# Psychologically Potent, Computationally Invisible: LLMs Generate Social-Comparison Triggers They Fail to Detect

Anonymous ACL submission

## Abstract

Social-comparison cues on lifestyle platforms are relational and often implied rather than explicitly comparative, yet they can shape readers' perceived standing and affect. We introduce XHS-SCoRE, a reader-grounded benchmark for detecting whether a text-only Xiaohongshu post elicits upward (poster better off), downward (worse off), or no/neutral comparison. Across multiple prompted LLMs as zero-shot classifiers, directional reliability drops sharply relative to in-domain fine-tuned Chinese encoders. The errors are structured: LLMs frequently neutralize comparison-triggering posts and, in some cases, systematically skew toward upward readings, with particularly weak sensitivity to downward cues. To validate that the construct is behaviorally real rather than an annotation artifact, we generate platform-style stimuli under corpus-derived constraints and show in a controlled lab study ($N$=29) that the generated posts reliably shift perceived standing and comparison-related emotion. Together, the results demonstrate a generation–detection dissociation: LLMs can produce psychologically potent comparison cues while computationally unreliable at detecting the same relational meaning, posing a blind spot for measurement, auditing, and platform governance.

## 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 are especially "rankable," intensifying comparison pressures (Jabłońska and Zajdel, 2020; Valkenburg and Peter, 2011). Comparison can quickly shift self-evaluations and emotions, contributing to anxiety, dissatisfaction, envy, and rumination (McComb et al., 2023; Xu and Li, 2024). Both directions can carry costs: upward comparisons heighten negative affect when targets feel advantaged or unattainable (Collins, 1996; McComb et al., 2023), while downward comparisons can elicit undesired affective responses (Buunk et al., 1990). In contemporary China, these dynamics intertwine with projects such as "involution" and "lying flat," reframing comparison as participation in performance norms (Wang et al., 2024; Deng et al., 2025). On Xiaohongshu, cues are often embedded in mundane narration—family talk, consumption choices, schooling, and peer success—rather than explicit comparatives.

These observations raise a concrete NLP question: can language models recover the direction of reader-perceived social comparison in naturally occurring posts? The question is 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 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 risk that surface fluency can obscure: LLMs may write comparison triggers that shape reader affect, yet fail to detect the same relational meaning in everyday, culturally grounded language. This gap matters for auditing, moderation, and measurement, because systematic errors can erase or distort socially grounded meaning (Hovy and Spruit, 2016).

We examine this potent-but-invisible dissociation on Xiaohongshu:

- We define reader-perceived comparison direction with a three-way label space (UPWARD/NEUTRAL/DOWNWARD) and introduce a benchmark dataset reflecting immedi-

ate browsing reactions.

- We compare prompted LLMs as zero-shot classifiers against supervised encoder baselines, showing structured failures including strong collapse of directional cases into NEUTRAL.
- We validate potency with a human pilot: LLM-generated posts constrained by corpus properties shift perceived direction and affect, showing that generation quality does not imply reliable relational understanding.

## 2 Related Work

### 2.1 Social Comparison on Platforms and Reader-Perceived Direction

Social comparison theory treats self-evaluation as relational (Festinger, 1954). Social media intensifies comparison by making targets abundant and salient while compressing interpretive context. Direction is therefore often inferred rather than marked: readers may experience a post as upward or downward without overt comparatives, deriving rank from stance, outcomes, and what looks "normal" in a feed.

Empirical work links comparison to affect. Meta-analytic evidence shows that upward-comparison exposure tends to reduce self-evaluation and increase negative affect (McComb et al., 2023). Platform studies similarly connect comparison to envy, dissatisfaction, and wellbeing outcomes (Appel et al., 2016; Fardouly and Vartanian, 2015). Costs are not confined to upward comparison: downward comparison can be affectively complex and may elicit distress, burden, or moralized negative emotions when others' misfortune becomes salient (Buunk et al., 1990; Verduyn et al., 2020). For NLP, this motivates a construct not reducible to sentiment polarity: direction of reader-perceived relational positioning.

The challenge is especially salient in contemporary Chinese social media, where comparison is often experienced as participation in broader social projects (e.g., Wang et al., 2024; Deng et al., 2025). On Xiaohongshu, cues frequently appear in mundane narration—family, consumption, schooling, peer success—where rankability is implied rather than asserted. Direction is thus carried by pragmatic framing, agency positioning, and evaluative stance rather than explicit comparative morphology.

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

NLP has long inferred psychological and social variables from text, with social media enabling large-scale measurement. Early work used lexicon-driven proxies (Tausczik and Pennebaker, 2010); later work emphasized supervised learning for outcomes such as mental health signals and temporal dynamics (Coppersmith et al., 2015; De Choudhury et al., 2013). Parallel lines operationalize socially grounded labels such as toxicity and personal attacks (Wulczyn et al., 2017).

Social comparison direction stresses two assumptions behind many such targets. First, many labels are author-centered, whereas direction is reader-centered: the same post can be experienced as upward, downward, or neutral depending on how it positions the reader. Second, direction often lacks stable lexical anchors; it can be encoded via narrative roles and agency framing, reported dialogue, evaluative listings, and platform-native registers of aspiration and hardship. These cues are robust to readers but hard to capture with surface heuristics, creating a gap between psychosocial evidence of impact and computational tools for detecting comparison triggers 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 social labels, but reliability is uneven for socially grounded constructs, especially when labels depend on pragmatics, implicit norms, or culturally embedded meaning (Ziems et al., 2024; Sravanthi et al., 2024). This is amplified for reader-centered constructs like comparison direction, where labels depend on situated interpretation.

At the same time, LLMs have advanced as generators of socially plausible, emotionally tuned content. LLM-generated language can be persuasive (Salvi et al., 2025), and humans can struggle to distinguish LLM text in social-media-like contexts (Dugan et al., 2024). Recent work also suggests increasing entanglement between social-network discourse and model development, including domain-specific post-training and evaluation targeting social platforms (Zhao et al., 2025).

Together, these strands motivate our evaluation problem: generation quality can outpace detection

reliability for the same socially consequential construct. If LLMs can write posts that induce upward or downward positioning but cannot stably recover that direction when classifying naturally occurring posts, they become potent generators yet unstable measurement instruments for auditing and moderation.

## 3 Task Definition

We study reader-perceived social comparison direction in text-only Xiaohongshu posts. Given a post text $x$, the task is to predict one label $y \in \{\text{UPWARD}, \text{NEUTRAL}, \text{DOWNWARD}\}$ from a first-person reader perspective, consistent with social comparison theory (Festinger, 1954). We map labels to numeric IDs for modeling and reporting: UPWARD $= 0$, NEUTRAL $= 1$, and DOWNWARD $= 2$. UPWARD indicates that the post positions the poster (or the life depicted) as better off than me; DOWNWARD indicates worse off than me; and NEUTRAL indicates similar to me or no clear invitation to compare.

Labels operationalize elicitation rather than sentiment: a post may be positive or negative in tone yet function as UPWARD, DOWNWARD, or NEUTRAL depending on the implied reader–poster relation. Because XHS-SCoRE is class-balanced, we report Accuracy and Macro-F1 as primary metrics. We additionally inspect per-class recall and confusion structure to diagnose directional failure modes, such as collapsing comparison-rich posts into NEUTRAL.

## 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 (from HK universities, age 18-24) 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); and (3) labels reflect the collector's immediate reader reaction during browsing, i.e., whether the post elicits comparison and, if so, in which direction.

The protocol treats comparison elicitation as partly reader-dependent: the same post may not elicit comparison for every reader. XHS-SCoRE therefore targets a psychologically grounded construct—reader-perceived comparison elicitation—rather than author intent.

### 4.2 Benchmark size and splits

XHS-SCoRE is balanced across the three labels (0=UPWARD, 1=NEUTRAL, 2=DOWNWARD). The benchmark contains 2,452,665 Chinese characters and 13,916 posts in total: 4,632 UPWARD, 4,631 NEUTRAL, and 4,653 DOWNWARD. Posts are randomized into fixed splits used for all comparisons. The TRAIN split contains 8,350 posts (2,780 UPWARD; 2,779 NEUTRAL; 2,791 DOWNWARD; 1,487,712 characters). The VAL split contains 2,783 posts (926 UPWARD; 926 NEUTRAL; 931 DOWNWARD; 496,996 characters). The TEST split contains 2,783 posts (926 UPWARD; 926 NEUTRAL; 931 DOWNWARD; 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. All non-raw, policy-compliant artifacts (label schema, prompts, scripts, aggregated results, and AI-generated examples) are released in a single repository; see Appendix A.

### 4.3 Corpus analysis

To characterize linguistic realizations of direction, we analyze an subset collected under the same elicitation protocol but restricted to UPWARD versus DOWNWARD. It contains 3,821 UPWARD posts (601,277 characters) and 3,734 DOWNWARD posts (679,307 characters). 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, UPWARD posts more often realize aspirational lifestyles through lexis and discourse associated with consumption, mobility, and positive evaluation, whereas DOWNWARD posts 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.

## 5 Models and Experimental Setup

### 5.1 Prompted LLM classifiers

We evaluate LLMs as zero-shot classifiers for UPWARD/NEUTRAL/DOWNWARD. Because labels are defined from a reader perspective ("the poster is better than me / similar to me / worse than me"), we prompt models with a first-person viewpoint aligned with a typical active youth user and require a single label based strictly on the post text. Prompts are in Simplified Chinese and constrained to JSON-only outputs; we set temperature $= 0.1$ to reduce sampling variability. For `gpt-5-2025-08-07`, we additionally set `reasoning.effort=minimal` to reduce extraneous reasoning tokens. The system instruction sets the reader persona, defines labels in plain language, and forbids explanations; the user message inserts the post text and repeats the JSON-only constraint. Full prompts appear in Appendix B.

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) as a state-of-the-art open comparison point (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 (BERT family)

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-style 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

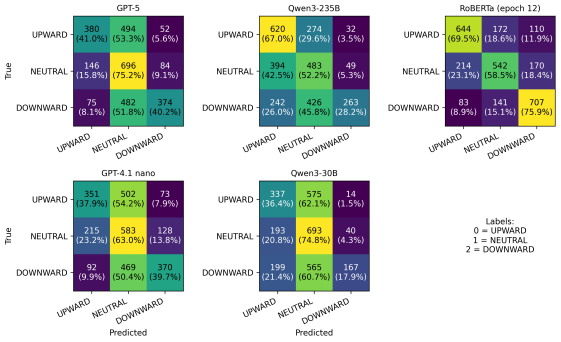


Figure 1: Confusion matrices for prompted LLM classifiers and the best-performing encoder baseline (test split).

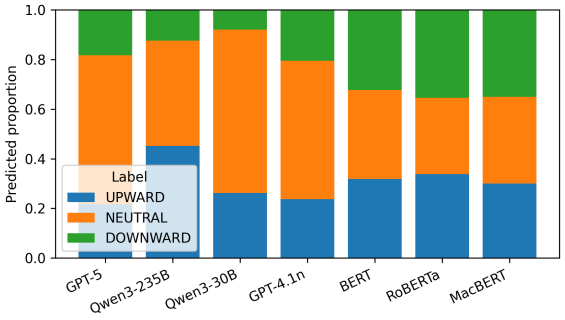


Figure 2: Predicted-label distribution by model (test split).

up to 15 epochs from pre-trained weights of Cui et al. (2020); we select the single best checkpoint by validation Macro-F1 and report its performance on TEST. Full hyperparameters, environment, and grid-search details are in Appendix D.

## 6 Classification Results

### 6.1 Main results on the test split

Table 1 reports test performance for prompted LLM classifiers and supervised encoder baselines; Figures 1–3 provide confusion matrices, predicted-label distributions, and per-class recall.

First, prompted LLMs underperform supervised encoders by a wide margin. The strongest LLM (GPT-5) reaches Accuracy $= .521$ and Macro-F1 $= .518$, while the best encoder (Chinese-RoBERTa-WWM-Ext) reaches Accuracy $= .680$ and Macro-F1 $= .679$. The signal is therefore learnable in-domain with standard encoders, yet remains unreliable under prompted LLM inference despite explicit perspective conditioning and strict output formatting.

Second, prompted LLMs exhibit systematic distortions that make comparison direction computa-

| Model | Type | Acc | Macro-F1 | Rec UP | Rec NEU | Rec DOWN | Pred 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 |
| C-BERT WWM | Encoder | 0.670 | 0.671 | 0.666 | 0.636 | 0.708 | 0.360 |
| C-RoBERTa WWM | Encoder | 0.680 | 0.679 | 0.695 | 0.585 | 0.759 | 0.307 |
| C-MacBERT Base | Encoder | 0.665 | 0.665 | 0.633 | 0.631 | 0.730 | 0.349 |

Table 1: Test performance. Labels: UPWARD= 0, NEUTRAL= 1, DOWNWARD= 2. "Pred NEUTRAL" is the proportion of outputs assigned NEUTRAL (diagnostic for neutral-collapse).

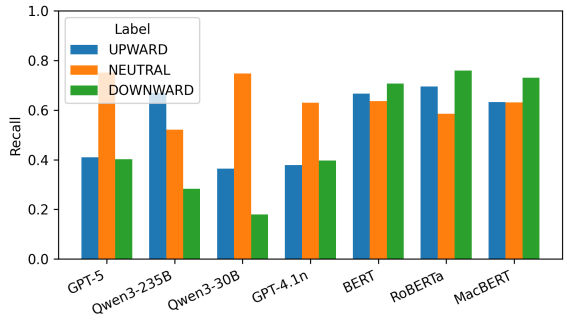


Figure 3: Per-class recall by model (test split).

tionally hard to detect, but the dominant shortcut differs by model. GPT-5 and GPT-4.1 nano are drawn to NEUTRAL, predicting it for 60.1% and 55.8% of items; Qwen3-30B shows the strongest NEUTRAL sink at 65.9% . Qwen3-235B instead over-predicts UPWARD (UPWARD outputs = 45.1%), misclassifying many NEUTRAL and DOWNWARD posts as UPWARD. These profiles indicate structured, model-specific mappings from implicit, platform cues to relational labels rather than undifferentiated noise.

### 6.2 Class-level behavior and the DOWNWARD sensitivity gap

Sensitivity to DOWNWARD triggers is a key practical concern because Xiaohongshu downward comparison often appears as low-agency hardship narratives and interpersonal conflict framing. Across LLMs, DOWNWARD recall is consistently weak: .402 (GPT-5), .397 (GPT-4.1 nano), .282 (Qwen3-235B), and .179 (Qwen3-30B). Thus, many true DOWNWARD posts are systematically reassigned to NEUTRAL (and, for Qwen3-235B, also to UPWARD), the pattern that would cause monitoring or auditing systems to miss psychologically potent downward-comparison cues.

Encoders are substantially stronger and more stable: DOWNWARD recall reaches .708 (Chinese-BERT), .759 (Chinese-RoBERTa), and .730 (Chinese-MacBERT), with Macro-F1 around .665–.679. They also avoid extreme NEUTRAL sinking: NEUTRAL prediction rates are 30.7–36.0%, close to balanced priors. This contrast supports the methodological conclusion implied by our title: LLM fluency and instruction following do not guarantee that pragmatics-heavy, reader-centered comparison cues are computationally detectable when labels encode implicit relational stance.

### 6.3 When comparison becomes invisible: error patterns beyond aggregate scores

#### 6.3.1 Neutralization of comparison triggers

For multiple prompted LLMs, the dominant error is to map comparison-triggering posts to NEUTRAL. For GPT-5, 53.4% of true UPWARD items (494/926) and 51.8% of true DOWNWARD items (482/931) are predicted NEUTRAL; GPT-4.1 nano is similar (54.2% UPWARD→NEUTRAL; 50.4% DOWNWARD→NEUTRAL). Qwen3-30B intensifies this pattern (62.1% UPWARD→NEUTRAL; 60.7% DOWNWARD→NEUTRAL). Even Qwen3-235B maps 45.8% of DOWNWARD items to NEUTRAL.

This matters because NEUTRAL is not an "unknown" label: it asserts no clear invitation to self–other ranking. Neutralization therefore makes comparison cues computationally invisible by collapsing implicit direction into an explicit non-comparison judgment. Encoders, by predicting NEUTRAL at about 0.31–0.36 (near priors), preserve higher recall for both UPWARD and DOWNWARD. The direction signal is present in text, but prompted LLM inference repeatedly fails to detect it.

#### 6.3.2 Directionality asymmetry and DOWNWARD sensitivity

LLM errors are also directionally asymmetric: DOWNWARD is consistently harder than UPWARD for several models. For Qwen3-235B

and Qwen3-30B, DOWNWARD recall falls to .282 and .179, indicating that most true DOWNWARD items are neutralized or redirected; even the stronger closed models plateau around .40. Encoder baselines maintain much higher DOWNWARD recall (≈.708–.759) on the same split.

This asymmetry is revealing because many DOWNWARD items encode comparison through agency and relational positioning rather than explicit comparative statements. The encoder advantage indicates learnable textual traces, while prompted LLMs often fail to detect the implied relational stance.

#### 6.3.3 A distinct LLM profile: UPWARD over-attribution

Not all failures are neutralization. Qwen3-235B over-predicts UPWARD, assigning UPWARD to 42.6% of true NEUTRAL items (394/926) and 26.0% of true DOWNWARD items (242/931). This suggests a shortcut where an aspirational platform register is treated as evidence of upward comparison even when the post is informational, self-contained, or not framed as a benchmark for the reader. Under this profile, comparison direction is not collapsed away; it is systematically skewed toward UPWARD.

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

Overall, the LLM gap is not only lower aggregate accuracy: it reflects structured confusions that compress reader-perceived relational meaning. For several models, comparison cues become computationally invisible through neutralization, turning comparison-triggering posts into non-comparison judgments; for others, cues remain visible but are directionally distorted through UPWARD over-attribution. Encoders provide an in-domain control showing that the signal is learnable and that LLM failures are not irreducible subjectivity. This is the sense in which comparison is "not detected": not because cues are absent, but because prompted LLM inference repeatedly collapses or skews the relational mapping the task requires.

## 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 our corpus-linguistic findings. The goal is construct-targeted generation rather than generic "realistic text": the UPWARD and DOWNWARD conditions are designed to differ in stance, agency framing, evaluative lexis, and list-like abundance cues, while remaining plausible as platform posts. Full prompts and constraint lists are in Appendix B, with examples in Appendix C.

Downward prompts reproduce a conflict- and low-agency narrative profile, including dense pronouns and reported speech, negation, negative-affect intensifiers, passive constructions (e.g., 被), and occasional explicit "others" anchoring. Upward prompts target aspirational abundance framing via positive evaluatives, superlatives, and list-like punctuation patterns. Neutral prompts suppress personal affect and self–other positioning, focusing instead on informational topics (e.g., weather, recipes, product descriptions). To avoid unnaturally polished outputs, we inject minor "authenticity noise" (small language and punctuation imperfections).

## 8 Human Validation: LLM-Generated Posts Elicit Social Comparison

Because social-comparison direction (UPWARD/DOWNWARD/NEUTRAL) 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 ($N = 29$) 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: UPWARD ($N = 10$), DOWNWARD ($N = 9$), and a NEUTRAL 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 that readers infer the intended comparison direction from the generated posts.

**Comparison-related emotions.** Downward assimilative emoition is substantially higher in the DOWNWARD condition than in the other two conditions combined (DOWNWARD: $M$=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: DOWNWARD $>$ NEUTRAL $>$ UPWARD. 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 UPWARD-, DOWNWARD-, or NEUTRAL-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 NEUTRAL, 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 results expose a dissociation between NLP and behavioral science. LLMs can generate Xiaohongshu-style posts that evoke comparison-related affect in human readers, yet fail to detect the same reader-grounded construct when asked to classify naturally occurring posts. On the detection side, prompted LLMs do not merely make occasional mistakes: they exhibit structured confusions—most prominently collapsing comparison-triggering posts into NEUTRAL and, for some models, systematically skewing predictions toward UPWARD. These are not random errors but stable mappings that compress or distort relational meaning.

The risk, therefore, is not simply that "LLMs are imperfect classifiers." It is that meaning can be behaviorally present while being computationally invisible under LLM-prompted detection. This matters because many downstream uses (auditing exposure, prioritizing moderation review, measuring platform psychological impact) depend on reliable identification of relational triggers, not on generation fluency. If detectors undercount where comparison is happening or misread its direction, they will misestimate prevalence and mischaracterize which content is likely to shape reader experience.

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

A common evaluation shortcut is to treat strong generation as evidence of understanding. Our find-

ings challenge that inference for reader-grounded relational meaning. The task is not sentiment or topic; it requires mapping text to a comparative relation between poster and reader ("better than me / worse than me / similar or unclear"), often expressed through stance and narrative positioning rather than explicit comparatives. In this regime, prompted LLMs frequently fail while supervised encoders trained in-domain recover the signal far more reliably.

Methodologically, this implies that evaluation for socially embedded constructs must probe where meaning disappears, not only how often predictions match labels. Two diagnostic failure patterns recur: (i) neutralization (systematic UPWARD/DOWNWARD→NEUTRAL collapse) and (ii) directional skew (e.g., UPWARD overattribution under aspirational registers). A detector that defaults to NEUTRAL may appear "cautious," but it is making a strong claim (no clear comparison) with predictable false negatives. For relational-meaning benchmarks, reporting class recall and confusion structure is therefore part of the evaluation target, not an optional analysis.

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

The dissociation points to a governance asymmetry: the cost of producing psychologically potent triggers is falling, while the reliability of detecting them remains uncertain. If LLMs can generate comparison-inducing posts at scale, comparison triggers may be amplified not only by recommender dynamics but also by synthetic content supply. Yet if automated monitors systematically neutralize these cues, platforms develop a blind spot precisely where intervention would matter.

This is especially salient for youth-dominated, lifestyle-oriented platforms where content is highly rankable and comparison cues are embedded in mundane narration (family, consumption, travel, schooling). Our human validation shows that brief exposure can shift perceived standing and produce measurable comparison-related affect. A monitoring pipeline that collapses such posts into NEUTRAL will undercount exposure and underestimate the psychological externalities of both organic and synthetic content. The operational implication is direct: comparison-trigger detection should be treated as a dedicated capability requiring validation, not assumed from general-purpose prompted LLM classification.

### 9.4 Implications for social science methods: controlled generation, human grounding

For computational social science and psychology, the results cut both ways. On one hand, LLMs enable scalable, controlled stimulus construction: prompts can be tuned to corpus-derived linguistic profiles (e.g., low-agency conflict narratives versus abundance/list framing) while remaining ecologically plausible. This offers a principled route from corpus analysis to experimental materials.

On the other hand, the same dissociation shows that LLM-in-the-loop research requires human grounding at two points. Generated stimuli must be validated to ensure they instantiate the intended construct (as we do via manipulation checks and emotion outcomes). And when LLMs are used as annotators or detectors, their agreement cannot be treated as ground truth for reader-defined labels. In short, LLMs can manufacture psychologically potent artifacts while remaining unreliable instruments for measuring the human meanings those artifacts induce. This divergence is precisely what makes reader-grounded benchmarks and human validation essential when NLP outputs support claims about psychological or societal impact.

## 10 Conclusion

We introduce XHS-SCoRE, a reader-grounded benchmark for detecting whether Xiaohongshu posts elicit UPWARD comparison, DOWNWARD comparison, or no clear comparison. Across four prompted LLM classifiers, we observe structured failures dominated by neutralization and directional skew, even with perspective conditioning and strict output constraints; supervised Chinese encoder baselines trained in-domain recover comparison direction substantially more reliably. A controlled human study further shows that LLM-generated Xiaohongshu-style posts, engineered from corpus-derived constraints, reliably shift perceived standing and comparison-related affect. Together, the evidence supports a central conclusion: LLMs can generate psychologically potent comparison cues that remain computationally unreliable to detect under prompting, so generation quality should not be treated as evidence of reliable relational understanding.

## Limitations

The human validation is a pilot (N=29) conducted in a lab setting with brief exposure; larger and more diverse samples and longer browsing sessions are needed for stronger external validity. The dataset targets Xiaohongshu discourse and Chinese sociocultural context; generalization across platforms and languages remains open. Because Xiaohongshu is multimodal, text-only posts omit images/video and engagement signals that may amplify comparison triggers. Prompted LLM classification can vary with model updates and prompt framing; we reduce variability with low temperature and strict outputs, but instability remains a methodological constraint.

## 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.

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.

Philippe Verduyn, Nino Gugushvili, Karlijn Massar, Karin Täht, and Ethan Kross. 2020. Social comparison on social networking sites. *Current Opinion in Psychology*, 36:32–37.

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-5.1-Codex-Max and ChatGPT-4o is being used to assist coding, all codes are reviewed and approved by the authors.

# B Prompts and Constraints

## B.1 Classification prompts

System prompt (reader persona and labels):

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 template (JSON-only output):

帖子：
`{post_text}`

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

## B.2 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 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=$2\times10^{-5}$, RoBERTa lr=$2.5\times10^{-5}$, MacBERT lr=$3\times10^{-5}$; 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.