Psychologically Potent, Computationally Invisible: LLMs Generate Social-Comparison Triggers 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 if a text-only Xiaohongshu (RedNote) post elicits UPWARD, DOWNWARD, or NEUTRAL/no clear social comparison from a first-person reader perspective. The task targets a socially meaningful relational signal that is behaviorally real yet not reducible to sentiment. Across prompted LLM classifiers and supervised Chinese encoder baselines, we find a consistent mismatch between generation fluency and reliable detection ability: the signal is textually learnable in-domain, but not robustly accessible to prompt-based classification. Prompted LLM classifiers exhibit stable, interpretable failure modes, especially neutralization of comparison-triggering posts and model-specific directional skew. A controlled pilot further 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 both 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 environment for social comparison: users infer standing by comparing themselves to others (Festinger, 1954). Attention-optimizing feeds amplify interpersonal cues, so ordinary posts can become implicit benchmarks (Wu et al., 2024). On highly visual, lifestyle-oriented platforms, achievements, bodies, consumption, and family narratives bacome "rankable" (Jabłońska and Zajdel, 2020; Valkenburg and Peter, 2011), including Xiaohongshu (Gu et al., 2026). These cues can quickly 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, these dynamics intertwine with "involution" and "lying flat," reframing comparison as participation in performance norms (Wang et al., 2024; Deng et al., 2025).

This creates a distinct NLP problem: can a model recover the direction of reader-perceived social comparison in naturally occurring posts? On Xiaohongshu, the relevant cues are often relational, implied, and culturally grounded rather than lexically explicit. The task is therefore not simply to detect sentiment or overt evaluation, but to determine whether a post invites comparison from a first-person reader perspective. The question is newly urgent because generative AI can now produce humanlike social-media text at scale, with persuasive, affect-calibrated style (Salvi et al., 2025). Recent NLP work also points to a tighter coupling between platform discourse and model development, including domain-specific post-training and the use of LLMs as instruments in computational social science (Zhao et al., 2025; Ziems et al., 2024). Together, these trends create a specific risk: models may generate comparison triggers that shape reader affect, yet fail to recover the same relational meaning in everyday posts. For auditing, moderation, and measurement, such failures matter because they can erase or distort socially grounded meaning (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 detecting comparison direction in text-only Xiaohongshu posts, targeting a psychologically consequential meaning dimension that is not reducible to sentiment; 2) show that this signal is textually learnable for strong in-domain encoders but remains unreliably accessible to prompted LLM classification, revealing a gap between socially resonant generation and reliable detection;

3) demonstrate that the resulting failures are not random but structured, with neutralization and directional skew emerging as repeatable error profiles across prompting regimes; 4) by combining corpus-constrained generation with a controlled human pilot, show that LLM-generated texts can successfully instantiate the same comparison-relevant signal in readers even when prompt-based detection remains fragile. Overall, the paper contributes both a benchmark and a reusable error-analysis template for studying cases in which generation fluency does not imply reliable detection ability for reader-grounded relational meaning.

2 Related Work

2.1 Social Comparison on Platforms 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 often narrated in socially rankable ways without overt comparison markers (Wang et al., 2024; Gu et al., 2026). For NLP, the key gap lies in modeling reader-perceived direction when relational positioning is implied rather than explicit.

2.2 NLP for Social and Psychological Measurement: Limits for Reader-Centered Meaning

NLP has long been used to infer 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). These works establish the feasibility of text-based social measurement, but they also reveal the assumptions: labels are author-centered and cues have relatively stable lexical or stylistic correlates. Social comparison direction departs from both 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 resulting gap is not simply one of domain shift, but of measurement form: psychosocial evidence shows that comparison cues matter, while existing NLP targets provide limited tools for recovering them as reader-grounded direction in the wild.

2.3 LLMs Under Pragmatic and Domain Shift: Detection Reliability Versus Generative Potency

LLMs are increasingly used as drop-in classifiers for socially grounded labels, yet reliability remains uneven when labels depend on pragmatics, implicit norms, or culturally embedded meaning (Ziems et al., 2024; Sravanthi et al., 2024). These difficulties are likely to be amplified for reader-centered constructs such as comparison direction, where labels depend on situated interpretation rather than overt lexical marking. At the same time, LLMs have become effective generators of socially plausible, emotionally calibrated text: generated language can be persuasive (Salvi et al., 2025), difficult for readers to distinguish from platform-native writing (Dugan et al., 2024), and increasingly intertwined with platform discourse through domain-specific post-training and evaluation (Zhao et al., 2025). What remains underexplored is the mismatch between these two capacities. Existing work does not directly test whether models that can successfully realize a socially consequential construct in generation can also recover that same construct reliably under prompted classification. XHS-SCoRE addresses the gap for reader-grounded comparison direction in naturally occurring posts.

3 Task Definition

Reader-perceived social comparison direction in text-only Xiaohongshu posts: given a post text x and a reader profile r, the task is to predict one label $y \in \{\textbf{UPWARD}, \textbf{NEUTRAL}, \textbf{DOWNWARD}\}$, corresponding to the immediate comparison response elicited in first-person reading. In XHS-SCoRE, r is not modeled at the individual level; instead, it is operationalized through a relatively homogeneous reader group and a standardized collection protocol. The resulting benchmark therefore targets a reader-grounded elicitation task rather than author intent or a population-invariant semantic label.

The three labels operationalize comparison elic-

itation rather than sentiment polarity. UP denotes posts that position the poster as better off, more successful, or otherwise advantaged relative to the reader, thereby inviting upward comparison. DOWN denotes posts that position the poster as worse off, more distressed, less resourced, or otherwise disadvantaged, thereby inviting downward comparison. NEU denotes posts that do not clearly invite reader-poster comparison. A post may therefore be affectively positive or negative while still being labeled NEU, and conversely may trigger comparison without containing overt evaluative language. Because XHS-SCoRE is nearly class-balanced, we use Accuracy and Macro-F1 as primary summary metrics, and analyze per-class recall, confusion structure, and predicted-label distributions to diagnose direction-specific failure modes. Section 4.4 provides examples in dataset.

4 Dataset

4.1 Collection protocol and labeling principle

XHS-SCoRE (**X**iao**h**ongshu **S**ocial **Co**mparison **R**eader **E**licitation) consists of text-only Xiaohongshu posts collected by young adult users under a standardized browsing protocol. Three requirements govern inclusion and labeling: 1) all items must come from Xiaohongshu; 2) only posts whose meaning is recoverable from text alone are included (items requiring images or video are excluded); 3) labels reflect the collector's immediate reader reaction during browsing (whether the post elicits comparison and if so, in which direction). The protocol treats elicitation as partly reader-dependent: a homogeneous reader group perceives certain comparisons. XHS-SCoRE therefore targets a psychologically grounded construct–reader-perceived comparison elicitation–rather than author intent.

4.2 Annotation regime and reliability

XHS-SCoRE uses a single-collector immediate-response protocol designed to preserve the first-person, reader-based nature of the construct. Each participant browsed Xiaohongshu on their own device for 7 days and identified 10 posts each that elicited UP, DOWN, and NEU/no clear comparison responses each day, yielding a target of 210 posts. They copied text-only content into structured spreadsheets and submitted entries daily. Participants were students ($N = 67$; $M_{\text{age}} = 21.75$; 85% female; native Chinese) from 8 publicly funded HK universities. We adopt this single-collector design because the task targets reader-perceived comparison elicitation under naturalistic browsing, rather than author intent or a population-invariant label.

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.

To calibrate this reader-grounded operationalization, we report three complementary reliability signals under the same text-only constraint used in modeling (Table 1). On a randomized set of 210 posts, additional in-group raters applying the same immediate-response scheme achieved 63.2-68.4% agreement, providing a conservative estimate of between-reader consistency when multimodal cues are removed. Repeated entries separated by at least 24 hours showed around 82% within-person stability, indicating that labels are reasonably stable for the same reader over time. Further, human raters re-labeled 20.4% of UP and 17.0% of DOWN posts as NEU. These indicate human ambiguity baselines when interpreting model-side NEU collapse.

4.3 Benchmark size and splits

XHS-SCoRE is balanced across the three labels, contains 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 used for all comparisons. The TRAIN split contains 8,350 posts (2,780 UP; 2,779 NEU; 2,791 DOWN; 1,487,712 characters). The VAL split contains 2,783 posts (926 UP; 926 NEU; 931 DOWN; 496,996 characters). The TEST split contains 2,783 posts (926 UP; 926 NEU; 931 DOWN; 467,957 characters).

Class balance reduces the likelihood that strong performance reflects majority-class behavior and makes Macro-F1 interpretable as directional reliability rather than label-frequency exploitation. See Appendix A for all non-raw, policy-compliant artifacts (label schema, prompts, scripts, aggregated results, and AI-generated examples).

4.4 Corpus analysis

To characterize linguistic realizations of direction, we compare distributions using Wmatrix 7 keyness analysis over word, part-of-speech, and semantic tags via log-likelihood statistics, followed by concordance inspection and inductive frame analysis to interpret pragmatic functions (Rayson,

2008). In brief, UP posts more often realize aspirational lifestyles through lexis/discourse associated with consumption, mobility and positive evaluation, whereas DOWN more often realize low-agency, conflict-centered narratives, including heavier use of negation, pronouns, reported speech and passive constructions. These corpus-derived cues later 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 → comparison against the poster's "perfect moment."

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 → relatively better domestic autonomy (readers').

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 instead of poster's own status/agency/experience → no clear comparison.

These examples illustrate the task targets reader-perceived comparison direction rather than sentiment polarity alone: affective language may accompany the labels, but the labeling 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, translation and linguistic interpretations are provided in Appendix D.

5 Models and Experimental Setup

5.1 Prompted LLM classification settings

We evaluate large language models as prompted classifiers for the three-way task. Our primary setting uses a zero-shot, first-person prompt aligned with the benchmark's reader profile and requires a single label based strictly on the post text. Prompts are in Simplified Chinese and constrained to JSON outputs. For all models, we set temperature=.1 to reduce sampling variability while retaining executable API behavior; for gpt-5-2025-08-07: reasoning.effort=minimal is set to limit unnecessary reasoning tokens. The system instruction defines the reader persona and label schema, and the user message inserts the post text and repeats the output constraint. Full prompt templates in Appendix B.

In addition to the primary zero-shot setting, we evaluate a small set of prompting variants as robustness conditions: persona-primed, few-shot, and cue-explicit prompts. These variants preserve the same text-only design and label inventory, and are intended to test if the observed error structure depends narrowly on one prompt formulation. Among these, the cue-explicit should be read as a diagnostic robustness probe rather than a routine classifier prompt, because it incorporates manually abstracted cue structure derived from prior corpus-linguistic analysis. Unless otherwise noted, the main comparison in Sections 6.1–6.3 uses the primary zero-shot setting; Section 6.4 summarizes robustness across different prompting 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 the family used in our stimulus generation pipeline, enabling direct generation–classification comparison (OpenAI, 2025b).

5.2 Supervised encoder baselines

To contrast prompted inference with supervised learning, 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 single best checkpoint by VAL Macro-F1 and report its performance 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 the main empirical contrast: the text-only directional signal is learnable in-domain for supervised encoder baselines, but remains unreliable under prompted LLM inference. Figure 1 shows that the gap is not merely one of aggregate performance but of error structure: several models absorb comparison-triggering posts into NEU, whereas Qwen3-235B exhibits a marked UP-skew profile. The core result is therefore not just weaker

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.

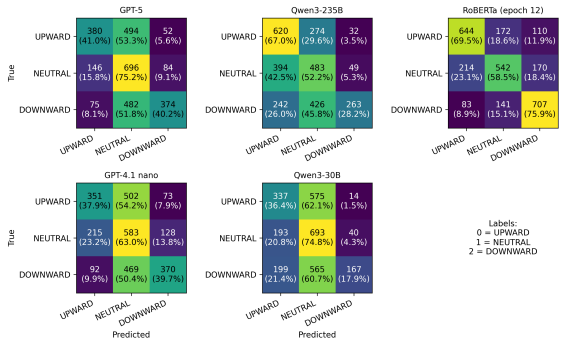

Figure 1: Confusion matrices (test): primary zero-shot LLM vs best encoder baseline. % row-normalized.

prompted classification, but interpretable failure tied to the relational structure of the task. Section 6.4 shows that this qualitative pattern remains broadly stable 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 substantially better. This is not just lower recall on one class: it identifies a systematic weakness in access to one side of the comparison space, and in practical terms would cause monitoring or auditing systems to miss precisely downward-comparison cues that matter most.

6.3 When comparison becomes invisible: error patterns beyond aggregate scores

6.3.1 Neutralization of comparison triggers

The most pervasive failure mode is neutralization: posts that in fact invite comparison are repeatedly reassigned to NEU. This pattern is visible across several prompted LLM settings and exceeds the human ambiguity baseline from Section 4.2. The severity of the error is not just that comparison-triggering language is missed, but that NEU is an explicit non-comparison judgment: it asserts no clear invitation to self–other ranking and thereby renders comparison cues computationally absent.

6.3.2 Directionality asymmetry and DOWN sensitivity

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

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 prompt-based 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 not simply error, but disappearance: comparison-triggering posts are repeatedly translated into non-comparison. Because NEU explicitly asserts the absence of a clear self–other ranking, such errors do more than weaken recovery of the construct; they erase it at the level of model output. The result is a systematic form of computational invisibility for reader-grounded comparison cues under prompted detection.

6.4 Prompt robustness and stability under alternative prompting regimes

Prompted classification can be sensitive to decoding and prompt stochasticity, stability of the primary zero-shot setting is tested with reruns under the same condition. Agreement with the original run is high and confusion-matrix drift is low, indicating that the qualitative error patterns in the zero-shot setting are stable properties of the tested prompting regime rather than one-off sampling noise. Full stats in Table A1 in Appendix A.

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

Table 3 shows a clear but non-uniform robustness pattern. Cue-explicit prompting is the strongest overall prompt condition tested and reduces neutralization sharply for some models. But this setting should not be read as the ordinary or default way of using LLMs as classifiers. 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 therefore do not dissolve the central difficulty; they expose it more clearly. The signal is recoverable, but under common prompting settings prompted LLM classifiers fail to access it robustly and require unusually strong external scaffolding to approach better directional recovery.

The broader conclusion remains unchanged. Some neutralization is expected even under human text-only re-reading, but several zero-shot LLM settings exceed that ambiguity baseline by a wide margin, especially for true DOWN posts. Cue-explicit prompting reduces this effect substantially for GPT-5 and Qwen3-235B/30B, yet recovery still trails the encoder baseline. The problem is not simply lower performance but a recurrent tendency for prompted classification to predict comparison-triggering language as no comparison. Hence, the construct is not absent from the text; it becomes computationally invisible by model 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 explicit constraints derived from corpus-linguistic analysis. The goal is construct-targeted generation rather than generic realism: the UP and DOWN conditions differ in stance, agency framing, evaluative lexis, and abundance-vs-hardship cues while remaining plausible as platform posts, whereas the NEU condition suppresses self-other positioning and affective framing. Full prompts and constraint lists in Appendix B, with stimulus examples in Appendix C.

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.

8 Human Validation: LLM-Generated Posts Elicit Social Comparison

Because social comparison direction is a reader-perceived psychological construct rather than a purely text-internal property, we conduct a lean human study to validate the potency side of the dissociation. The question is whether LLM-generated posts, produced under corpus-derived constraints and minimally edited for naturalness, can elicit (i) the intended comparison direction and (ii) downstream affective responses. The study is intended as construct validation rather than a standalone psychological contribution.

8.1 Participants, design, and procedure

Participants ($N = 29$; similar profile to corpus collectors but no overlap) were recruited via university advertisements and completed the study in a computer lab 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 followed 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. Participants then completed post-exposure measures of comparison-related emotions (Smith,

2000) and general affect (Watson et al., 1988). An instructional attention check was administered at the end, followed by a debrief disclosing the study purpose and the 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 predicting perceived standing from condition shows a strong condition effect (adjusted $R^2 = .570$, $p < .001$), indicating readers infer the intended comparison direction from generated posts.

Comparison-related emotions. DOWN assimilative emoition 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 by condition (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 affective consequences aligned with comparison direction.

8.3 Link to NLP: potency vs detectability

The human results show that our label space is behaviorally grounded: when the generator is instructed to produce UP-, DOWN-, or NEU-comparison posts under corpus-derived constraints, readers infer the intended self–other standing relation and show corresponding shifts in comparison-related emotion. This clarifies what is at stake in the classification failures reported earlier. When a detector labels a comparison-triggering post as NEU, it is not merely making a small semantic error; it is failing to detect content that systematically induces standing judgments and emotion profiles for readers.

This is the sense in which comparison can become computationally invisible: socially consequential relational meaning is present in text as an elicited signal, yet collapses or becomes unreliable under model-based detection, especially when expressed through implicit stance and narrative positioning rather than explicit comparative markers. Together with the controlled generation setup, the pilot supports the paper's central dissociation: LLMs can produce posts that shift human judgments and emotions while still failing to reliably detect the same cues when "reading" them. This asymmetry creates a concrete risk for pipelines that use prompted LLMs to audit, monitor, or moderate comparison-related harms: the system may scale the production of psychologically potent triggers while under-detecting their presence and direction in real platform text.

9 Discussion

9.1 The generation–detection dissociation: what it is and why it matters

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 in human readers, yet prompted LLM classifiers remain unreliable detectors of the same reader-grounded construct when classifying naturally occurring posts. The failure is not undifferentiated: some models repeatedly collapse comparison-triggering posts into NEU, while others skew systematically toward the wrong direction. Because the supervised encoder baselines recover the signal substantially more reliably, the central issue is not the absence of learnable textual cues, but the instability of prompt-based detection for this task under the tested regimes.

The importance of that gap is methodological as much as empirical. The paper does not merely show that one detector family is weaker than another; it identifies a mismatch between behaviorally consequential meaning and robust prompt-based access to that meaning. That mismatch is what makes the benchmark useful: it turns a broad concern about "LLMs being worse" into a more precise account of where reader-grounded inference fails and how that failure is structured.

9.2 Implications for evaluation: generation quality is not evidence of understanding

Generation fluency is not sufficient evidence of reliable detection for reader-grounded relational meaning. In the pilot, corpus-constrained LLM-generated posts shifted perceived standing and comparison-related emotions in human readers, showing that the models can realize the construct effectively in generation. Yet when prompted to classify naturally occurring posts into UP, NEU, or

DOWN, the same family of models remained error-prone in stable and interpretable ways. The relevant contrast is therefore not between generation and "understanding" in the abstract, but between construct realization in text and construct recovery under prompted detection.

That contrast has a direct consequence for evaluation. For socially embedded constructs, aggregate accuracy alone is not enough; the evaluation target includes where meaning disappears and how it is distorted. In XHS-SCoRE, two recurrent error modes dominate: neutralization of comparison-triggering posts and directional skew in the recovered label. For reader-grounded meaning tasks, confusion structure is therefore part of the phenomenon to be explained, not merely an auxiliary diagnostic.

9.3 Implications for governance and platforms: scalable triggers, unreliable monitors

The error profile exposed by XHS-SCoRE is consequential not merely because prompted LLM classifiers underperform, but because their failures are structured in application-relevant ways. Comparison-triggering posts are repeatedly absorbed into NEU or displaced toward the wrong direction, creating systematic blind spots for any downstream use of prompt-based detectors. In settings such as platform monitoring, auditing, or intervention design, the relational cues most worth tracking may therefore be precisely those most likely to be washed out. Comparison-trigger detection cannot be treated as a by-product of fluency or generic classification strength; it is a separate capability that has to be validated directly.

9.4 Implications for 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 be useful for constructing controlled materials that shift perceived standing and comparison-related affect in readers. Yet the same result makes the detection failure more informative, not less: the models can help realize the construct in generation while remaining unreliable instruments for classifying it in naturally occurring posts.

Stimulus usefulness, human grounding and detector reliability are thus not interchangeable signs of model competence. XHS-SCoRE is useful precisely as it forces those dimensions apart and makes their relationship empirically testable. For LLM-assisted social-science workflows, that distinction matters: generated materials may be usable under human validation, but detector outputs cannot be treated as validated measures of reader-grounded meaning without task-specific evaluation.

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. The central result is therefore sharper than the familiar observation that one detector family underperforms another: models can generate texts that successfully trigger social-comparison responses in readers while still failing to reliably detect the same comparison-triggering signal. XHS-SCoRE thus contributes not only a benchmark for the present Xiaohongshu setting, but also a reusable evaluation logic for other reader-conditioned inference tasks in NLP.

10 Conclusion

We introduced Xiaohongshu Social Comparison Reader Elicitation (XHS-SCoRE), a reader-grounded benchmark for detecting whether text-only Xiaohongshu posts elicit UP comparison, DOWN comparison, or no clear comparison from a first-person reader perspective. The benchmark reveals a consistent gap between generation fluency and reliable detection ability: supervised Chinese encoder baselines recover comparison direction substantially more reliably, whereas prompted LLM classifiers exhibit 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 of the same construct remains fragile. Together, these results identify a reader-grounded meaning dimension that is behaviorally consequential and textually recoverable, yet still only partially accessible to prompt-based inference under the tested regimes.

Taken together, the results support a narrow but consequential conclusion: in this text-only, reader-grounded Xiaohongshu setting, LLMs can generate psychologically potent social-comparison triggers that remain only partially visible to prompt-based detection under the tested regimes.

Limitations

This study should be read as a text-first benchmark and diagnostic baseline, not as 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 truths. Third, the behavioral validation is a pilot study (N=29) intended to establish 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 present contribution is therefore best understood as a strong benchmark, a measurement-calibrated comparison between detector families, and an error-analysis template that can be extended to multimodal, cross-platform, and more heterogeneous-reader settings in future work.

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 is 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-trigger 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.

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.

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`

ChatGPT-5.3-Codex, ChatGPT-5.1-Codex-Max and ChatGPT-4o is being used to assist coding, all codes are 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.

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"}

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.

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"}}

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”}

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”}}

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

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。

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

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。” (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。” (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。” (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.