CAN LLMS THINK LIKE CONSUMERS? BENCHMARKING CROWD-LEVEL REACTION RECONSTRUCTION WITH CONSUMERSIMBENCH

Anonymous authors
Paper under double-blind review

ABSTRACT

LLMs are increasingly used as "digital consumers" to simulate public opinion, pre-test marketing decisions, and anticipate audience response. However, existing evaluations rarely ask whether a model can reconstruct the concrete reaction patterns that real consumers surface in public discourse. We introduce CONSUMERSIMBENCH, a benchmark built from 1,553 real Chinese social-media topics and 23,122 atomic, rule-audited criteria spanning four reaction families. Rather than scoring open-ended generations with a holistic preference judge, CONSUMERSIMBENCH decomposes each task into auditable yes-no decisions over concrete reaction points, raising three-judge agreement from 65.8% to 92.1% with 98.4% agreement between pointwise judge decisions and human-majority labels. Across 13 frontier generators, the strongest model, Gemini-3.1-Pro, covers only 42.5% of real reaction criteria, while GPT-5.2 and Claude-4.6 trail far behind despite their strength on technical benchmarks. The failures reveal a sharp gap between technical-benchmark performance and socially grounded consumer intuition. Our structured prompt, CONSUMERSCF, yields 0.6-3.8 percentage points higher coverage, while generate-reflect multi-agent refinement improves MiMo-V2.5-Pro from 32.9% to 37.6%. CONSUMERSIMBENCH reframes consumer simulation as a forecasting problem over real public-discourse reactions, showing that frontier LLMs remain far from reliably predicting what consumers will actually care about in high-context Chinese consumer discourse.

1 INTRODUCTION

> *"If an LLM can truly think like a consumer,*
> *it should be able to predict what a real user would say about a brand."*

The vision of LLMs acting as *digital consumers* is rapidly moving from demos to deployment. Teams increasingly use LLM agents to simulate social opinion (Yang et al., 2024a; Piao et al., 2025; Zhang et al., 2025b), pre-test marketing messages (Wang et al., 2025a; Chu et al., 2025), and design social strategy (Mou et al., 2024; Qu & Wang, 2024). The motivation is straightforward: traditional consumer research is slow and expensive, while large-scale scalar signals such as click-through rate (CTR) are post-hoc, platform-specific, and often fail to explain *why* consumers react the way they do.

However, downstream decision-makers still lack a principled understanding of **how closely LLM simulations align with actual human behaviors** in high-context, strategic, and socially intensive environments (Wang et al., 2025c). For example, on lifestyle-oriented user-generated-content (UGC) platforms such as RedNote or Youtube, creators often rely on pragmatic moves such as reverse signaling, humblebragging, and anxiety induction to shape how audiences interpret a post. In such settings, surface-level understanding (e.g., *the text expresses sadness*) is insufficient; what matters is **strategic social intelligence** (e.g., *the YouTuber creates a sad atmosphere for his advertisement*), which requires understanding the narrative stance and how it is meant to shape the audience's inference. This can fundamentally reshape the marketing industry, because reliably predicting consumer reactions before publication could defuse the next viral crisis instead of apologizing for it, and identify breakout angles before chasing them, which remains an extremely difficult mission even for marketing experts.

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

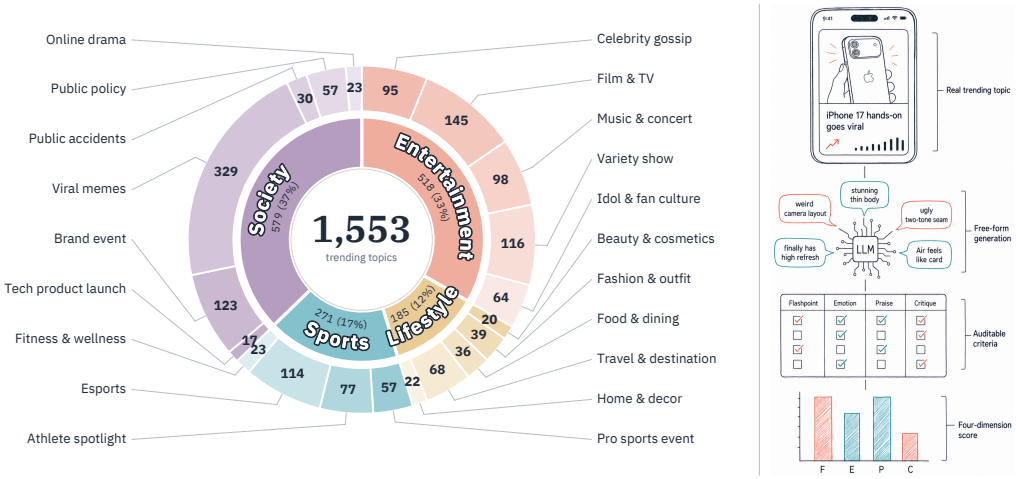


Figure 1: **Overview of CONSUMERSIMBENCH .** *Left:* The 1,553 trending topics span four super-categories and twenty consumer-facing sub-fields. *Right:* One task instance. Given a real trending topic, a generator must produce a free-form bundle of consumer comments that collectively cover an audited set of atomic criteria across four reaction families (flashpoints, emotion, praise, critique).

Existing benchmarks are inadequate for three reasons. **First**, real UGC does not follow the logic in standard ToM-style tests. ToM examples are typically built around a strict reasoning trace (e.g. *I stole the hat. He did not see. So, he doesn't know the hat is missing.*) (Chen et al., 2025b), **where the key facts are explicit and the answer space is closed**. In contrast, real UGC's meanings can be layered, implicit, and involve satire or boasting (e.g. *It is so annoying to decide which house to decorate!*). **Second,** many benchmarks emphasize the prediction of a choice, distribution, scalar number, or fixed user trace (e.g., CTR, SSR (Chen et al., 2025a; Maier et al., 2025), public-opinion distributions (Miranda & Balbi, 2025), social-media engagement labels and conditioned comment prediction (Bojic et al., 2026; Schwager et al., 2026), popularity prediction (Wu et al., 2023a), etc.), which cannot evaluate models' open-ended humanities-side capacity (Hu et al., 2025). Practitioners care about *what real people will say and why*; most benchmarks only measure *what label a model predicts*. **Third,** many evaluations rely on an LLM to directly assign a holistic final score for an open-ended output, which can be subjective, unstable, and vulnerable to systematic biases (Tu et al., 2024; Liu et al., 2024a; Gu et al., 2024; Chen et al., 2024a). This makes it difficult to interpret what a numeric score means and whether it is reliable across models, prompts, and domains.

Meanwhile, despite LLMs' top performance on quantitative leaderboards (Wang et al., 2024c; Jimenez et al., 2023; Merrill et al., 2026), **social-media users continue to complain that these models are weak at the "humanities" side of language**—reading people, registering social nuance, and showing human touch. Part of this perceived gap is a benchmarking artifact, where very few benchmarks audit a model's humanities-side skills.

To bridge these gaps, we introduce CONSUMERSIMBENCH, a benchmark designed to evaluate whether LLMs can simulate **authentic consumer behaviors** in the wild. We use RedNote as a major consumer-facing UGC platform where real reactions to products, brands, entertainment releases, lifestyle trends, and public events are observable at scale. Its dense, high-context consumer discourse makes it a representative and practically important setting for testing whether models can anticipate what consumers will notice, praise, attack, and emotionally amplify. Concretely, given a trending topic and its event description, a model is required to generate realistic comments. We audit coverage over four reaction families: **(i)** *sentiment flashpoints*, the concrete social triggers; **(ii)** *emotion keywords*, the overall sentiment of the real crowd; **(iii)** *positive aspects*, what users praise; **(iv)** *negative aspects*, what users attack. Every criterion is atomic, defined, example-anchored, and rule-audited. We find that even the strongest model (Gemini-3.1-Pro: 42.5%) performs poorly on this humanities task, and top-tier models that look comparable on technical benchmarks diverge sharply on CONSUMERSIMBENCH (GPT-5.2: 28.7%, Claude-Opus-4.6: 27.0%). This gives measurable feedback on a humanities-side capability that current LLM training and evaluation often underemphasize. **Our main contributions are:**

- To study whether an LLM can *truly think like a consumer*, we propose **CONSUMER-SIMBENCH**, a UGC-behavior simulation benchmark that operationalizes this question as

crowd-level reaction reconstruction: given a real trending topic and event description, can the model reconstruct the reaction patterns that real consumers surfaced in public discourse?

- To make open-ended consumer-likeness measurable, CONSUMERSIMBENCH contains 1,553 real topic instances and 23,122 atomic scoring criteria that jointly evaluate emotion keywords and emotional bursts **at the macro level**, and content-faithful expressions of individual consumer feedback **at the micro level**.
- For automatic evaluation, we construct detailed criteria that turn open-ended scoring from holistic "judge the answer" decisions into **auditable yes-no questions**. This yields 92.1% three-judge agreement, reaches 98.4% agreement with human-majority labels, and outperforms holistic LLM-as-Judge scoring (65.8%) and traditional similarity metrics such as Jaccard and N-gram overlap.
- On a 100-topic subset, our structured prompt, CONSUMERSCF, yields 0.6-3.8 percentage points higher coverage without changing model ordering. Generate-reflect multi-agent refinement improves MiMo-V2.5-Pro from 32.9% to 37.6%.

2 RELATED WORK

Table 1: Positioning by *unit of inference*. Consumer simulation requires open-form, crowd-level reconstruction of real public discourse with auditable targets. Among the benchmark families summarized here, CONSUMERSIMBENCH is the only one that satisfies all five criteria jointly.

Benchmark family	Unit of inference	Real public discourse	Free-form generation	Strategic targets	Auditable eval
ToM (Kim et al., 2023; Xu et al., 2024; Chen et al., 2025b)	Dyad (Alice / Bob)	✗ (fictional)	✗ (MCQ)	△	△
CTR (scalar prediction)	Aggregate scalar	✓	✗	✗	✓
Persona Simulation (Park et al., 2023; Wang et al., 2025a)	Individual trajectory	△	✓	△	✗
CONSUMERSIMBENCH(Ours)	**Crowd reaction**	✓	✓	✓	✓

Persona simulation. Recent work has explored using LLMs as proxies for human behavior in opinion, market, and psychological studies (Aher et al., 2023; Lin, 2026; Qi et al., 2025; Ludwig et al., 2026; Wang et al., 2026). Generative Agents (Park et al., 2023) demonstrated that LLM-powered agents with memory, planning, and reflection can exhibit emergent social behaviors in a simulated town. Subsequent work has extended this paradigm to social-science experiments (Hewitt et al., 2024; Anthis et al., 2025; Zakazov et al., 2024; Xiao & Wang, 2025; Slumbers et al., 2025), user-behavior simulation (Wang et al., 2025a; Chu et al., 2025; Yang et al., 2024a; Piao et al., 2025; Zhang et al., 2025b; Yang et al., 2024b; Zhou et al., 2023), cognitive reasoning and human-response modeling (Zhou et al., 2025; Hwang et al., 2025; Binz et al., 2025), customer simulation (Wang et al., 2025d; Wang et al.), marketing simulation (Brand et al., 2023; Agarwal, 2025; Yang et al., 2025; Li et al., 2024; Yu et al., 2024; Su et al., 2026), opinion dynamics and pluralism (Chuang et al., 2025; Poole-Dayan et al., 2025), multi-turn user simulation (Dou et al., 2025; Seshadri et al., 2026), war simulation (Hua et al., 2023), comprehensive systems and agentic public-opinion analysis (Gao et al., 2023; Tang et al., 2025; Zhang et al., 2025a; Chen et al., 2024c; Wang et al., 2025b; Liu et al., 2025a), and others (Gao et al., 2024; Chen et al.). Critical studies also warn that synthetic respondents can fail systematically as replacements for human survey data (Bisbee et al., 2024). Across these works, the similarity between LLM agents and real human social behavior is mostly evaluated through classical psychology, survey, or social-science experiments, which risk data contamination or being solved by general common sense rather than genuine open-ended reaction reconstruction.

Evaluating LLMs' mindset. Current benchmarks mainly rely on multiple-choice (Sap et al., 2019; Lee et al., 2025; Zeng, 2025; Chen et al., 2024b; Marraffini et al., 2024; Anonymous, 2025; Yong et al., 2025; Sabour et al., 2024; Kang et al., 2025), classification (Hendrycks et al., 2021), LLM-as-a-Judge (Shao et al., 2023; Tu et al., 2024; Wang et al., 2024b; Samuel et al., 2025; Klinkert et al., 2024; Ong et al., 2025; Mou et al., 2025), and similarity metrics like Rouge-L (Wang et al., 2024a) and BERTScore (Emelin et al., 2021). Beyond emotional and cognitive benchmarks (Chen et al., 2024d; Liu et al., 2024b; Gandhi et al., 2023; Echterhoff et al., 2024; Jia et al., 2024), a parallel line of work uses Theory-of-Mind (ToM) tasks (Chen et al., 2025b) to track what one agent believes another believes, in stylized, closed-form, typically multiple-choice scenarios (Shinoda et al., 2025; Kim et al., 2023; Xu et al., 2024; Strachan et al., 2024; Wu et al., 2023b; Shi et al., 2025; Liu et al., 2025b; Zhang et al., 2024). While these benchmarks illuminate a cognitive competence, they test reasoning ability rather than the actual emulation of human social behavior, and their scoring is

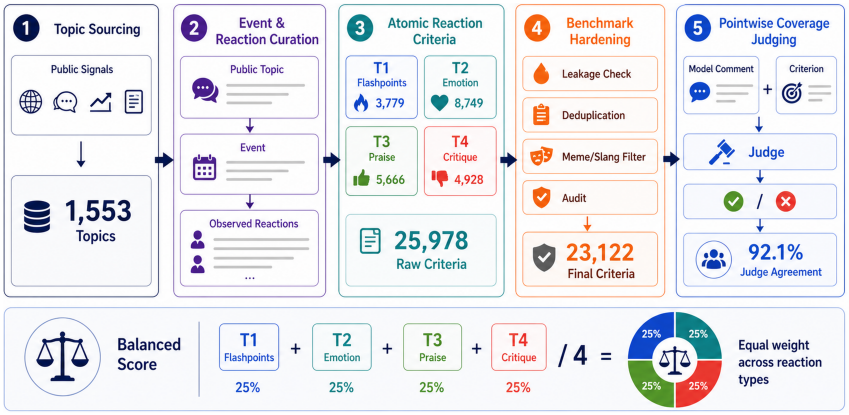


Figure 2: **CONSUMERSIMBENCH construction and evaluation pipeline.** Public trend signals are curated into topic-event records and abstracted observed reactions; these reactions are converted into four families of atomic reaction criteria, hardened, and finally used for pointwise coverage judging of model-generated comments. The final score gives equal weight to the four reaction families.
largely subjective. Our work advances benchmarking on humanities-side capabilities with auditable criteria even when the answer is not unique.

Position. As shown in Table 1, CONSUMERSIMBENCH targets a different competence from ToM. ToM asks whether a model can ***reason*** about a mind; CONSUMERSIMBENCH asks whether a model can ***reconstruct*** what a crowd will think, feel, and say about a real brand event.

3 CONSUMERSIMBENCH DATASET DESIGN

3.1 DESIGN GOALS

CONSUMERSIMBENCH targets four design goals. **1. Authenticity over generic plausibility:** Rather than synthesizing diverse scenarios, we ground every record in real public reactions. Models must reconstruct the salient reaction points real consumers surfaced about a real brand, product, or public event, not merely sound emotional; this evaluates public-trend reaction coverage rather than median survey sentiment. **2. Free-form outputs:** A model should not merely classify sentiment into a pre-set category (Argyle et al., 2023); it should simulate what a crowd of consumers would plausibly say. **3. Derived artifacts:** For responsible release, we publish derived, anonymized artifacts rather than raw user histories or identifiers. **4. Auditable scoring:** Instead of asking a vague *"How good is this answer?"* question like what LLM-as-Judge benchmarks do ("40% human-like" is meaningless), we change it into a series of auditable decisions: *"Did the generated comments cover this particular reaction point?"*, where every point is an inspectable criterion with definitions, examples, and rules. The result is far more stable and easier to interpret than a global similarity score.

3.2 TASK FORMULATION

Each topic has a topic keyword k and a neutral event description d. Given the i-th topic $\{k_i, d_i\}$, an LLM π is required to produce n_i (10-30) realistic comments. The comments will be graded from four dimensions, each contributing 25% of the total score:

- **T1 sentiment flashpoints**: The emotional trigger that real consumers resonate with, which can be one or many phrases. This is usually the reason that pushes the topic onto the trending list. Accurately predicting these triggers, both positive and negative, can help brands avoid reputation crises while identifying possible viral hits. For example, the topic *Arc'teryx's fireworks* has a flashpoint named *environmental protection*, where people's concern about the fireworks' environmental impact later led to a public-relations crisis. The grading process judges how many flashpoints are clearly identified, and the score is the ratio.
- **T2 emotion keywords**: The affective vocabulary that best describes people's attitude in the comments (such as *resentful*, *conflicted*, *put on the spot*). They can cover many aspects, and

all the words are open-form. There are no preset categories. The grading process identifies how many emotions are present in the generated posts, using semantic coverage rather than exact word matching.

- **T3 positive aspects:** The specific *details* appearing in the original event that consumers like, praise, or upvote (such as *cost-effectiveness* and *emotional value*). The marketing industry can use this to test if their proposed selling point will lead to the intended response.
- **T4 negative aspects:** Similar to T3, these are the details on which criticism occurs against the original event (such as *discrimination* or *affectation*). A capable model on this can help the marketing industry avoid inappropriate phrasing in their post.

Within each section, the score is

$$s_{i,j} = \frac{\#\text{ of matched criteria in } T_j}{\#\text{ of total criteria in } T_j} \quad (1)$$

and the final total score is

$$s = \frac{1}{N}\sum_{i=1}^{N} \frac{s_{i,1} + s_{i,2} + s_{i,3} + s_{i,4}}{4}. \quad (2)$$

Atomic Criteria. Each of the four dimensions decomposes into atomic criteria that can be judged independently. As shown in Table 2, every atomic criterion contains five fields: **1.** the corresponding trigger / keyword / aspect; **2.** a detailed definition; **3.** positive examples that meet the criterion; **4.** negative examples that do not; **5.** a one-sentence judgment rule that disambiguates edge cases. A single comment may cover multiple criteria, and evaluation is over the concatenated comment set rather than one-to-one post matching.

For each atomic criterion, we ask a judge model whether the generated comments cover that criterion. To keep evaluation fair across generators that produce different numbers or lengths of comments, we concatenate each model's comments and truncate the judge-visible text to 3,600 characters, the estimated empirical 99% coverage bound, and adapt n_i to follow the empirical comment-volume. The judge receives all the above criteria, and returns a binary score and an explanation. The benchmark score is interpretable because every point corresponds to a concrete, inspectable decision.

Latent reaction points. CONSUMERSIMBENCH is a forecasting benchmark, not an entailment benchmark. We do not restrict criteria to reactions directly implied by the event description alone. We therefore include reaction points that real users surfaced through platform conventions, prior brand knowledge, local memes, demographic identity, and recent discourse.

Table 2: Example T1 flashpoint criterion for a product-recall topic. All 23,122 criteria follow this five-field structure.

Field	Content
Trigger	"Apology lacks sincerity"
Definition	Users question whether the public apology is genuine or merely a tactic.
Positive example	"This reads like their PR team wrote it"
Negative example	"I don't like this brand"
Judgment rule	Credit only if the comment questions the *sincerity or authenticity* of the apology, not general brand criticism

3.3 DATASET COLLECTION AND CONSTRUCTION

Data Collection. We source data from RedNote (Xiaohongshu), a Chinese social platform with 300M+ monthly active users focused on lifestyle and consumer content, providing rich naturalistic context for consumer behavior. RedNote's trending-topic lists and topic summaries are published by multiple third-party public aggregators (e.g., `tophub.today`, `qian-gua.com`). We obtain hot-topic keywords and their public summaries from these aggregators and through manual collection. We only use publicly available content. We do not scrape or crawl RedNote, do not use any undocumented platform API, and do not access raw user posts, account identifiers, or any logged-in views. Each topic is verified by an actual public-trending search on the corresponding date, and the resulting artifact contains no quoted user text and no information by which an individual user could be identified. Collection complies with each aggregator's posted terms of service.

After this collection step we apply LLM-based filtering to retain only topics that are related to consumer reactions, and we manually sample 100 topics to verify the filtering quality.

Criteria drafting. For each topic, an LLM proposes T1/T2/T3/T4 criteria with definitions, positive/negative examples, and judgment rules, along with objective topic descriptions. Each proposed criterion is anchored in observed source reactions, while an LLM standardizes those reactions into auditable records with explicit decision boundaries in Table 2. Appendix D summarizes the operational drafting prompt.

Criteria hardening. We harden the drafted criteria to remove event-restatement leakage, semantic duplication, over-pruning, and judge-fragile meme points. The process merges similar topics and criteria, drops criteria that collapse to the event description, checks within-topic independence, and audits pruned items. Appendix E gives the full hardening protocol and examples. After hardening and pruning, the released **evaluation set contains 23,122 criteria**. All headline leaderboard scores and error analyses in the paper are reported on this final evaluation set.

4 Experiments and Results

4.1 Dataset Statistics

We collected hot topics between March and September 2025. The final release contains 1,553 topics with 23,122 criteria. We release the canonical Chinese benchmark, English sidecars for reviewer inspection, judge prompts, and the evaluation harness under CC BY-NC. Official evaluation uses the Chinese criteria for judge consistency, and the released criteria contain abstracted reaction-type descriptions rather than raw user content or account identifiers. Additional statistics are shown in Table 3.

Table 3: Benchmark statistics.

Statistic	Value
Topic instances	1,553
Atomic criteria (total)	23,122
Sentiment flashpoints	3,779
Emotion keywords	8,749
Positive aspects	5,666
Negative aspects	4,928
Average criteria per topic	14.9
Median criteria per topic	14

Topic coverage. The 1,553 topics span four categories and twenty consumer-facing sub-fields (Figure 1, left). Society and public-issue topics dominate (37%) because they generate the highest engagement on the platform, followed by entertainment (33%), sports and gaming (17%), and lifestyle goods (12%). This breadth makes domain-specific strengths visible rather than letting any single topic family dominate the leaderboard.

Candidates. We evaluate Gemini-3.1-Pro, Gemini-3-Flash, Qwen3.5-397B, Grok-4.2, DeepSeek-V3, Doubao-Seed-1.8-Thinking, Kimi-K2.5, GPT-5.2, MiniMax-M2.5, Claude-Opus-4.6, MiMo-V2.5-Pro, Claude-Sonnet-4.6, and GPT-4o.

Judge. We use GPT-5.2 as the fixed judge for cost-effectiveness; §5.3 validates judge reliability with cross-judge and human-audit checks.

4.2 Main Results

We now answer the opening question empirically. Can LLMs think like consumers, in the sense of reconstructing what a crowd will say in real public discussion? The headline is **not yet, and the remaining gap is not uniform**.

Figure 3 shows the full leaderboard. The best model, Gemini-3.1-Pro, reaches 42.5% overall, followed by Gemini-3-Flash at 41.7%. Grok-4.2, Kimi-K2.5, Qwen3.5-397B, DeepSeek-V3, and Doubao-Seed-1.8 form a tight 37-40% middle tier. MiMo-V2.5-Pro ranks eighth at 34.4%, followed by GPT-5.2 at 28.7%; MiniMax-M2.5, Claude-Opus-4.6, Claude-Sonnet-4.6, and GPT-4o stay at 22-28% in this setting. Rank order is stable under bootstrap resampling, and the broad tiering is stable under the criteria-weighted micro-average. The gap is not regional, and the hard sections (T1 flashpoints and T4 criticisms) remain difficult across the leaderboard.

Despite having competitive performance on technical benchmarks, top-tier models (such as Gemini-3.1-Pro, GPT-5.2 and Claude-Opus-4.6) **have a significant performance gap** in mimicking consumer behaviors. Gemini is 13.8 points above GPT-5.2 and 15.5 points above Claude-Opus-4.6, giving measurable form to a common user-facing complaint about emotional value, namely that some

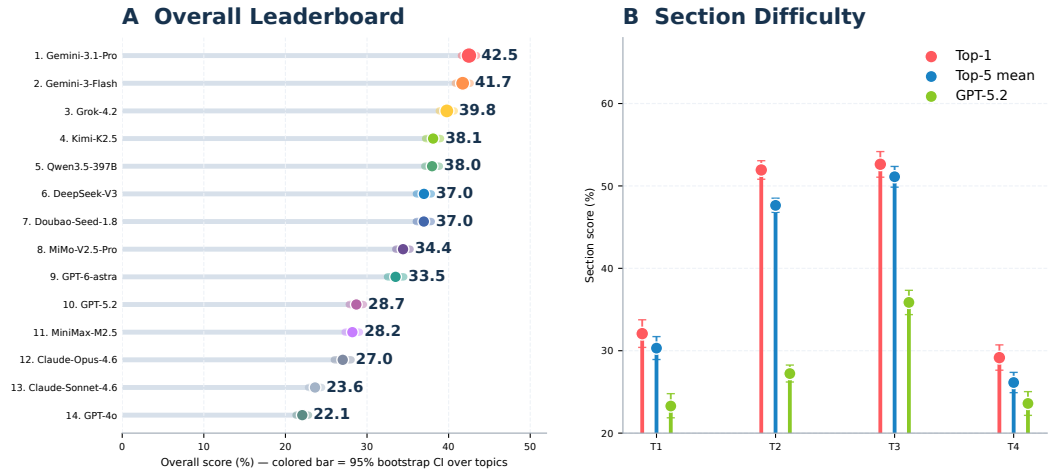


Figure 3: Main results. Left: overall leaderboard on the final full benchmark. Right: section-wise difficulty, comparing the top model, the top-5 mean, and GPT-5.2. Flashpoints and negative aspects remain the hardest families.

Table 4: Full leaderboard on the final benchmark (1,553 topics, 23,122 criteria). Scores are percentages (judge GPT-5.2). Overall scores include 95% bootstrap CIs over topics; the broad tier separation is stable under topic bootstrap and criteria-weighted micro-averaging.

	Model	AVG(95% CI)	T1	T2	T3	T4
1	Gemini-3.1-Pro	**42.5** [41.6, 43.4]	32.1	**51.9**	52.6	29.2
2	Gemini-3-Flash	41.7 [40.8, 42.6]	**32.3**	48.9	**55.6**	25.3
3	Grok-4.2	39.8 [38.9, 40.7]	29.0	48.1	50.4	25.4
4	Kimi-K2.5	38.1 [37.2, 39.0]	29.2	44.2	45.9	29.1
5	Qwen3.5-397B	38.0 [37.1, 38.9]	29.0	45.1	51.1	21.7
6	DeepSeek-V3	37.0 [36.0, 37.9]	27.0	42.9	44.6	**29.4**
7	Doubao-Seed-1.8	37.0 [36.0, 37.9]	28.0	45.6	49.4	20.1
8	MiMo-V2.5-Pro	34.4 [33.5, 35.3]	27.4	39.7	45.3	20.1
9	GPT-6-astra	33.5 [32.5, 34.5]	26.8	37.7	36.4	25.3
10	GPT-5.2	28.7 [27.8, 29.6]	23.3	27.2	35.9	23.6
11	MiniMax-M2.5	28.2 [27.3, 29.1]	20.2	33.0	36.3	16.9
12	Claude-Opus-4.6	27.0 [26.0, 28.0]	17.8	31.0	32.1	16.9
13	Claude-Sonnet-4.6	23.6 [22.8, 24.4]	15.5	24.9	33.6	13.3
14	GPT-4o	22.1 [21.3, 22.8]	14.3	27.8	30.4	8.9

frontier assistants can produce fluent comments while still missing the affective targets and criticism vectors that make public reactions feel human. Moreover, **all models remain far from the ceiling**. Even the strongest model leaves over half of real consumer reactions uncovered.

4.3 THE ASYMMETRIC FAILURE: TONE WITHOUT TARGET

We find that **current LLMs can partially *"sound"* like consumers, while failing to predict what consumers will focus on or attack.** Across all 13 models, average T2/T3 coverage exceeds average T1/T4 coverage by 8-24 points. For the top model, section scores have a 23-point spread. It is a *crowd-forecasting error*, where models often sound socially fluent, but fail to recover the concrete objects of attention, such as the meme phrases, product details, regional constraints, or business-model facts that become the public hook. For downstream marketing use cases, this is the decisive failure, because crisis forecasting and pre-launch risk assessment depend precisely on anticipating those hooks. The same T1/T4 gap persists across model tiers. The same asymmetric pattern appears in a second-platform YouTube experiment (Appendix G, Figure 6), where the construction is instantiated on 1,000 predominantly English trending videos and again finds positive aspects easiest while flashpoints and negative aspects remain hardest.

Figure 4 further checks whether the failure is an artifact of one domain, and finds the gap is broad rather than domain-local. Model scores vary across the 20 sub-fields, but every sub-field remains far from ceiling, and the relative rankings are stable across domains. This suggests that the gap is not reducible to a single topic family such as celebrity gossip or tech launches.

5 Ablations and Robustness

5.1 Where does the error come from?

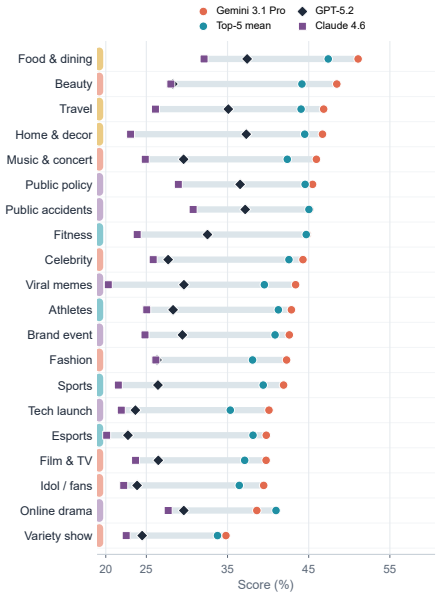


Figure 4: Ranked sub-field gap plot across 20 consumer-facing domains. Rows are sorted by Gemini-3.1-Pro score; horizontal spans show the cross-model range.

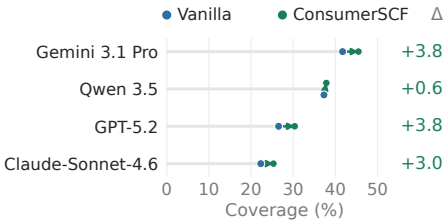


Figure 5: ConsumerSCF achieves higher coverage. Arrows connect vanilla prompting (blue) and ConsumerSCF (green).

We classify all 197,790 missed criteria across the 13 generators by parsing judge explanations into an 8-category taxonomy (Appendix B; the categorization is derived from keyword patterns in judge text plus a 50-sample manual audit). Using this taxonomy, we find that **missing-content patterns dominate the error taxonomy**. Generations are topically relevant but do not surface the specific reaction trigger (44.5% of misses), section-specific vocabulary at T2-T4 (28.6%) or the flashpoint anchor itself (7.1%). Combined, 80.2% of misses pattern-match to lexical / anchor specificity rather than semantic disagreement. Meme repetition without pragmatic transfer (8.8%) concentrates in T1; opposite-valence patterns (7.1%) are most prevalent in T3. On the final 23,122-criterion benchmark, 6,677 criteria (28.9%) are covered by zero of the 13 models. The two "hard" families have higher zero-coverage rates (T1 = 49.3%, T4 = 41.6%) than the easier two (T2 = 19.4%, T3 = 18.8%), consistent with flashpoint anticipation and criticism forecasting being the frontier challenges. These errors point to a more specific cognitive gap than generic sentiment modeling: models often recognize the event topic and produce plausible affect, but fail at *attention-anchor selection*—deciding which concrete detail will become the object of public repetition, moralization, joking, or complaint. This also explains why T1 and T4 are hardest: they require stance-taking around a socially charged anchor, not merely naming an emotion or listing a product attribute. Across six models, neither packing reactions nor explicitly discouraging repetition yields a statistically significant coverage gain under pairwise character matching (Appendix N).

5.2 Does direct SCF prompting help?

We test whether a specially designed prompt can improve coverage. A standard "few-shot" control is not clean because the task has no canonical ground-truth response. There are no so-called "reference comments". Instead, we develop ConsumerSCF, an expanded structured social-cognitive prompt based on consumer psychology. Before generation, it asks the model to internally reason through three social-theory lenses—self-presentation strategy, displayed capital, and framing (Goffman, 1959; Bourdieu, 1984; Kahneman & Tversky, 1979)—and operationalizes each lens with concrete participation motives, identity signals, resource cues, and narrative frames; the full prompt is shown in Appendix J. This makes ConsumerSCF stronger than a vague role prompt, since it expands social cognition into participation motives, identity and resource signals, and narrative packaging choices. We randomly select 100 topics across four generators that span the leaderboard (Gemini-3.1-Pro, Qwen3.5-397B, GPT-5.2, Claude-Sonnet-4.6). ConsumerSCF yields 3.8, 0.6, 3.8, and 3.0 percentage points higher coverage for Gemini-3.1-Pro, Qwen3.5-397B, GPT-5.2, and Claude-Sonnet-4.6, respectively, while

preserving their relative ordering (Figure 5). Differences across models remain substantial: coverage under SCF ranges from 25.3% to 45.5%.

We also evaluate a generate-reflect multi-agent pipeline based on MEGAAGENT (Wang et al., 2025b) on the same 100-topic slice. It first produces the naive comment set, then lets a CEO agent autonomously define subordinate agents, exchange files and messages, identify missing reaction angles, and refine the draft. Recorded coverage rises for MiMo-V2.5-Pro from 32.9% to **37.6%** (+4.7 pt; 54 wins, 22 ties, 24 losses) and GPT-5.2 from 30.1% to **31.8%** (+1.8 pt; 41 wins, 29 ties, 30 losses). These gains show that CONSUMERSIMBENCH can reward iterative refinement beyond a single call.

Generic chain-of-thought is difficult to isolate for current frontier APIs, where hidden reasoning is already part of the serving stack. Text classifiers, ABSA, and opinion-mining models are natural baselines for fixed-label reaction prediction (Bojic et al., 2026), but they predict labels for given text, whereas CONSUMERSIMBENCH requires open-ended comments for a new topic and then audits whether those comments recover unseen reaction criteria.

5.3 ARE POINTWISE JUDGES RELIABLE?

Our main experiments use GPT-5.2 as the fixed judge for cost-effectiveness. We validate the scoring pipeline at three levels: inter-judge agreement, intra-judge stability, and human audit of pointwise labels. The audit asks annotators to assign their own yes/no label for each generated-comment/criterion pair, allowing us to compare frozen judge labels against independent human-majority labels.

For inter-judge reliability, we employ three independent judges (GPT-5.2, Gemini-3-Pro, and Claude-4.5-Opus) on 100+ criteria from 10 random topics with the same rules. The three judges achieve 92.1% agreement. Compared with holistic LLM-as-Judge scoring, whose agreement rate collapses to 65.8%, **our yes-no criteria yield substantially higher agreement**. We also show the necessity of our five-field criteria by keeping only the trigger itself: the agreement rate drops from 92.1% to 84.8%, showing that the judgment rule carries real signal. For intra-judge stability, we run GPT-5.2 three times on the same set and obtain 91.8% agreement. Together, these checks support pointwise judging as a relatively stable method.

Finally, we connect inter-judge reliability to human validity by asking three human annotators to label the same 100+ criteria. The audit finds 98.4% agreement between pointwise judge decisions and human-majority labels. On the hardest subset—the three-judge-disagreement cases—human majority still agrees with the LLM-judge majority in 84.6% of cases. The remaining disagreements are semantic-boundary keyword cases, suggesting that errors are localized to ambiguous criterion boundaries rather than systematic judge misalignment. We also reduce length bias by judging only the first 3,600 characters of each generation and reduce style bias by asking the judge to apply criterion-specific definitions, examples, negative examples, and semantic-equivalence rules rather than rewarding rhetorical fluency. As a post-hoc check, verbosity alone does not explain the leaderboard: Gemini-3.1-Pro is the top model without being the longest generator, and MiMo-V2.5-Pro outperforms several longer generators.

6 CONCLUSION

Can LLMs think like consumers? On CONSUMERSIMBENCH, the answer is **not yet**. The best system reconstructs only 42.5% of what real public discussion surfaces, and the failure is asymmetric: models can sound emotional while failing to anticipate which trigger will dominate or which dimension users will attack. Well-designed prompts and multi-agent refinement can improve coverage, but closing the gap likely requires more than prompt-level scaffolding alone.

The broader message is a reframing, from *dyadic belief inference* in fictional vignettes to *crowd-level reaction reconstruction* in real public discourse. Models that sound consumer-like are not automatically models that anticipate consumer criticism, and for downstream marketing users, the latter is what matters. With top-tier LLMs surging on technical benchmarks, CONSUMERSIMBENCH gives measurable form to a complementary frontier of open-form humanities problems, where models must anticipate socially grounded reactions rather than solve formally specified tasks. This echoes users' complaint that today's models often lack real emotional value, not just logical competence.

AI USE STATEMENT

Large Language Models were used as research and writing assistants. They supported language polishing and code debugging. LLM assistance was also used in benchmark construction and evaluation, including criteria drafting, topic filtering, and pointwise judging.

ETHICS STATEMENT

For responsible release, we use only publicly available topic-level materials and do not crawl RedNote posts, access undocumented platform APIs, or release raw user threads or account identifiers.

The public release at `https://anonymous.4open.science/r/ConsumerSimBench-DCDE/` ships the benchmark JSONL, three judge prompts, English sidecars for reviewer inspection, and the evaluation harness. Its intended use is marketing-facing reaction forecasting: anticipating praise targets, criticism vectors, emotional flashpoints, and breakout angles before release. Raw RedNote threads are not included. A high score indicates reaction-forecasting capability, while deployment should still account for platform, product, and governance context (Appendix F).

The benchmark targets capability auditing of LLMs, not surveillance or behavioral targeting.

REPRODUCIBILITY STATEMENT

All main leaderboard evaluations use the same harness and fixed GPT-5.2 judge. Evaluation settings and full prompts are documented in Appendices I and D.

REFERENCES

Arya Agarwal. The silicon sample: Benchmarking synthetic users against human respondents in market research. *Available at SSRN 5835122*, 2025.

Gati V Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 337–371. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/aher23a.html`.

Anonymous. Social-r1: Enhancing social intelligence in LLMs through human-like reinforced reasoning. In *Submitted to The Fourteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=3qAzQyOOnA`. under review.

Jacy Reese Anthis, Ryan Liu, Sean M Richardson, Austin C Kozlowski, Bernard Koch, James Evans, Erik Brynjolfsson, and Michael Bernstein. Llm social simulations are a promising research method. *arXiv preprint arXiv:2504.02234*, 2025.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K Eckstein, Noémi Éltető, et al. A foundation model to predict and capture human cognition. *Nature*, 644(8078):1002–1009, 2025.

James Bisbee, Joshua D Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M Larson. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32 (4):401–416, 2024.

Ljubisa Bojic, Alexander Felfernig, Bojana Dinic, Velibor Ilic, Achim Rettinger, Vera Mevorah, and Damian Trilling. Llm agents predict social media reactions but do not outperform text

classifiers: Benchmarking simulation accuracy using 120k+ personas of 1511 humans. *arXiv preprint arXiv:2604.19787*, 2026.

Pierre Bourdieu. *Distinction: A Social Critique of the Judgement of Taste*. Harvard University Press, 1984.

James Brand, Ayelet Israeli, and Donald Ngwe. Using llms for market research. *Harvard Business School Marketing Unit Working Paper*, (23-062), 2023.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. Humans or LLMs as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8301–8327. Association for Computational Linguistics, 2024a.

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Gao Xing, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, and Fei Huang. Socialbench: Sociality evaluation of role-playing conversational agents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2108–2126, 2024b.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, et al. From persona to personalization: A survey on role-playing language agents. *Transactions on Machine Learning Research*.

Luyu Chen, Quanyu Dai, Zeyu Zhang, Xueyang Feng, Mingyu Zhang, Pengcheng Tang, Xu Chen, Yue Zhu, and Zhenhua Dong. Recusersim: A realistic and diverse user simulator for evaluating conversational recommender systems. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25, pp. 133–142, New York, NY, USA, 2025a. Association for Computing Machinery. ISBN 9798400713316. doi: 10.1145/3701716.3715258. URL https://doi.org/10.1145/3701716.3715258.

Ruirui Chen, Weifeng Jiang, Chengwei Qin, and Cheston Tan. Theory of mind in large language models: Assessment and enhancement. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31539–31558, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1522. URL https://aclanthology.org/2025.acl-long.1522/.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, et al. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *ICLR*, 2024c.

Yuyan Chen, Songzhou Yan, Sijia Liu, Yueze Li, and Yanghua Xiao. Emotionqueen: A benchmark for evaluating empathy of large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 2149–2176, 2024d.

Man-Lin Chu et al. Llm-based multi-agent system for simulating and analyzing marketing and consumer behavior. In *IEEE International Conference on e-Business Engineering (ICEBE)*, 2025.

Yun-Shiuan Chuang, Ruixuan Tu, Chengtao Dai, Smit Vasani, Binwei Yao, Michael Henry Tessler, Sijia Yang, Dhavan Shah, Robert Hawkins, Junjie Hu, et al. Debate: A large-scale benchmark for role-playing llm agents in multi-agent, long-form debates. *arXiv preprint arXiv:2510.25110*, 2025.

Yao Dou, Michel Galley, Baolin Peng, Chris Kedzie, Weixin Cai, Alan Ritter, Chris Quirk, Wei Xu, and Jianfeng Gao. Simulatorarena: Are user simulators reliable proxies for multi-turn evaluation of ai assistants? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 35200–35278, 2025.

Jessica Maria Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. Cognitive bias in decision-making with llms. In *Findings of the association for computational linguistics: EMNLP 2024*, pp. 12640–12653, 2024.

Denis Emelin, Ronan Le Bras, Jena D Hwang, Maxwell Forbes, and Yejin Choi. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 698–718, 2021.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. Understanding social reasoning in language models with language models. *Advances in Neural Information Processing Systems*, 36:13518–13529, 2023.

Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. S3: Social-network simulation system with large language model-empowered agents. *arXiv preprint arXiv:2307.14984*, 2023.

Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications*, 11(1):1–24, 2024.

Erving Goffman. *The Presentation of Self in Everyday Life*. Doubleday, 1959.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on LLM-as-a-Judge. *arXiv preprint arXiv:2411.15594*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning {ai} with shared human values. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=dNy_RKzJacY.

Luke Hewitt, Ashwini Ashokkumar, Isaias Ghezae, and Robb Willer. Predicting results of social science experiments using large language models. *Preprint*, 2024.

Tiancheng Hu, Joachim Baumann, Lorenzo Lupo, Nigel Collier, Dirk Hovy, and Paul Röttger. Simbench: Benchmarking the ability of large language models to simulate human behaviors. *arXiv preprint arXiv:2510.17516*, 2025.

Wenyue Hua, Lizhou Fan, Lingyao Li, Kai Mei, Jianchao Ji, Yingqiang Ge, Libby Hemphill, and Yongfeng Zhang. War and peace (waragent): Large language model-based multi-agent simulation of world wars. *arXiv preprint arXiv:2311.17227*, 2023.

EunJeong Hwang, Yuwei Yin, Giuseppe Carenini, Peter West, and Vered Shwartz. Infusing theory of mind into socially intelligent llm agents. *arXiv preprint arXiv:2509.22887*, 2025.

Jingru Jia, Zehua Yuan, Junhao Pan, Paul E McNamara, and Deming Chen. Decision-making behavior evaluation framework for llms under uncertain context. *Advances in neural information processing systems*, 37:113360–113382, 2024.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*, 2023.

Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979.

Zhaolu Kang, Junhao Gong, Jiaxu Yan, Wanke Xia, Yian Wang, Ziwen Wang, Huaxuan Ding, Zhuo Cheng, Wenhao Cao, Zhiyuan Feng, et al. Hssbench: Benchmarking humanities and social sciences ability for multimodal large language models. *arXiv preprint arXiv:2506.03922*, 2025.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. Fantom: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14397–14413, 2023.

Lawrence J. Klinkert, Steph Buongiorno, and Corey Clark. Evaluating the efficacy of llms to emulate realistic human personalities. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 20(1):65–75, Nov. 2024. doi: 10.1609/aiide.v20i1.31867. URL https://ojs.aaai.org/index.php/AIIDE/article/view/31867.

Seungbeen Lee, Seungwon Lim, Seungju Han, Giyeong Oh, Hyungjoo Chae, Jiwan Chung, Minju Kim, Beong-woo Kwak, Yeonsoo Lee, Dongha Lee, et al. Do llms have distinct and consistent personality? trait: Personality testset designed for llms with psychometrics. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 8397–8437, 2025.

Nian Li, Chen Gao, Mingyu Li, Yong Li, and Qingmin Liao. Econagent: large language model-empowered agents for simulating macroeconomic activities. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 15523–15536, 2024.

Zhicheng Lin. Large language models as psychological simulators: A methodological guide. *Advances in Methods and Practices in Psychological Science*, 9(1):25152459251410153, 2026.

Jiaheng Liu, Zehao Ni, Haoran Que, Tao Sun, Zekun Wang, Jian Yang, Jiakai Wang, Hongcheng Guo, Zhongyuan Peng, Ge Zhang, et al. Roleagent: Building, interacting, and benchmarking high-quality role-playing agents from scripts. *Advances in Neural Information Processing Systems*, 37:49403–49428, 2024a.

Jing Liu, Xinxing Ren, Yanmeng Xu, and Zekun Guo. Can ai automatically analyze public opinion? a llm agents-based agentic pipeline for timely public opinion analysis. *arXiv preprint arXiv:2505.11401*, 2025a.

Ziyi Liu, Abhishek Anand, Pei Zhou, Jen-tse Huang, and Jieyu Zhao. Interintent: Investigating social intelligence of llms via intention understanding in an interactive game context. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6718–6746, 2024b.

Zizhou Liu, Ziwei Gong, Lin Ai, Zheng Hui, Run Chen, Colin Wayne Leach, Michelle R Greene, and Julia Hirschberg. The mind in the machine: A survey of incorporating psychological theories in llms. *arXiv preprint arXiv:2505.00003*, 2025b.

Stephan Ludwig, Peter J Danaher, Xiaohao Yang, Yu-Ting Lin, Ehsan Abedin, Dhruv Grewal, and Lan Du. Extracting consumer insight from text: A large language model approach to emotion and evaluation measurement. *arXiv preprint arXiv:2602.15312*, 2026.

Benjamin F Maier, Ulf Aslak, Luca Fiaschi, Nina Rismal, Kemble Fletcher, Christian C Luhmann, Robbie Dow, Kli Pappas, and Thomas V Wiecki. Llms reproduce human purchase intent via semantic similarity elicitation of likert ratings. *arXiv preprint arXiv:2510.08338*, 2025.

Giovanni Franco Gabriel Marraffini, Andrés Cotton, Noe Fabian Hsueh, Axel Fridman, Juan Wisznia, and Luciano Del Corro. The greatest good benchmark: Measuring llms' alignment with utilitarian moral dilemmas. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 21950–21959, 2024.

Mike A Merrill, Alexander G Shaw, Nicholas Carlini, Boxuan Li, Harsh Raj, Ivan Bercovich, Lin Shi, Jeong Yeon Shin, Thomas Walshe, E Kelly Buchanan, et al. Terminal-bench: Benchmarking agents on hard, realistic tasks in command line interfaces. *arXiv preprint arXiv:2601.11868*, 2026.

Fernando Miranda and Pedro Paulo Balbi. Simulating public opinion: Comparing distributional and individual-level predictions from llms and random forests. *Entropy*, 27(9):923, 2025.

Xinyi Mou, Xuanwen Ding, Qi He, Liang Wang, Jingcong Liang, Xinnong Zhang, Libo Sun, Jiayu Lin, Jie Zhou, Xuanjing Huang, et al. From individual to society: A survey on social simulation driven by large language model-based agents. *arXiv preprint arXiv:2412.03563*, 2024.

Xinyi Mou, Jingcong Liang, Jiayu Lin, Xinnong Zhang, Xiawei Liu, Shiyue Yang, Rong Ye, Lei Chen, Haoyu Kuang, Xuanjing Huang, and Zhongyu Wei. AgentSense: Benchmarking social intelligence of language agents through interactive scenarios. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4975–5001, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.257. URL https://aclanthology.org/2025.naacl-long.257/.

Keane Ong, Wei Dai, Carol Li, Dewei Feng, Hengzhi Li, Jingyao Wu, Jiaee Cheong, Rui Mao, Gianmarco Mengaldo, Erik Cambria, et al. Human behavior atlas: Benchmarking unified psychological and social behavior understanding. *arXiv preprint arXiv:2510.04899*, 2025.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23. Association for Computing Machinery, 2023.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, et al. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. *arXiv preprint arXiv:2502.08691*, 2025.

Elinor Poole-Dayan, Jiayi Wu, Taylor Sorensen, Jiaxin Pei, and Michiel A Bakker. Benchmarking overton pluralism in llms. *arXiv preprint arXiv:2512.01351*, 2025.

Weihong Qi, Fan Huang, Jisun An, and Haewoon Kwak. A cross-cultural comparison of llm-based public opinion simulation: Evaluating chinese and us models on diverse societies. *arXiv preprint arXiv:2506.21587*, 2025.

Yao Qu and Jue Wang. Performance and biases of large language models in public opinion simulation. *Humanities and Social Sciences Communications*, 11(1):1–13, 2024.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. Emobench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5986–6004, 2024.

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik R Narasimhan, and Vishvak Murahari. PersonaGym: Evaluating persona agents and LLMs. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 6999–7022, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.368. URL https://aclanthology.org/2025.findings-emnlp.368/.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social iqa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4463–4473, 2019.

Nils Schwager, Simon Münker, Alistair Plum, and Achim Rettinger. Towards simulating social media users with llms: Evaluating the operational validity of conditioned comment prediction. In *The Proceedings for the 15th Workshop on Computational Approaches to Subjectivity, Sentiment Social Media Analysis (WASSA 2026)*, pp. 208–221, 2026.

Preethi Seshadri, Samuel Cahyawijaya, Ayomide Odumakinde, Sameer Singh, and Seraphina Goldfarb-Tarrant. Lost in simulation: Llm-simulated users are unreliable proxies for human users in agentic evaluations. *arXiv preprint arXiv:2601.17087*, 2026.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-llm: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158*, 2023.

Haojun Shi, Suyu Ye, Xinyu Fang, Chuanyang Jin, Leyla Isik, Yen-Ling Kuo, and Tianmin Shu. Muma-tom: Multi-modal multi-agent theory of mind. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1510–1519, 2025.

Kazutoshi Shinoda, Nobukatsu Hojo, Kyosuke Nishida, Saki Mizuno, Keita Suzuki, Ryo Masumura, Hiroaki Sugiyama, and Kuniko Saito. Tomato: Verbalizing the mental states of role-playing llms for benchmarking theory of mind. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1520–1528, 2025.

Oliver Slumbers, Joel Z Leibo, and Marco A Janssen. Using large language models to simulate human behavioural experiments: Port of mars. *arXiv preprint arXiv:2506.05555*, 2025.

James WA Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, et al. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285–1295, 2024.

Xuanbo Su, Wenhao Hu, Le Zhan, Yanqi Yang, and Leo Huang. Sell more, play less: Benchmarking llm realistic selling skill. *arXiv preprint arXiv:2604.07054*, 2026.

Jiakai Tang, Heyang Gao, Xuchen Pan, Lei Wang, Haoran Tan, Dawei Gao, Yushuo Chen, Xu Chen, Yankai Lin, Yaliang Li, et al. Gensim: A general social simulation platform with large language model based agents. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pp. 143–150, 2025.

Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. Charactereval: A chinese benchmark for role-playing conversational agent evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11836–11850, 2024.

Lei Wang, Jingsen Zhang, Hao Yang, Zhi-Yuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Hao Sun, Ruihua Song, et al. User behavior simulation with large language model-based agents. *ACM Transactions on Information Systems*, 43(2):1–37, 2025a.

Mengxin Wang, Dennis J Zhang, and Heng Zhang. Large language models for market research: A data-augmentation approach. *Marketing Science*, 2026.

Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14743–14777, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.878. URL https://aclanthology.org/2024.findings-acl.878/.

Qian Wang, Tianyu Wang, Zhenheng Tang, Qinbin Li, Nuo Chen, Jingsheng Liang, and Bingsheng He. Megaagent: A large-scale autonomous llm-based multi-agent system without predefined sops. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 4998–5036, 2025b.

Qian Wang, Jiaying Wu, Zichen Jiang, Zhenheng Tang, Bingqiao Luo, Nuo Chen, Wei Chen, and Bingsheng He. Llm-based human simulations have not yet been reliable. *arXiv preprint arXiv:2501.08579*, 2025c.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, et al. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 1840–1873, 2024b.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37: 95266–95290, 2024c.

Ziyi Wang, Yuxuan Lu, Yimeng Zhang, Jing Huang, and Dakuo Wang. Customer-r1: personalized simulation of human behaviors via rl-based llm agent in online shopping. In *First Workshop on Multi-Turn Interactions in Large Language Models*.

Ziyi Wang, Yuxuan Lu, Yimeng Zhang, Jing Huang, and Dakuo Wang. Customer-r1: Personalized simulation of human behaviors via rl-based llm agent in online shopping. *arXiv preprint arXiv:2510.07230*, 2025d.

Bo Wu, Peiye Liu, Wen-Huang Cheng, Bei Liu, Zhaoyang Zeng, Jia Wang, Qiushi Huang, and Jiebo Luo. Smp challenge: An overview and analysis of social media prediction challenge. In *Proceedings of the 31st ACM International Conference on Multimedia*, 2023a.

Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. Hi-tom: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10691–10706, 2023b.

Feng Xiao and XT XiaoTian Wang. Evaluating the ability of large language models to predict human social decisions. *Scientific Reports*, 15(1):32290, 2025.

Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. Opentom: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8593–8623, 2024.

Diyi Yang, Caleb Ziems, William Held, Omar Shaikh, Michael S Bernstein, and John Mitchell. Social skill training with large language models. *arXiv preprint arXiv:2404.04204*, 2024a.

Yuzhe Yang, Yifei Zhang, Minghao Wu, Kaidi Zhang, Yunmiao Zhang, Honghai Yu, Yan Hu, and Benyou Wang. Twinmarket: A scalable behavioral and social simulation for financial markets. *arXiv preprint arXiv:2502.01506*, 2025.

Ziyi Yang, Zaibin Zhang, Zirui Zheng, Yuxian Jiang, Ziyue Gan, Zhiyu Wang, Zijian Ling, Jinsong Chen, Martz Ma, Bowen Dong, et al. Oasis: Open agent social interaction simulations with one million agents. *arXiv preprint arXiv:2411.11581*, 2024b.

Xixian Yong, Jianxun Lian, Xiaoyuan Yi, Xiao Zhou, and Xing Xie. Motivebench: How far are we from human-like motivational reasoning in large language models? *arXiv preprint arXiv:2506.13065*, 2025.

Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yuechen Jiang, Yupeng Cao, Zhi Chen, Jordan Suchow, Zhenyu Cui, Rong Liu, et al. Fincon: A synthesized llm multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. *Advances in Neural Information Processing Systems*, 37:137010–137045, 2024.

Ivan Zakazov, Mikolaj Boronski, Lorenzo Drudi, and Robert West. Assessing social alignment: Do personality-prompted large language models behave like humans? *arXiv preprint arXiv:2412.16772*, 2024.

Min Zeng. Psychcounsel-bench: Evaluating the psychology intelligence of large language models. *arXiv preprint arXiv:2510.01611*, 2025.

Xinnong Zhang, Jiayu Lin, Xinyi Mou, Shiyue Yang, Xiawei Liu, Libo Sun, Hanjia Lyu, Yihang Yang, Weihong Qi, Yue Chen, et al. Socioverse: A world model for social simulation powered by llm agents and a pool of 10 million real-world users. *arXiv preprint arXiv:2504.10157*, 2025a.

Yiqun Zhang, Xiaocui Yang, Xingle Xu, Zeran Gao, Yijie Huang, Shiyi Mu, Shi Feng, Daling Wang, Yifei Zhang, Kaisong Song, et al. Affective computing in the era of large language models: A survey from the nlp perspective. *arXiv preprint arXiv:2408.04638*, 2024.

Zeyu Zhang, Jianxun Lian, Chen Ma, Yaning Qu, Ye Luo, Lei Wang, Rui Li, Xu Chen, Yankai Lin, Le Wu, et al. Trendsim: Simulating trending topics in social media under poisoning attacks with llm-based multi-agent system. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 2930–2949, 2025b.

Jinfeng Zhou, Zheyu Chen, Shuai Wang, Quanyu Dai, Zhenhua Dong, Hongning Wang, and Minlie Huang. Think socially via cognitive reasoning. *arXiv preprint arXiv:2509.22546*, 2025.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*, 2023.