secondary labeling issue.

6.3.3 A distinct profile: UP over-attribution

Qwen3-235B shows a different but structured profile: rather than defaulting mainly to NEU, it over-attributes UP. This shows prompted failure is not reducible to a single "cautious classifier" pattern; different models distort in different ways.

6.3.4 Summary: "invisibility" as predictable collapse, not random noise

The central problem is disappearance, not simply error: comparison-eliciting posts are repeatedly translated into non-comparison. Because NEU asserts no clear self–other ranking, such errors do more than weaken construct recovery; they erase it at model output. The result is systematic computational invisibility for reader-grounded comparison cues under prompted detection.

6.4 Prompt robustness and stability under alternative prompting regimes

Because prompted classification can be sensitive to decoding and prompt stochasticity, we test primary zero-shot stability with same-condition reruns. Agreement with the original run is high and confusion-matrix drift is low, indicating that qualitative zero-shot error patterns are stable under the tested prompting regime rather than one-off sampling noise. Full statistics appear in Table A1 in

Appendix A.

We then evaluated three prompt variants: few-shot, persona-primed, and cue-explicit. Table 3 summarizes their behavior alongside the human text-only ambiguity baseline and strongest supervised in-domain reference. A compact prompt-family summary (Table B1) and verbatim prompts are in Appendix B.

Table 3 shows clear but non-uniform robustness. Cue-explicit prompting is the strongest tested prompt condition and sharply reduces neutralization for some models. This setting should not be read as ordinary or default LLM classification. Unlike zero-shot, persona-primed, or few-shot prompting, the cue-explicit prompt depends on substantial prior corpus-linguistic analysis and manual abstraction of comparison-relevant cues into an explicit rubric. Its gains show that comparison-relevant information is present and can be made more accessible, not that routine prompted detection is adequate. The signal is recoverable, but common prompting settings fail to access it robustly without unusually strong external scaffolding.

Condition	**F1**	**PredN**	$U \to N$	$D \to N$
Human re-rating	–	–	20.4	17.0
RoBERTa	67.9	30.7	18.6	15.1

Model	**Metric**	**ZS**	**PP**	**FS**	**CE**
GPT-5	F1	51.8	51.6	50.2	55.1
	PredN	60.1	65.9	59.4	33.4
	$U \to N$	53.3	62.2	47.8	26.0
	$D \to N$	51.8	51.9	56.2	27.8
GPT-4.1nano	F1	46.9	38.0	38.1	44.9
	PredN	55.8	77.4	69.8	68.2
	$U \to N$	54.2	79.9	62.0	68.6
	$D \to N$	50.4	65.8	72.0	57.8
Qwen-235B	F1	48.0	46.0	51.4	54.8
	PredN	42.5	66.7	54.5	30.8
	$U \to N$	29.6	50.7	44.8	24.6
	$D \to N$	45.8	68.0	49.0	24.0
Qwen-30B	F1	40.0	47.8	46.9	53.7
	PredN	65.9	49.1	39.7	28.0
	$U \to N$	62.1	40.6	31.2	25.7
	$D \to N$	60.7	46.1	36.5	19.6

Table 3: Neutralization under alternative prompting regimes. ZS: zero-shot, PP: persona-primed, FS: few-shot, CE: cue-explicit. Human re-rating: text-only ambiguity baseline, RoBERTa: supervised in-domain. $U \to N$ and $D \to N$ indicate the percentage of original UP and DOWN items predicted as NEU; PredN is the overall percentage of predictions assigned to NEU.

The overall pattern is clear. Some neutralization is expected even under human text-only re-reading, but several zero-shot LLM settings exceed that ambiguity baseline, especially for original DOWN posts. Cue-explicit prompting reduces this effect for GPT-5 and Qwen3-235B/30B, yet recovery still trails the encoder baseline. The problem is not lower performance alone, but a recurrent tendency to predict comparison-eliciting language as non-comparison. Hence, the construct is not absent from the text; it becomes computationally invisible under routine prompting.

7 Controlled LLM Generation of Social-Comparison Stimuli

To test the "psychologically potent" side of the dissociation, we generate Xiaohongshu-style stimuli with GPT-4.1 nano under corpus-derived constraints. The goal is construct-targeted generation rather than generic realism: UP and DOWN differ in stance, agency framing, evaluative lexis, and abundance-vs-hardship cues while remaining plausible as platform posts, whereas NEU suppresses self-other positioning and affective framing. Full prompts and constraint lists are in Appendix B, with stimulus examples in Appendix C.

8 Human Validation: LLM-Generated Posts Elicit Social Comparison

Because comparison direction is reader-perceived rather than purely text-internal, we conduct a lean human study to validate the potency side of the dissociation. We ask whether LLM-generated posts, produced under corpus-derived constraints and minimally edited for naturalness, elicit the intended comparison direction and downstream affective responses; the study is construct validation rather than a standalone psychological contribution.

8.1 Participants, design, and procedure

Participants ($N = 29$; recruited under the same basic eligibility requirements as the corpus collectors, with no overlap) were young adult active Xiaohongshu users who completed the computer-lab study after informed consent. The experiment was implemented in jsPsych (de Leeuw, 2015) with a between-subject design: UP ($N = 10$), DOWN ($N = 9$), and a NEU control ($N = 10$). Assignment used a pre-generated randomization plan to maintain balance under small N.

After demographics and baseline measures, participants read seven short posts from their assigned condition in randomized order. After each post, they completed manipulation checks measuring perceived relative standing and self–other similarity, then post-exposure measures of comparison-

related emotions (Smith, 2000) and general affect (Watson et al., 1988). They then completed an attention check and debrief disclosing the study purpose and AI-generated nature of the stimuli, with support resources provided as needed.

8.2 Results: manipulation success and affective potency

Directional manipulation check. A regression of perceived standing on condition shows a strong effect (adjusted $R^2 = .570$, $p < .001$), indicating readers infer the intended comparison direction from generated posts.

Comparison-related emotions. DOWN assimilative emotion is substantially higher in the DOWN condition than in the other two conditions combined (DOWN: M=6.33, SD=1.41; Others: M=1.85, SD=1.90), $t(27) = 6.312$, $p < .001$, $d = 2.534$. Undesirable comparison-related emotions differ across conditions, $F(2, 26) = 14.26$, $p < .001$, $\eta_p^2 = .52$, with a monotonic pattern: DOWN > NEU > UP. General affect also shifts (positive affect: $F(2, 26) = 5.942$, $p = .007$, $\eta_p^2 = .314$; negative affect: $F(2, 26) = 3.616$, $p = .041$, $\eta_p^2 = .218$).

Taken together, the pilot supports potency in the precise sense needed here: the texts shift perceived standing and induce measurable affect aligned with comparison direction.

8.3 Link to NLP: potency vs detectability

The human results ground the label space behaviorally: when the generator produces UP, DOWN, or NEU stimuli under corpus-derived constraints, readers infer the intended self–other standing relation and show corresponding shifts in comparison-related emotion, clarifying the classification stakes. When a detector labels a comparison-eliciting post as NEU, it is not merely making a small semantic error; it assigns a substantive non-comparison label to content that can induce standing judgments and emotion profiles.

A diagnostic closed-loop check sharpens this interpretation. We classified the 21 generated stimuli with the same prompted LLM classifiers: Qwen3-235B and GPT-4.1-nano classified all generated stimuli correctly, GPT-5 misclassified one item, and Qwen3-30B achieved 71.4% accuracy and 0.70 Macro-F1. The generated items were produced under explicit corpus-derived constraints, so they are cleaner and more scaffolded than natural posts. Thus, the issue is not that prompted models cannot apply the label schema when comparison direction is deliberately instantiated; failure appears when recovering reader-grounded cues from natural platform text, where the signal is implicit, culturally situated, and pragmatically mixed.

This is how comparison becomes computationally invisible: socially consequential relational meaning is present in text as an elicited signal, yet collapses under model-based detection, especially when expressed through implicit stance and narrative positioning rather than explicit comparatives. Together with controlled generation, the pilot supports the central dissociation: LLMs can produce posts that shift human judgments and emotions while failing to reliably detect the same cues in natural posts. This asymmetry creates a risk for pipelines using prompted LLMs to audit, monitor, or moderate comparison-related harms: systems may scale psychologically potent comparison-eliciting content while under-detecting its presence and direction in real platform text.

9 Discussion

9.1 The generation–detection dissociation

Our findings isolate a dissociation within the tested text-only Xiaohongshu setting and prompting regimes. LLMs can generate platform-style posts that shift perceived standing and comparison-related affect, yet prompted LLM classifiers remain unreliable detectors of the same reader-grounded construct in natural posts. The failure is structured: some models collapse comparison-eliciting posts into NEU, while others skew toward the wrong direction. Because supervised encoders recover the signal substantially better, the central issue is not absence of learnable textual cues, but instability of prompt-based detection under the tested regimes.

The gap is methodological as well as empirical: the paper identifies a mismatch between behaviorally consequential meaning and robust prompt-based access, not just one detector family underperforming another. That mismatch turns a broad concern about "LLMs being worse" into a precise account of where reader-grounded inference fails and how that failure is structured.

9.2 Evaluation: generation quality is not detection reliability

Generation fluency is not evidence of reliable detection for reader-grounded relational meaning. In the pilot, corpus-constrained LLM-generated posts

shifted perceived standing and comparison-related emotions, showing that models can realize the construct in generation. Yet when classifying natural posts into UP, NEU, or DOWN, the same family remained stably error-prone. The relevant contrast is not generation versus "understanding" in the abstract, but construct realization in text versus construct recovery under prompted detection.

That contrast has a direct evaluation consequence. For socially embedded constructs, aggregate accuracy alone is insufficient; the target includes where meaning disappears and how it is distorted. In XHS-SCoRE, neutralization and directional skew dominate. For reader-grounded meaning tasks, confusion structure is part of the phenomenon to explain, not merely an auxiliary diagnostic.

9.3 Governance: scalable elicitation, unreliable monitors

The XHS-SCoRE error profile matters because prompted LLM failures are structured in application-relevant ways. Comparison-eliciting posts are repeatedly absorbed into NEU or displaced toward the wrong direction, creating systematic blind spots for downstream prompt-based detectors. In platform monitoring, auditing, or intervention design, the cues most worth tracking may be those most likely to be washed out. Detecting comparison-eliciting cues is not a by-product of fluency or generic classification strength; it is a separate capability requiring direct validation.

9.4 Social science methods: controlled generation, human grounding

The dissociation identified here is methodological as well as empirical. The pilot shows that LLM-generated Xiaohongshu-style texts can serve as controlled materials that shift perceived standing and comparison-related affect. This makes detection failure more informative: models can help realize the construct in generation while remaining unreliable instruments for classifying it in natural posts.

Stimulus usefulness, human grounding, and detector reliability are not interchangeable signs of model competence. XHS-SCoRE forces those dimensions apart and makes their relationship empirically testable. For LLM-assisted social-science workflows, generated materials may be usable under human validation, but detector outputs require task-specific evaluation before being treated as measures of reader-grounded meaning.

In this sense, XHS-SCoRE contributes both a benchmark and a diagnostic logic. Empirically, it isolates a reader-grounded relational meaning task that is behaviorally real, textually learnable, and socially consequential, yet not robustly recovered by prompted LLM classification under the tested regimes. The central result is sharper than one detector family underperforming another: models can generate texts that elicit social-comparison responses while failing to detect the same signal reliably in natural platform text. While specific to text-only Xiaohongshu posts and the sampled reader population, the transferable contribution is an evaluation logic for reader-perceived constructs that are behaviorally real, group-dependent, and textually learnable, yet fragile under ordinary prompted inference. For such constructs, generation fluency and controlled stimulus usefulness should not be treated as evidence of reliable detection in naturalistic data.

10 Conclusion

We introduced Xiaohongshu Social Comparison Reader Elicitation (XHS-SCoRE), a reader-grounded benchmark for detecting whether text-only Xiaohongshu posts elicit UP, DOWN, or no clear comparison from a first-person reader perspective. The benchmark reveals a consistent generation–detection gap: supervised Chinese encoders recover comparison direction more reliably, whereas prompted LLM classifiers show stable distortions dominated by neutralization and directional skew. A controlled human pilot further shows that LLM-generated Xiaohongshu-style texts can shift perceived standing and comparison-related affect even when prompt-based detection remains fragile. Together, these results identify a reader-grounded meaning dimension that is behaviorally consequential and textually recoverable, yet only partially accessible to prompt-based inference.

Taken together, the results support a narrow but consequential conclusion: in this text-only, reader-grounded Xiaohongshu setting, LLMs can generate psychologically potent texts that elicit social-comparison responses in the target reader group while those cues remain only partially visible to prompt-based detection.

Limitations

This study is a text-first benchmark and diagnostic baseline, not a general account of social-

comparison detection across platforms or reader populations. First, XHS-SCoRE uses text-only posts drawn from a strongly multimodal platform, so images, video, layout, and engagement context may alter comparison elicitation in ways not captured here. Second, labels are defined relative to a specific reader profile and collection regime, not as population-invariant labels. Third, the behavioral validation is a pilot study (N=29) for construct grounding rather than broad external validity. Fourth, although we include stronger prompting baselines and robustness checks, the detector conclusions remain scoped to the tested prompting regimes rather than all possible prompt designs or future model versions. The contribution is best understood as a strong benchmark, a measurement-calibrated detector comparison, and an error-analysis template for multimodal, cross-platform, and more heterogeneous-reader settings.

Ethics Statement

Human subjects protections. Participants provided informed consent and were recruited via university advertisements. The study used a cover story to reduce demand characteristics for the social-comparison manipulation. Participants were debriefed at the end of the study with disclosure of the true purpose and the AI-generated nature of the stimuli. Support resources were made available in case the content elicited discomfort or distress. Ethical approval was obtained from the university ethics board prior to the research.

Data privacy and platform policy compliance. All collected and processed data were handled with care to minimize privacy risk. The dataset is not released in raw form due to platform policies and the potential for re-identification or unauthorized redistribution. Any released artifacts are designed to be policy-compliant and privacy-preserving (e.g., paraphrased examples, aggregated statistics, code and evaluation scripts) and avoid exposing usernames, personal identifiers, or direct reproductions of user content.

Dual-use considerations. This work demonstrates that LLMs can generate posts that are psychologically potent with respect to social comparison. Such capability could be misused to mass-produce content that manipulates comparison emotions or exacerbates distress. We mitigate this risk by (i) limiting the release of high-fidelity generation recipes to what is necessary for scientific transparency, (ii) framing the generation component as controlled construct validation rather than a "how-to" guide, and (iii) emphasizing that deployment-facing systems should not treat prompted LLM detection as sufficient for risk monitoring. Where appropriate, we recommend platform-facing evaluation focus on detecting comparison-eliciting cues and assessing false-negative rates arising from neutralization.

Fairness and vulnerable populations. Social comparison harms can disproportionately affect vulnerable users, including youth and individuals with elevated anxiety or depressive symptoms. While this paper does not stratify results by demographic vulnerability, the findings motivate targeted auditing: platforms and researchers should assess whether detection failures and exposure risks are unevenly distributed across user groups and content topics. Human-grounded evaluation is especially important when interventions affect minors or psychologically sensitive populations.

References

Helmut Appel, Alexander L Gerlach, and Jan Crusius. 2016. The interplay between facebook use, social comparison, envy, and depression. *Current Opinion in Psychology*, 9:44–49. Social media and applications to health behavior.

B. P. Buunk, R. L. Collins, S. E. Taylor, N. W. VanYperen, and G. A. Dakof. 1990. The affective consequences of social comparison: either direction has its ups and downs. *J Pers Soc Psychol*, 59(6):1238–49. Buunk, B P Collins, R L Taylor, S E VanYperen, N W Dakof, G A CA 36409/CA/NCI NIH HHS/United States MH 00311/MH/NIMH NIH HHS/United States MH 42258/MH/NIMH NIH HHS/United States etc. Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S. United States 1990/12/01 J Pers Soc Psychol. 1990 Dec;59(6):1238-49. doi: 10.1037//0022-3514.59.6.1238.

Rebecca L. Collins. 1996. For better or worse: The impact of upward social comparison on self-evaluations. pages 51–69.

Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015. From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 1–10, Denver, Colorado. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained

models for Chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 657–668, Online. Association for Computational Linguistics.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, page 3267–3276, New York, NY, USA. Association for Computing Machinery.

Joshua R. de Leeuw. 2015. jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, 47(1):1–12.

Yiheng Deng, Alexander Scott English, and Yuting Li. 2025. New elements of career construction for china's youth: Analyzing 'lying flat' and work involution among emerging adults. *Emerging Adulthood*, 13(1):131–145.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. RAID: A shared benchmark for robust evaluation of machine-generated text detectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12463–12492, Bangkok, Thailand. Association for Computational Linguistics.

Jasmine Fardouly and Lenny R. Vartanian. 2015. Negative comparisons about one's appearance mediate the relationship between facebook usage and body image concerns. *Body Image*, 12:82–88.

Leon Festinger. 1954. A theory of social comparison processes. *Human Relations*, 7(2):117–140. Doi: 10.1177/001872675400700202.

Michelle Mingyue Gu, Shuting Zhang, Yue Zhang, Hua Zhao, Jiapei Gu, Jincheng Ding, Xiaofei Lu, and Chetwyn Che-hin Chan. 2026. The emotional valence paradox in social media: A computational linguistic analysis of how sentiment, emotion, and content domain predict social comparisons on rednote. *International Journal of Applied Linguistics*, n/a(n/a).

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

M. R. Jabłońska and R. Zajdel. 2020. Artificial neural networks for predicting social comparison effects among female instagram users. *PLoS One*, 15(2):e0229354. 1932-6203 Jabłońska, Marta R Orcid: 0000-0001-6004-6228 Zajdel, Radosław Journal Article United States 2020/02/26 PLoS One. 2020 Feb 25;15(2):e0229354. doi: 10.1371/journal.pone.0229354. eCollection 2020.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Carly A. McComb, Eric J. Vanman, and Stephanie J. Tobin. 2023. A meta-analysis of the effects of social media exposure to upward comparison targets on self-evaluations and emotions. *Media Psychology*, 26(5):612–635. Doi: 10.1080/15213269.2023.2180647.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

OpenAI. 2025a. Gpt-5 system card. System card, OpenAI.

OpenAI. 2025b. Introducing gpt-4.1 in the api. Accessed: 2026-01-05.

Paul Rayson. 2008. From key words to key semantic domains. *International Journal of Corpus Linguistics*, 13(4):519–549.

Francesco Salvi, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West. 2025. On the conversational persuasiveness of gpt-4. *Nature Human Behaviour*, 9(8):1645–1653.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, and Lyle H. Ungar. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLOS ONE*, 8(9):1–16.

Richard H. Smith. 2000. *Assimilative and contrastive emotional reactions to upward and downward social comparisons*, pages 173–200. The Plenum series in social/clinical psychology. Kluwer Academic Publishers.

Settaluri Sravanthi, Meet Doshi, Pavan Tankala, Rudra Murthy, Raj Dabre, and Pushpak Bhattacharyya. 2024. PUB: A pragmatics understanding benchmark for assessing LLMs' pragmatics capabilities. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12075–12097, Bangkok, Thailand. Association for Computational Linguistics.

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Patti M. Valkenburg and Jochen Peter. 2011. Online communication among adolescents: An integrated model of its attraction, opportunities, and risks. *Journal of Adolescent Health*, 48(2):121–127.

Feng Wang, Yanchao Yang, and Tianxue Cui. 2024. Development and validation of an academic involution scale for college students in china. *Psychology in the Schools*, 61(3):847–860. (Robin).

D. Watson, L. A. Clark, and A. Tellegen. 1988. Development and validation of brief measures of positive and negative affect: the panas scales. *J Pers Soc Psychol*, 54(6):1063–70. Watson, D Clark, L A Tellegen, A Journal Article United States 1988/06/01 J Pers Soc Psychol. 1988 Jun;54(6):1063-70. doi: 10.1037//0022-3514.54.6.1063.

Qingyue Wu, Lei Gu, Mingxiao Zhang, and Huimei Liu. 2024. Understanding dual effects of social network services on digital well-being and sustainability: A case study of xiaohongshu (red). *Sustainability*, 16(15).

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1391–1399, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Lijuan Xu and Li Li. 2024. Upward social comparison and social anxiety among chinese college students: a chain-mediation model of relative deprivation and rumination. *Frontiers in Psychology*, Volume 15 - 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Fei Zhao, Chonggang Lu, Wangyue, Zheyong Xie, Ziyan Liu, Haofu Qian, Jianzhao Huang, Fangcheng Shi, Zijie Meng, Hongcheng Guo, Mingqian He, Xinze Lyu, Zheyu Ye, Weiting Liu, Boyang Wang, and Shaosheng Cao. 2025. RedOne: Revealing domain-specific LLM post-training in social networking services. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 2648–2674, Suzhou (China). Association for Computational Linguistics.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

A Release Artifacts

Inventory of policy-compliant artifacts to be released in the repository: label schema, prompts, scripts, predictions, aggregated results, paraphrased examples, and repository link for XHS-SCoRE: https://anonymous.4open.science/r/XHS-SCoRE-B0B7/ Core files (under CC BY-NC 4.0):

- **Data**: `data/AIGC_posts.csv` (policy-compliant synthetic posts).
- **Scripts/configs**: `scripts/` (e.g., BERT configs and runners).
- **Results**: `results/` (e.g., `test_split_bertclassifier.csv`, `test_split_llmclassifier.csv`).
- **README**: overview of structure and usage instructions in the repository root.
- **Transcribed video instruction**: `videoinstruction.md`

These artifacts are intended to support procedural reproducibility for constructing comparable reader-grounded datasets, rather than to reproduce the unreleased raw Xiaohongshu corpus. The released label schema, collection instructions, prompt templates, scripts/configs, prediction files, aggregated results, paraphrased examples, synthetic posts, and README specify the operational target, reader perspective, text-only constraint, modeling setup, and evaluation outputs used in XHS-SCoRE. Together, they allow researchers to instantiate the same workflow with a new policy-compliant sample, a different reader population, or another platform domain, while preserving the distinction between original reader responses, calibration checks, model predictions, and aggregate error analyses. Raw human-collected platform posts are not released; reproducibility is therefore supported at the level of protocol, schema, code, prompts, synthetic/paraphrased examples, and aggregate diagnostics.

ChatGPT-5.5, ChatGPT-5.3-Codex, ChatGPT-5.1-Codex-Max, and ChatGPT-4o were used to assist coding and formatting; all code was reviewed and approved by the authors.

A.1 Rerun stability of zero-shot prompted classification

This subsection reports rerun agreement and confusion-drift statistics for the original zero-shot prompting setup, complementing the summary reported in Section 6.4.

B Prompts and Constraints

B.1 Prompt variants used in robustness evaluation

Appendix B documents the prompting regimes used in the main and robustness evaluations, first through a compact summary of prompt families and then through the full verbatim prompt templates. English translations are provided below the Chinese prompt text for readability; label names are kept in their original uppercase form.

B.2 Zero-shot prompt

System prompt

作为一名 18-24 岁的典型活跃社交媒体用户的视角，仅根据提供的帖子文本将其分类为且仅为一个标签：
- UPWARD：帖主比我更好
- DOWNWARD：帖主比我更糟
- NEUTRAL：与我差不多，或没有/不清晰的比较

User prompt

帖子：
{post_text}

Output constraint

仅输出 JSON：
{"label":"UPWARD|DOWNWARD|NEUTRAL"}

English translation

> From the perspective of a typical active social-media user aged 18–24, classify the provided post text into one and only one label: UPWARD, where the poster seems better off than me; DOWNWARD, where the poster seems worse off than me; or NEUTRAL, where the poster seems similar to me, or where no comparison is present or clear. The user message supplies the post as {post_text}, and the model must output JSON only.

B.3 Persona-primed prompt

System prompt

你将以“在群体画像内的读者”的视角来判断帖子是否会引发社交比较及其方向。你的读者画像如下：
- 年龄：18-24岁
- 身份：香港在读大学生
- 背景：中国大陆成长背景
- 性别：女性
- 使用习惯：小红书/RedNote 典型活跃用户（高频浏览生活方式内容）
- 社会经济地位：中上（高于中位数）

任务：仅根据提供的【帖子文本】（不考虑图片/视频/评论/作者主页等任何外部信息），将其分类为且仅为一个标签：
- UPWARD：这条帖子会让我感觉“帖主比我更好/更优越/更成功/更幸福/更有资源”，从而引发向上比较。
- DOWNWARD：这条帖子会让我感觉“帖主比我更糟/更弱势/更失败/更不幸/更缺资源”，从而引发向下比较。
- NEUTRAL：这条帖子与我差不多，或不清晰/不构成比较邀请（例如纯信息、说明、广告，无明显自我-他人定位）。

User prompt

帖子：
{post_text}

Output constraint

仅输出 JSON：
{{“label”:“UPWARD|DOWNWARD|NEUTRAL”}}

English translation

> Judge whether the post elicits social comparison, and in which direction, from the standpoint of a reader within the target group. The reader profile is: aged 18–24; a university student in Hong Kong; raised in mainland China; female; a typical active Xiaohongshu/RedNote user who frequently browses lifestyle content; and upper-middle socioeconomic status.
>
> The task is to classify the post, using only its text and ignoring images, video, comments, author profile pages, or any other external information, into one and only one label. UPWARD means the post would make me feel that the poster is better off, more advantaged, more successful, happier, or better resourced than I am, thereby eliciting upward comparison. DOWNWARD means the post would make me feel that the poster is worse off, more vulnerable, less successful, less fortunate, or less resourced than I am, thereby eliciting downward comparison. NEUTRAL means the post seems similar to my own situation, or does not clearly

Model	ReRun 1 agreement	ReRun 2 agreement	Mean agreement	Mean abs. confusion drift	SD confusion drift
qwen235b	97.33%	97.33%	97.33%	0.67 pp	0.94 pp
qwen30b	91.00%	91.00%	91.00%	2.33 pp	3.09 pp
gpt5	90.67%	92.33%	91.50%	1.67 pp	2.00 pp
gpt41nano	93.33%	94.33%	93.83%	1.89 pp	2.38 pp

Table A1: Stability of the original zero-shot prompting setup across two additional full reruns per model. Agreement is computed relative to the original run. "Mean abs. confusion drift" is the mean absolute change in row-normalized confusion-matrix cells across reruns, summarized in percentage points (pp). The low drift values indicate that the qualitative error profiles discussed in Section 6.4 are stable properties of the tested zero-shot configuration rather than one-off sampling noise.

Prompt family	Main modification relative to zero-shot	Purpose in this paper	Main observed pattern
Zero-shot	Base first-person reader prompt with label definition and JSON output constraint	Primary benchmark condition	Strong neutralization for several models; main comparison setting
Persona-primed	Makes the reader profile explicit in the system prompt	Tests whether stronger reader-profile anchoring improves directional recovery	Often leaves neutralization intact; sometimes worsens it
Few-shot	Adds labeled exemplars before inference	Tests whether demonstration-based prompting stabilizes classification	Mixed gains; does not reliably remove neutralization
Cue-explicit	Adds corpus-informed heuristic cue inventory for comparison direction	Tests whether substantial construct-specific scaffolding derived from prior corpus analysis can reduce neutralization	Strongest overall prompt variant; reduces neutralization substantially for several models, but functions as a diagnostic rather than routine classifier prompt

Table B1: Prompt families evaluated in addition to the primary zero-shot setting. The robustness variants preserve the same text-only task definition and output label inventory, but differ in how much reader-profile, exemplar, or cue-level guidance is provided to the model.

invite comparison, as in purely informational, explanatory, or advertising content with no clear self–other positioning. The user message supplies the post as `{post_text}`, and the model must output JSON only.

B.4 Few-shot prompt

System prompt

任务：仅根据【帖子文本】输出且仅输出一个标签：

- UPWARD：帖主比我更好（引发向上比较）
- DOWNWARD：帖主比我更糟（引发向下比较）
- NEUTRAL：与我差不多，或没有/不清晰的比较邀请

User prompt preamble

下面是带标签的示例：

Labeled examples 1–6, inference target, and output constraint

【示例 1】
帖子：
这次去海岛旅游真的是我最近最幸福的回忆！海水巨蓝，沙滩上的沙子超级细腻踩上去，而且几天都是蓝天白云的好天气，像在童话世界一样。在这里我心情超级放松，拍了很多漂亮的人生照片。每次想到这次的旅行经历，心里很满足很快乐，觉得自己是世界上最最幸福的旅行家！尤其是感受到海风的轻拂，温暖的阳光，心情真的特别舒畅。这次的经历让我深深体会到大自然的美丽景观才是最打动人心的，真心希望下一次还能遇到这么完美的风景和天气，我要把这些瞬间定格在心里！！
输出：
`{"label":"UPWARD"}`

【示例 2】
帖子：
今天去了家小红薯上超多人打卡的网红甜品店，点了店里最受欢迎的芒果慕斯。慕斯味道超级浓郁，每一口都像是在吃软绵绵的芒果味的云朵！奶油细腻滑顺，芒果香味浓郁，让我觉得每口吃下去都很丝滑，口感很美妙，感觉自己拥有了超棒的一次甜品体验，完全没有踩雷，所以吃的时候心情也特别开心。这个甜点的颜值太高了，出了很多片，味道也是堪称完美哈哈哈～以后我也会常常来这里打卡更多别的甜品，享受让人治愈的幸福时光，也希望每个喜欢甜食的朋友都能尝到这份幸福，感受到甜蜜带来的快乐～

输出：
{“label”:“UPWARD”}

【示例 3】
帖子：
今天的天气非常适合外出，天空晴朗，阳光充足，没有云层遮挡。空气清新，没有明显的雾霾或污染，温度适中，大约在20℃左右，既不冷也不热。这样的天气非常适合散步、骑行或者进行一些户外运动。建议出门时携带防晒霜和太阳帽，保护皮肤免受紫外线的伤害。同时，注意补充水分，避免长时间暴露在阳光下。未来几天的天气基本保持稳定，适合安排一些户外活动，享受好天气带来的愉悦心情。
输出：
{“label”:“NEUTRAL”}

【示例 4】
帖子：
最近几天天气一直处于阴天状态，天空布满厚厚的云层，没有阳光照射，大约在15℃左右。空气湿度较高，感觉有些潮湿，偶尔还会有细雨或毛毛雨飘落。这样的天气比较适合待在室内工作或休息，不太适合进行剧烈运动。外出时建议携带雨具，穿着防水外套，以免淋湿。阴天的天气虽然没有阳光明媚，但也带来一种宁静的感觉。未来几天可能还会持续多云或阴天，请根据天气情况合理安排出行计划。
输出：
{“label”:“NEUTRAL”}

【示例 5】
帖子：
今天又和妈妈吵起来 她说我总是不懂事 我说我已经很努力了但是没有被看见 自己的安排总是被她一句话推翻 还被责怪得特别难受 总说别家的孩子怎么怎么的更有出息 而我是总被拿来做比较的那个 时间长了我也不想顶嘴 本来以为也习惯了 但她时不时的冷嘲热讽总让我感觉到窒息 今天的晚饭没做好 她又说你怎么这么笨 我没说话一个人在收拾 感觉自己像个被动挨打的失败者 家里房门也不让我锁 我连喘息的空间都没有 别人都说父母是港湾 我为什么总觉得我是在被风浪反复拍打呢?
输出：
{“label”:“DOWNWARD”}

【示例 6】
帖子：
爸爸说我回家太晚，我说只是加班，他却说我不孝顺，老是出去鬼混，我被他骂得特别委屈，回房间想锁房门，却发现钥匙被他收走了，行李也被他翻得乱七八糟。别人家的父母会先问“你累不累”，我家只有“你怎么这么没出息，怪不得不找对象，谁看得上你”。我也想好好沟通，可他一句“别找借口”把我的话堵死了。我的工资卡也被要求上交，我没有反抗的勇气。明明我有把自己的行程提前报备，也从来不去酒吧这些，但他还是不听。我真的太难过了，又不知道该怎么办。
输出：
{“label”:“DOWNWARD”}

现在请分类下面这条帖子（只输出JSON）：
帖子：
{post_text}

仅输出JSON：
{“label”:“UPWARD|DOWNWARD|NEUTRAL”}

English translation

The system instruction asks the model to output exactly one label based only on the post text: UPWARD, where the poster seems better off than me and elicits upward comparison; DOWNWARD, where the poster seems worse off than me and elicits downward comparison; or NEUTRAL, where the poster seems similar to me, or no comparison invitation is present or clear. The prompt then introduces labeled examples.

Example 1. This island trip has truly become one of my happiest recent memories. The sea was intensely blue, the sand was fine and soft underfoot, and for several days the sky was blue with white clouds, like a fairytale world. I felt completely relaxed there and took many beautiful “photos of a lifetime.” Whenever I think back on the trip, I feel satisfied and happy, as if I were the happiest traveler in the world. The sea breeze and warm sunlight were especially soothing. The experience made me feel how moving nature can be, and I truly hope I can encounter such perfect scenery and weather again and keep these moments in my heart. **Output:** UPWARD.

Example 2. Today I went to a popular Xiaohongshu dessert shop and ordered its signature mango mousse. The mango flavor was rich; each bite felt like eating a soft mango cloud. The cream was smooth, the fragrance was full, and the texture was silky and lovely. I felt that I had enjoyed a wonderful dessert experience without disappointment, and I was especially happy while eating it. The dessert looked beautiful in photos and tasted almost perfect. I will keep coming back to try more desserts and enjoy these small healing moments, and I hope everyone who likes sweets can taste this happiness too. **Output:** UPWARD.

Example 3. The weather today is very suitable for going out: the sky is clear, the sunlight is bright, and no clouds are blocking it. The air is fresh, with no obvious haze or pollution, and the temperature is mild, around 20 degrees Celsius. This weather is well suited for walking, cycling,

or other outdoor activities. It is advisable to bring sunscreen and a sun hat, protect the skin from ultraviolet rays, drink enough water, and avoid prolonged sun exposure. The weather should remain stable over the next few days, making it a good time to plan outdoor activities and enjoy the pleasant conditions. **Output:** NEUTRAL.

Example 4. The weather has been cloudy for the past few days, with the sky covered by thick clouds and no direct sunlight. The temperature is around 15 degrees Celsius, and the high humidity makes it feel somewhat damp. There may occasionally be drizzle or light rain. This kind of weather is better suited to staying indoors to work or rest, and less suitable for strenuous exercise. When going out, it is best to carry rain gear and wear a waterproof jacket. Although cloudy weather lacks bright sunshine, it also brings a quiet feeling. The next few days may remain cloudy, so travel plans should be arranged according to the weather. **Output:** NEUTRAL.

Example 5. I argued with my mother again today. She said I am always immature; I said I have already been trying hard, but no one sees it. My own plans are always overturned by a single sentence from her, and I am blamed in ways that hurt. She keeps saying other people's children are more promising, while I am always the one being compared. Over time I stopped wanting to talk back. I thought I had become used to it, but her sarcasm still makes me feel suffocated. Dinner was not done well today, and she said again that I was stupid. I said nothing and cleaned up alone, feeling like a passive failure being beaten down. I am not even allowed to lock my bedroom door, and I have no space to breathe. People say parents are a harbor; why do I feel as if I am being battered by waves again and again? **Output:** DOWNWARD.

Example 6. My father said I came home too late. I said I was only working overtime, but he accused me of being unfilial and always going out to fool around. I felt deeply wronged. I went back to my room and wanted to lock the door, only to find that he had taken the key; my luggage had also been searched and left in a mess. Other people's parents ask first, "Are you tired?" Mine only say, "Why are you so useless? No wonder you do not have a partner. Who would want you?" I want to communicate properly, but one sentence – "stop making excuses" – blocks everything I try to say. I am also required to hand over my salary card, and I do not have the courage to resist. I had clearly reported my schedule in advance and never go to bars, but he still refuses to listen. I am so sad and do not know what to do. **Output:** DOWNWARD.

After the examples, the prompt asks the model to classify a new post, supplied as `{post_text}`, and to output JSON only.

B.5 Cue-explicit prompt

System prompt

任务：仅根据【帖子文本】判断其最可能引发的比较方向，输出且仅输出一个标签（UPWARD/DOWNWARD/NEUTRAL）。

分类线索（启发式，不是硬规则；以"是否邀请比较"+"方向"综合判断）：

A）更可能为 DOWNWARD（向下比较：读者感觉自己相对更好）
· 叙事常呈现冲突/受挫/被指责/被控制：与父母/伴侣/他人争吵、被骂、被否定、被干涉。
· 低能动性/受害者叙事：大量被动句或"被...”结构、无力反抗、被夺走选择权。
· 强负面情绪与强化：难受/委屈/崩溃/窒息 等 + 很/特别/超级 等强化。
· 否定词密集：不、没、没有、别、从来不 等。
· 报告动词/转述冲突：说、骂、问、逼 等，突出指责对话。
· 明示"别人更好而我更差"：如"别人家的...都...我却...”（注意：这里"别人更好"是用来凸显帖主更差，从而让读者产生向下比较）。

B）更可能为 UPWARD（向上比较：读者感觉对方相对更好）
· 生活方式的"丰裕/理想/高光"框架：旅行、美食打卡、外貌管理、消费/购买体验、精致生活。
· 积极评价词与最高级/极值表达：好看、可爱、幸福、完美、最...、超级...、堪称... 等。
· 适度感叹号与列举：用"！"增强兴奋感；用顿号/冒号/清单罗列美好事物或成就。
· 语气展示满足与拥有感：强调"我拥有/我体验到/我达成"的高位体验。

C）更可能为 NEUTRAL（中性/无比较邀请）
· 信息型、说明型、教程型、天气/菜谱/客观建议/广告介绍为主。
· 低情绪或无个人立场；缺乏自我-他人定位；不暗示谁更好/更差。
· 即使出现积极/消极词，但不指向"我与他人/我与帖主"的相对地位。

User prompt

帖子：
`{post_text}`

Output constraint

仅输出 JSON：
`{{"label":"UPWARD|DOWNWARD|NEUTRAL"}}`

English translation

The task is to judge, using only the post text, which comparison direction the post is most likely to elicit, and to output exactly one label: `UPWARD`, `DOWNWARD`, or `NEUTRAL`. The following cues are heuristic rather than hard rules; classification should combine whether the post invites comparison with the likely direction of that comparison.

More likely `DOWNWARD`. The reader feels relatively better off. Narratives often involve conflict, frustration, blame, or control, such as quarrels with parents, partners, or others, being scolded, negated, or interfered with. Low-agency or victim-positioned narration may appear through passive constructions, inability to resist, or having one's choices taken away. Strong negative emotions may be intensified by words such as very, especially, or super. Negation may be dense, with forms such as not, no, never, or do not. Reporting verbs such as say, scold, ask, or force may foreground accusatory dialogue. The post may also explicitly contrast "others are better while I am worse," for example "other people's families all..., but I..."; in such cases, others being better highlights the poster's worse position and can elicit downward comparison.

More likely `UPWARD`. The reader feels the other person is relatively better off. Posts may frame lifestyle abundance, ideal moments, or peak experiences, including travel, food check-ins, appearance management, consumption, purchases, or refined living. They often use positive adjectives and superlative or extreme expressions, such as beautiful, cute, happy, perfect, the most, super, or almost flawless. Exclamation marks may heighten excitement, while enumeration with pauses, colons, or lists can accumulate desirable experiences or achievements. The tone often conveys satisfaction and possession, emphasizing what "I have," "I experienced," or "I achieved."

More likely `NEUTRAL`. The post is mainly informational, explanatory, tutorial-like, weather-related, recipe-based, advisory, or promotional. It has low emotion or little personal stance, lacks clear self–other positioning, and does not imply who is better or worse. Even if positive or negative words appear, they do not point to the reader's relative standing with the poster or with others. The user message supplies the post as `{post_text}`, and the model must output JSON only.

B.6 Generation constraints and prompts

To align generation with corpus-derived cues, we constrain prompts by label.

DOWNWARD (向下比较) replicate conflict/low-agency narrative:

- 7 posts; text only; length 190–200 chars.
- Topics: 4 interpersonal (parents), 2 shopping/purchase, 1 education/academics.
- Heavy personal pronouns and reporting verbs (e.g., 说) to depict quarrels.
- Include negation (不，没) and negative emotional adjectives with intensifiers.
- Use 被and passive patterns to signal low agency/victimhood.
- Include comparisons where "others" are better off than the poster.
- Add 1 extra post with no comparison: text-only, 170–180 chars, product ad, no personal affect.
- Intentionally add minor language/punctuation errors for authenticity.

Full downward prompt (Chinese):

按照以下要求，生成7条会令读者产生向下比较的小红书帖子：
1. 只要求文字；
2. 帖子长度在190-200字左右；
3. 4条帖子的主题为人际关系，话题主要围绕父母；2条帖子的主题为购物或购买；1条帖子的主题为教育和学业；
4. 帖子中多使用人称代词，如"我"，"她"，"他"，和报告动词，如"说"，来描绘争吵；
5. 融合一些否定词，如"不"，"没有"，"有"；
6. 使用一些情绪形容词来表达负面情绪，并加入一些增强词来增强负面情绪的表达；
7. 句子中可以包含与"别人"的比较，并且"别人"的情况要比发帖人好；
8. 使用一些被动句来构建一种低能动性和受害者的叙事；
9. 生成1条不会令读者产生任何比较的小红书帖子：只要求文字；帖子长度在170-180字左右；话题围绕在产品介绍（广告）；帖子不涉及个人情感色彩
10. 故意手动加了语言使用错误，看上去更加真实

English translation

Generate seven Xiaohongshu posts that would lead readers to make downward comparisons. Use text only. Each post should be around 190–200 Chinese characters. Four posts should focus on interpersonal relationships, mainly involving parents; two should focus on shopping or purchases; and one should focus on education or academics. Use many personal pronouns, such as "I," "she," and "he," and reporting verbs such as "said" to portray arguments. Incorporate words of negation or absence, and use emotional adjectives with intensifiers to express stronger negative feeling. Sentences may include comparisons with "others," where others are better off than the poster.

Use passive constructions to build a low-agency, victim-positioned narrative. In addition, generate one post that does not elicit any comparison: text only, around 170–180 characters, centered on product introduction or advertising, without personal emotion. Add minor language or punctuation errors deliberately so that the posts look more authentic.

UPWARD (向上比较) mimic aspirational/abundance framing:

- 7 posts; text only; length 170–180 chars.
- Topics: 4 travel/food, 2 appearance, 1 shopping/purchase.
- Use positive adjectives (e.g., 可爱，好看) and superlatives.
- Use exclamation marks sparingly; use enumeration commas/colons for lists.
- Include minor language/punctuation errors for authenticity.

Full upward prompt (Chinese):

按照以下要求，生成7条会令读者产生向上比较的小红书帖子：
1．只要求文字；
2．帖子长度在170-180字左右；
3． 4条帖子话题主要围绕旅行和美食打卡；2条帖子的主题为外貌；1条帖子的主题为购物或购买；
4．用积极的形容词来形容经历或拥有的事物，如“可爱”，“好看”；
5．融入一些最高级表达来形容经历或拥有的事物；
6．用感叹号来加强情感表达，但不要过度使用；
7．用顿号和冒号来罗列积极的事物，但不要过度使用；
8．加入一些语言使用和标点错误来模仿真实小红书帖子。

English translation

Generate seven Xiaohongshu posts that would lead readers to make upward comparisons. Use text only. Each post should be around 170–180 Chinese characters. Four posts should focus on travel and food check-ins; two should focus on appearance; and one should focus on shopping or purchases. Use positive adjectives to describe experiences or possessions, such as “cute” and “beautiful.” Include superlative expressions to describe what the poster has experienced or owns. Use exclamation marks to strengthen emotional expression, but not excessively. Use enumeration marks and colons to list positive things, again without overusing them. Add minor language and punctuation errors to imitate authentic Xiaohongshu posts.

NEUTRAL (中性) informational, low-affect control:

- 7 posts; text only; length 170–180 chars.
- Topics: weather, recipes, ads, etc.
- No personal affect or self/other positioning.

Full neutral prompt (Chinese):

生成7条不会令读者产生任何比较的小红书帖子：
1．只要求文字；
2．帖子长度在170-180字左右；
3．话题围绕在天气，菜谱，广告等；
4．帖子不涉及个人情感色彩

English translation

Generate seven Xiaohongshu posts that would not lead readers to make any comparison. Use text only. Each post should be around 170–180 Chinese characters. Topics should revolve around weather, recipes, advertisements, and similar informational content. The posts should not involve personal emotional coloring.

C Generated Examples

Illustrative Xiaohongshu-style outputs from the constrained generation recipes (full set in `data/AIGC_posts.csv`):

- **UPWARD (class 0)**: “这次去海岛旅游真的是我最近最幸福的回忆！海水巨蓝，沙滩上的沙子超级细腻踩上去，而且几天都是蓝天白云的好天气，像在童话世界一样。” *Translation:* “This island trip has truly become one of my happiest recent memories. The sea was vividly blue, the sand was wonderfully fine and soft underfoot, and for several days the sky stayed bright blue with white clouds, as if I were in a fairytale.” (aspirational travel, peak-experience framing).
- **NEUTRAL (class 1)**: “今天的天气非常适合外出，天空晴朗，阳光充足.....建议出门时携带防晒霜和太阳帽，保护皮肤免受紫外线的伤害。” *Translation:* “The weather today is ideal for going out: the sky is clear and the sunlight is abundant. It is recommended to bring sunscreen and a sun hat to protect the skin from ultraviolet rays.” (informational weather, no self–other positioning).
- **DOWNWARD (class 2)**: “今天又和妈妈吵起来她说我总是不懂事我说我已经很努力了

但是没有被看见……总说别家的孩子怎么怎么的更有出息，而我是总被拿来做比较的那个。” *Translation:* “I argued with my mother again today. She said I am always immature; I said I have already been trying hard, but no one sees it. She keeps saying that other people’s children are so much more promising, while I am always the one being compared.” (conflict, low-agency family narrative).

D Full Illustrative Examples

D.1 UPWARD example

ID: uc_01095
Split: train
Full text:

亚庇|拍到了人生照片
日落、海浪、沙滩、夕阳、烟花，
一切美好的元素都在这一刻集齐了。
太浪漫了～～#那是我们一直想去的海边#人生再也拍不出来的照片#海边日落#亚庇#亚庇旅游#亚庇日落#人生照片#马来西亚#落日

Translation:

Kota Kinabalu | Captured the photo of a lifetime
Sunset, waves, beach, twilight, fireworks—
all the beautiful elements came together in this single moment.
So romantic~~ #the-seaside-we-always-wanted-to-visit #a-photo-of-a-lifetime-that-can-never-be-recreated #seaside-sunset #KotaKinabalu #KotaKinabaluTravel #KotaKinabaluSunset #photo-of-a-lifetime #Malaysia #sunset

Interpretation: This elicits upward comparison through travel-related semantic content, asyndetic listing of high-aesthetic experiential elements, and the superlative framing of “人生照片” (“the photo of a lifetime”). Together, these features construct an idealized lifestyle moment and a peak emotional experience, inviting readers to compare their own more ordinary reality against the poster’s “perfect moment.”

D.2 DOWNWARD example

ID: dc_01364
Split: val
Full text:

拉黑我妈后，我的人生才刚刚开始
那是平淡无奇的一天
按部通勤的地铁上，高楼一栋栋从窗户里掠过，白云低得仿佛站树梢上抬起手就能触碰
深圳的天气还是一样的好

没有争吵，甚至没有发生任何事
我拿出手机，在那个原本刺眼的头像上，按了拉黑

那一刻终于明白，我的前半生一直都在被人操纵
大到人生选择，生活方式
小到今天穿什么衣服，做什么事
在我父母那里得到的永远是反对，贬低与道德绑架

而如今
走过27年光景，当一切切不堪往事再重提之时
心态有了改变
也终于意识到他们对我实施操纵的手段

《操纵心理学》里描述得很对
他们利用情感纽带裹挟亲人，有意无意地影响他人思想、情感或行为
很多阅历尚浅的人，根本无法免于操纵
操纵者往往会把操纵行为掩饰在一层层谎言之下，让人无法察觉
而由于他们是亲人的关系，我们倾注了很多感情，难以割舍，而操纵行为就有了温床
特别建议和曾经的我一样，陷入原生家庭困扰情感操纵的人去阅读《操纵心理学》
它是实用性强的工具书
教你认识容易被人得寸进尺的7种人格特征，防范9种操纵类型，学习摆脱操纵的7个诀窍

如果你不想成为操纵者的目标，就要成为一个自我导向的人，愿你活出自我，自己把握人生
#家庭关系#父母沟通#操纵心理学#INFP精神世界#个人成长#原生家庭#情感树洞

Translation:

Only after blocking my mother did my life truly begin.
It was an utterly ordinary day.
On the subway during my usual commute, high-rises swept past the window one after another, and the clouds hung so low it felt as if I could touch them just by standing on tiptoe and reaching up.
The weather in Shenzhen was as good as ever.

There was no argument; in fact, nothing had happened at all.
I took out my phone and tapped “block” on that profile picture that had always felt so piercing.

At that moment, I finally understood that the first half of my life had been controlled by others.
From major life choices and ways of living
to something as small as what to wear today and what to do,
all I ever received from my parents was opposition, belittlement, and emotional coercion.

And now,
after 27 years, when all those unbearable past events are brought up again,
my mindset has changed,
and I have finally recognized the ways they used to control me.

The Psychology of Manipulation describes it exactly right.
They use emotional bonds to coerce those close to them, intentionally or unintentionally influencing other people’s thoughts, feelings, and behavior.
Many people with little life experience simply cannot avoid being manipulated.
Manipulators often conceal their behavior beneath layers of lies, making it hard for others to notice.

> And because they are family, we invest so much emotion in them and find it hard to cut ties, which gives manipulation fertile ground.
> I especially recommend *The Psychology of Manipulation* to people who, like my former self, are trapped in emotional manipulation and distress rooted in their family of origin.
> It is a highly practical guide.
> It teaches you to recognize seven personality traits that make people vulnerable to being taken advantage of, guard against nine types of manipulation, and learn seven strategies for breaking free.
>
> If you do not want to become a manipulator's target, you must become a self-directed person. May you live as your true self and take charge of your own life.
> #family-relationships #parent-child-communication #the-psychology-of-manipulation #INFP-inner-world #personal-growth #family-of-origin #emotional-confessions

Interpretation: This elicits downward comparison through parallelism, high-intensity negative affect, and a cluster of low-agency terms tied to the "family of origin" theme. These features together construct a narrative of restricted agency and emotional distress, inviting readers to perceive their own domestic autonomy as relatively superior.

D.3 NEUTRAL example

ID: nc_02285
Split: val
Full text:

> AI最新排行榜｜夸克增速狂飙
> 夸克这波出息了，appstore排名超越元宝，看来新上线的深度思考功能很好用啊！大家有用过的嘛评论区说说看！#国产ai #AI榜单#夸克#deepseek #人工智能#智能科技#ai助手#夸克AI深度思考

Translation:

> Latest AI rankings | Quark's growth is surging
> Quark has really made a name for itself this time—its App Store ranking has surpassed Yuanbao. It seems the newly launched Deep Thinking feature is really useful! Has anyone tried it? Tell me in the comments!
> #domestic-AI #AI-rankings #Quark #deepseek #artificial-intelligence #smart-technology #AI-assistant #Quark-AI-Deep-Thinking

Interpretation: This is classified as neutral because the semantic focus is shifted toward third-party technological entities and rankings. The post uses impersonal evaluative language and emphasizes functional utility rather than the poster's own lifestyle, agency, or status relative to the reader, so it does not clearly invite reader–poster comparison.

E BERT Training Details

Additional supervised encoder details for reproducibility (files in `bert_train/` retained locally; key settings summarized here), pretrained weight from Cui et al. (2020):

- **Environment**: Linux (6.8.x), Python 3.13.5, PyTorch 2.8.0+cu128, Transformers 4.55.2, Datasets 4.0.0, GPU: RTX 4090 D (CUDA-capable); HF models: `hfl/chinese-bert-wwm-ext`, `hfl/chinese-roberta-wwm-ext`, `hfl/chinese-macbert-base`.
- **Final configs** (also mirrored under `scripts/bert training config/`): BERT lr=2×10^{-5}, RoBERTa lr=2.5×10^{-5}, MacBERT lr=3×10^{-5}; max_length=512; batch=16; grad_accum=2; epochs=15; weight_decay=0.01; warmup ratio 0.2/0.2/0.15; label smoothing 0.15/0.15/0.1; start/end lr ratios and logit_scale per model.
- **Grid search plan**: per model, four runs varying learning rate (2e-5, 2.5e-5, 3e-5), warmup ratio (0.15/0.2), label smoothing (0.0–0.15), start/end lr ratios (e.g., 0.05–0.3), and logit_scale (0.85–0.9) over 5 epochs; best checkpoint selected by validation Macro-F1, then a full 15-epoch train with the chosen hyperparameters.