LongSocialBench: Do Long-Context LLMs Understand Online Discussion Threads?

Anonymous ACL submission

Abstract

Long-context LLMs can now ingest entire online discussion threads, but understanding the nuances of the social discourse they express requires more than reading a long document: models must jointly track parent–reply relations, chain-level turning points, scoped topic subtrees, cross-branch contrasts, and participant topic-specific attitude trajectories over time. To test this structure-aware social reasoning, we introduce **LongSocialBench**, a benchmark of $1{,}462$ verified human-authored multiple-choice items over 94 complete Hacker News, Stack Exchange, and Reddit r/ChangeMyView episodes (median $\approx$ 72K tokens per episode). Each item pairs a complete serialized discussion and reply structure with a four-option question, requiring models to recover structured social evidence before answering. Released items are verified for answerability, option uniqueness, and evidence grounding. Across 18 models and 29 evaluation settings, deployable workflows (Baseline, CoT variants, and retrieval variants) achieve low exact-match accuracy; even the best single setting, Claude-Opus-4.7 with self-reflection, reaches only 63.0%, compared with 72.4% for independent human readers. An oracle-scope diagnostic that gives models the relevant thread region raises average accuracy only to 54.3%, indicating that failures persist even after evidence localization. LongSocialBench shows that the missing capability is not context access or prompting alone, but social understanding over structured reply trees.

1 Introduction

Online discussion threads are a central setting for public social interaction: participants debate claims, challenge one another, shift positions, form sub-discussions, and reveal roles through reply structure [Abbott et al., 2016, Tan et al., 2016, Fabbri et al., 2021, Niu et al., 2024]. Understanding the social dynamics of discussion episodes is therefore not the same as summarizing a long document. These dynamics are expressed through reply links, participant identities, branch boundaries, and temporal turning points, so a model must decide which part of the thread matters, interpret how comments relate, and combine evidence spread across the episode to properly interpret the social discourse. Long-context LLMs now make it possible to place a complete Hacker News thread, Stack Exchange page, or Reddit debate in context [Gemini Team, 2024, Bai et al., 2024, 2025, Hsieh et al., 2024, Yen et al., 2025], but it remains unclear whether current systems can understand the discussion as a structured social discourse episode. This paper asks: *When the complete discussion is already in context, can current LLM workflows recover the relevant evidence and determine the supported social interpretation?*

The challenge is not retrieving a salient span. In a r/ChangeMyView discussion about middle-class alienation, one LongSocialBench item asks where the episode's unique longest reply chain first moves from a class-and-privilege debate into a corporate wage-distribution argument. The answer is not determined by repeated words such as *class*, *privilege*, or *wages*; it depends on isolating the specified chain and tracking when the exchange shifts from housing and wealth-hoarding claims to concrete claims about company revenue, GDP per worker, and worker pay. Solving the item requires tracing the specified reply chain, integrating the claims made across turns, and judging when the exchange shifts at a higher discourse level, rather than matching repeated keywords.

Existing benchmarks each cover part of this setting but not the combination. Long-context benchmarks evaluate large inputs organized as documents, passages, code, tables, or synthetic structures rather than branching social interaction [Shaham et al., 2023, Bai et al., 2024, Li et al., 2024, Bai et al., 2025, Hsieh et al., 2024, Yen et al., 2025].

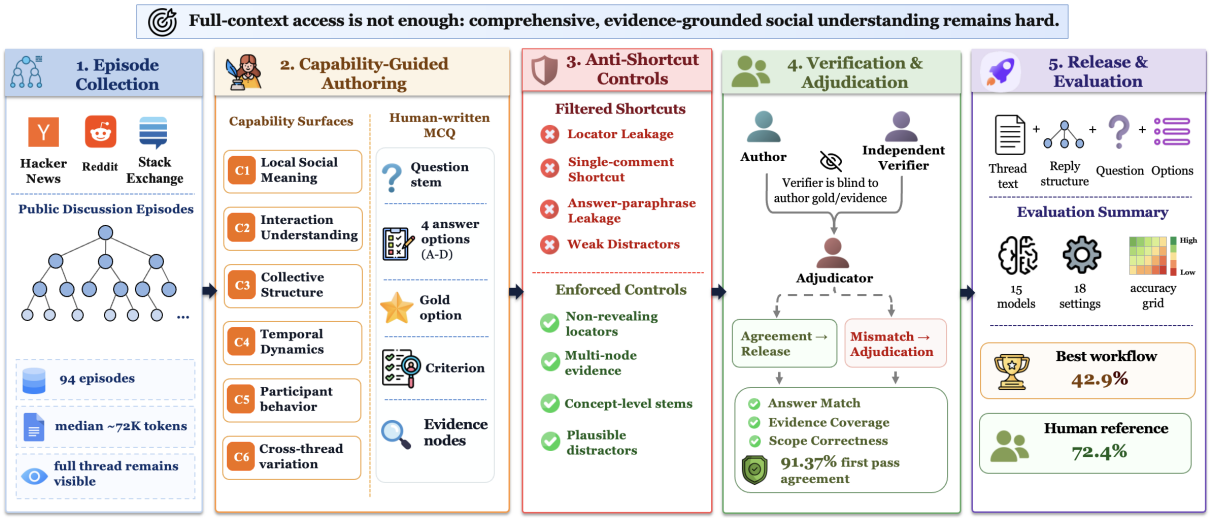


Figure 1: **LongSocialBench construction and verification pipeline (§3.1).** We sample long public discussions, author capability-guided items, apply anti-shortcut controls, independently verify answers and evidence, and adjudicate residual disputes.

Social-reasoning and theory-of-mind benchmarks test pragmatic or interpersonal inference, but in short or already-localized contexts where the relevant interaction is given [Sap et al., 2019, Chen et al., 2024]. Long-conversation benchmarks stress memory over user–assistant histories [Maharana et al., 2024], while online-discussion resources provide stance labels, persuasion outcomes, or summaries rather than objective items over complete episodes [Abbott et al., 2016, Tan et al., 2016, Fabbri et al., 2021, Niu et al., 2024]. The missing problem is evidence-grounded social reasoning over a full reply tree: a model must first recover the relevant branch, chain, participant trajectory, or comparison set, then decide which social interpretation is supported.

We introduce **LongSocialBench**, 1,462 human-authored multiple-choice items over 94 real Hacker News, Stack Exchange, and Reddit r/ChangeMyView episodes (median ≈72K tokens). Episodes are complete public discussions selected to avoid truncation confounds while covering complementary forum formats and topic ranges: technical and policy threads from Hacker News, social and policy debates from Reddit r/ChangeMyView, and expert-domain Q&A from Stack Exchange. Released items are verified with explicit human-written criteria and node-level evidence.

LongSocialBench operationalizes a premise supported by work on social inference, stance, persuasion, conversational stance, and discussion summarization: social meaning in threaded discussions appears not only in isolated comments, but also in reply relations, debate positions, participant behavior, and evolution dynamics of sub-discussions [Sap et al., 2019, Chen et al., 2024, Abbott et al., 2016, Tan et al., 2016, Niu et al., 2024, Fabbri et al., 2021]. We organize these recurring evidence patterns into six capability surfaces: **C1** local social meaning, **C2** reply-pair interaction, **C3** collective structure, **C4** temporal dynamics, **C5** participant profile, and **C6** cross-subdiscussion variation.

Empirically, LongSocialBench exposes a large gap between full-context access and reliable social understanding. To separate deployable performance from diagnostic explanations, we evaluate 29 settings spanning full-context baselines, reasoning prompts, retrieval workflows, validity probes, and scope diagnostics. Across 18 models, the deployable workflows (Baseline, CoT variants, and retrieval variants) cluster in a narrow 38.5–46.5% average-accuracy range, far below the 72.4% mean of five independent human readers under the same input condition. The best individual deployable setting, Claude-Opus-4.7 with self-reflection, reaches 63.0%, still 9.4 pp below the human mean. An oracle-scope diagnostic that gives models the relevant thread region still yields only 54.3% average accuracy, suggesting that the missing capability involves social understanding over structured reply trees, not context access alone.

Our contributions are:

- **A reliable benchmark for long-episode social reasoning.** LongSocialBench turns complete real

discussion episodes into carefully verified QA items, organized by six capability surfaces that test structured social evidence from local replies to participant trajectories and cross-branch dynamics.

- **A comprehensive stress test of long-context workflows.** We evaluate 18 models across 29 evaluation settings, separating deployable workflows (Baseline, CoT variants, and retrieval variants) from diagnostic settings (validity probes and scope diagnostics). The failure is not limited to one prompt or method; it persists across the deployable workflows, showing that current systems remain unreliable on complete discussion episodes.
- **A diagnosis of why models fail.** We separate incorrect evidence-scope selection from reasoning failures conditioned on the correct scope. In a scope diagnostic, even after models are given the relevant thread region, they average only 54.3%. This shows that the bottleneck is not context length alone, but also the ability to apply social-reading criteria over chains, subtrees, participant trajectories, and branch comparisons.

Together, LongSocialBench provides a challenging, reliable, and diagnostic testbed for future research on long-context models that must understand real online discussions as social interaction rather than flat text.

2 Related Work

LongSocialBench targets a missing combination in prior evaluation: verifiable social reasoning over complete real reply-thread episodes. Prior benchmarks cover adjacent settings such as long-context QA, dialogue memory, localized social inference, argument or stance labels, and discussion summarization, but they rarely require models to read a complete discussion, recover the relevant thread scope, and make an objectively verifiable social judgment from that evidence. Table 1 summarizes this area-level gap; representative cited benchmarks are listed in Appendix A.1, Table 4.

2.1 Long-context evaluation and memory.

Long-context benchmarks evaluate QA, aggregation, synthetic context use, code, structured data, and multi-hop reasoning over long inputs [Shaham et al., 2023, Bai et al., 2024, Li et al., 2024, Bai et al., 2025, Hsieh et al., 2024, Yen et al., 2025, Bertsch et al., 2025]. Related work also studies long-term dialogue memory over user–assistant histories [Maharana et al., 2024]. These benchmarks establish that long-context use is difficult, but their typical inputs are documents, passages, code, tables, synthetic structures, or dialogue histories rather than complete public reply-thread episodes. LongSocialBench instead tests whether models can judge how replies, participants, positions, and sub-discussions develop across a full discussion thread.

Table 1: **Compact area-level comparison.** The table summarizes typical capabilities of evaluation areas; cited benchmark-level examples are provided in Table 4.

Area	Long	Thread	Social	Verif.
Long-context eval./agg.	✓	▲	▲	✓
Long-memory dialogue	✓	▲	▲	▲
Social commonsense/ToM	✗	✗	✓	✓
Argument / stance	▲	✓	✓	▲
Discussion summarization	▲	▲	✓	✗
Ours	✓	✓	✓	✓

✓ = central; ▲ = partial; ✗ = generally absent. Long = long-context input; Thread = reply-tree structure; Social = social interpretation; Verif. = objective QA or gold-answer verification.

2.2 Social-discourse and threaded-discussion evaluation.

SocialIQA [Sap et al., 2019], MuTual [Cui et al., 2020], ToMi [Le et al., 2019], and ToMBench [Chen et al., 2024] test social inference, but the model is given a short, localized interaction. Threaded-discussion resources such as IAC 2.0 [Abbott et al., 2016], ChangeMyView persuasion studies [Tan et al., 2016], MT-CSD [Niu et al., 2024], and ConvoSumm [Fabbri et al., 2021] study real online discussions through stance labels, persuasion outcomes, conversational-stance labels, or summaries. LongSocialBench instead evaluates complete-episode social reasoning: models must read a full long discussion and make verified judgments from evidence distributed across the reply-structured episode.

3 Task and Construction

3.1 Task Overview

A LongSocialBench item is a verifiable multiple-choice question about one complete real discussion episode. We use *episode* for the full public discussion instance, such as one Hacker News thread, one Reddit r/ChangeMyView discussion, or one Stack Exchange Q&A page. For each item, the model receives the episode as a serialized thread with reply links, together with a question stem and four

answer options.

LongSocialBench items are designed to make whole-episode social comprehension measurable. Throughout, we use *social* in an interactional and observable sense, focusing on what participants *do* to one another through replies, stances, trajectories, and sub-discussions rather than on private mental-state attribution, so that items remain grounded in thread behavior rather than reader speculation about hidden beliefs. Prior work on social inference, stance, persuasion, conversational stance, and discussion summarization has studied pieces of this problem, such as what an utterance implies, how replies relate, how positions form, and how discussions develop [Sap et al., 2019, Chen et al., 2024, Abbott et al., 2016, Tan et al., 2016, Niu et al., 2024, Fabbri et al., 2021]. We turn these recurring evidence patterns into six capability surfaces, summarized in Table 2. Together, these surfaces make the benchmark coverage systematic rather than ad hoc, linking each item to a specific evidence pattern and social-reading decision. Appendix C.1, Table 7 provides full decision descriptions and example items.

Table 2: **Compact map of the six capability surfaces.** Each surface corresponds to a recurring evidence pattern in reply-structured discussions; full descriptions and examples are provided in Appendix C.1, Table 7.

Capability surface	Core evidence pattern
C1: Local social meaning	One comment's role in its nearby context.
C2: Reply-pair interaction	How a reply relates to the parent comment it answers.
C3: Collective structure	Positions or roles organizing a scoped discussion region.
C4: Temporal dynamics	Turning points in an ordered exchange or reply chain.
C5: Participant profile	One author's stance or strategy across multiple comments.
C6: Cross-subdiscussion variation	Similarity or change across separate branches or sub-discussions.

3.2 Data Collection

Episode sampling. We collect 94 complete public discussion episodes from Hacker News, Reddit r/ChangeMyView, and Stack Exchange. The sources cover different forum focuses: Hacker News contributes many computing, AI, economy, and policy discussions; Reddit r/ChangeMyView contributes social, policy, labor, health, education, and climate debates; and Stack Exchange contributes expert-domain Q&A across computing, science, engineering, health, law, economics, and the humanities (Appendix B, Table 5). The released set is genuinely long-context, containing ≈6.81M serialized tokens, with a median episode length of ≈72K tokens and a maximum length of 116K tokens. Episodes are kept within evaluated context windows to avoid truncation confounds. Per-source length and structural statistics are reported in Appendix B, Table 6.

Item design and safeguards. Twenty invited contributors authored items under written guidelines for systematic capability coverage and shortcut-resistant question design. Appendix I.1 summarizes the authoring protocol, contributor training, and role separation. Each author read the full episode, aligned the item with one of the six capability surfaces, and wrote the question stem, four options, gold option, plain-language decision criterion, and supporting evidence nodes.

A candidate item is released only if the gold option is recoverable from the episode alone, uniquely supported under the author-written criterion, and sufficiently grounded in cited evidence, including key structural premises such as scope, reply relation, counting unit, participant trajectory, or comparison set. Appendix C.3, Table 8, details the release criteria.

The guidelines also control wording, option balance, distractor logic, and contamination risk. Stems avoid answer-revealing locators; options use parallel wording; and distractors are written as plausible wrong readings, such as wrong scope, wrong criterion, boundary errors, lexical matches, or aggregation mistakes. Following contamination concerns in long-context benchmark construction such as LongBench v2 [Bai et al., 2025], authors check distinctive stem/option phrases and named events against open-web sources; items recoverable outside the supplied episode are rewritten or excluded. Appendix C.4, Table 9, details the shortcut and contamination controls.

3.3 Verification and Release

After authoring, every released item passes independent verification or adjudication. Appendix H shows the portal workflow used to enforce role separation and record authoring, verification, and adjudication status.

Phase 1 is authoring: a contributor writes the

item, criterion, gold option, and evidence nodes. **Phase 2** is independent cross-verification: a different trained contributor sees only the model-facing input, answers from scratch, and checks whether the authored gold option is supported. **Phase 3** adjudicates verifier disagreements and author-flagged edge cases: a third reviewer re-checks the criterion, evidence, and option uniqueness, then confirms the item, requires revision and a new Phase-2 verification pass, or removes it.

Among 1,553 parseable Phase-2 records, 1,419 match the author gold option on the first pass, yielding 91.37% candidate-level agreement before adjudication. Of the 134 mismatches, 43 are retained after adjudication and 91 unresolved cases are removed. Another 24 author-flagged borderlines pass Phase 3. The released benchmark contains 1,462 items. A 100-item stratified release audit finds 98/100 pass, 2/100 wording-level corrections, and no residual gold-option errors (Appendix D.2, Table 11). Public-data licensing, contributor protocol, and release artifacts are documented in Appendix I and Appendix J.

4 Experimental Setup

Models. We evaluate 18 contemporary LLMs spanning closed-source frontier systems, strong open-weight models, and within-family progressions. Closed-source models include Claude Opus 4.7, Claude Sonnet 4.5/4/3.7, GPT-5, and Gemini 2.5 Pro; open-weight models include Mistral-Large-3, Llama-4 Scout/Maverick, Llama-3.1/3.2, Qwen3-Next/32B, GLM-5, DeepSeek-R1, Gemma-3, and OpenAI's gpt-oss models. This pool tests whether complete-thread social reasoning is solved by frontier scale, open-weight progress, or reasoning-oriented design. Appendix E.1 reports the full model list and context windows; Appendix E.2 reports workflow construction and prompt instructions.

Workflows and metrics. We evaluate 29 settings: 22 deployable long-context workflows and 7 diagnostic settings. The deployable workflows include full-context prompting; five reasoning prompts (CoT, evidence enumeration, temporal tracking, option elimination, and self-reflection); and retrieval-based methods, including sparse BM25/E5 retrieval, RAG with anchor notes, subtree retrieval, and LLM-as-retriever. The diagnostic settings include validity probes (options-only, question-only, title/root-only, shuffled-thread, and open-answer) and scope diagnostics (gold-scope and infer-scope). These diagnostics test shortcut artifacts, multiple-choice format effects, and the role of evidence localization rather than serving as standard deployment methods. Responses are scored by exact match on the parsed A–D answer over the 1,462 released items, with additional C1–C6 capability breakdowns reported as macro means. Appendix E.2, Table 13 gives the exact list of settings, input construction, and prompt instructions.

5 Results

5.1 Main result: current workflows remain far from reliable

The main result is simple: current LLMs remain unreliable on long-episode social reasoning. Across 18 models, the main tested settings, including Baselines, CoT variants, and retrieval variants, stay in a low 38.5% to 46.5% average-accuracy band (Appendix F.1, Table 15, Fig. 6). Even the best individual setting, Claude-Opus-4.7 with self-reflection, reaches only 63.0%. By contrast, five independent human readers average 72.4%, and the weakest reader still reaches 70.0% (Appendix F.2, Table 16).

This gap appears before any fine-grained diagnosis: standard Baselines, CoT variants, and retrieval variants all remain far from reliable, and even the strongest model-workflow pair remains below the human reference. The rest of this section explains the gap by analyzing capability profiles, CoT effects, retrieval behavior, validity probes, and the two-stage reading process of finding the right evidence scope and applying the social criterion.

5.2 Local meaning is easier than thread structure

Table 3 shows a clear capability split. The average full-context baseline is highest on **C1 local social meaning**, reaching 59.1%. This is the most local capability: the model mainly judges the role of one comment given its nearby context.

Performance drops when the question requires a judgment over multiple comments. For **C2** to **C6**, the average baseline ranges only from 39.2% to 46.3%. The weakest capabilities are **C3 collective structure** and **C6 cross-subdiscussion variation**, where the model must summarize a discussion region or compare behavior across branches. Models therefore interpret individual comments more reliably than they build higher-level social readings

Table 3: **Per-model capability accuracy under Baseline and CoT.** Cells report accuracy; **Avg** is the C1–C6 macro mean. CoT superscripts show lift over Baseline, with green/red for positive/negative change. **Bold** marks the best model per column; C1–C6 are defined in Table 7.

	Baseline (Full-context)							CoT (with Δ vs Baseline)						
Model	C1	C2	C3	C4	C5	C6	Avg	C1	C2	C3	C4	C5	C6	Avg
(Closed-source frontier)														
Claude-Opus-4.7	70.9	**61.8**	**50.4**	**58.4**	**51.4**	41.5	**55.7**	73.9 $^{+3.0}$	66.3 $^{+4.5}$	**58.2** $^{+7.9}$	**66.5** $^{+8.1}$	56.4 $^{+5.0}$	51.7 $^{+10.2}$	**62.2** $^{+6.5}$
Claude-Sonnet-4.5	61.2	47.2	45.3	56.2	**51.4**	**53.4**	52.4	70.9 $^{+9.7}$	56.2 $^{+9.0}$	53.7 $^{+8.4}$	63.2 $^{+7.0}$	**59.6** $^{+8.3}$	**61.0** $^{+7.6}$	60.8 $^{+8.3}$
GPT-5	**80.0**	58.4	44.8	51.4	50.0	40.7	54.2	**81.8** $^{+1.8}$	65.2 $^{+6.7}$	52.1 $^{+7.3}$	59.5 $^{+8.1}$	56.4 $^{+6.4}$	52.5 $^{+11.9}$	61.3 $^{+7.0}$
Gemini-2.5-Pro	70.9	50.6	47.9	43.2	45.9	39.0	49.6	74.5 $^{+3.6}$	58.4 $^{+7.9}$	56.2 $^{+8.3}$	51.4 $^{+8.1}$	53.2 $^{+7.3}$	46.6 $^{+7.6}$	56.7 $^{+7.1}$
Claude-4	72.7	52.8	42.8	51.9	45.0	39.8	50.8	73.3 $^{+0.6}$	60.7 $^{+7.9}$	52.7 $^{+9.9}$	59.5 $^{+7.6}$	53.2 $^{+8.3}$	49.2 $^{+9.3}$	58.1 $^{+7.3}$
Claude-3.7-Sonnet	65.5	59.6	43.8	48.1	45.0	38.1	50.0	69.7 $^{+4.2}$	**68.5** $^{+9.0}$	50.8 $^{+7.0}$	57.8 $^{+9.7}$	52.3 $^{+7.3}$	44.9 $^{+6.8}$	57.3 $^{+7.3}$
(Strong open-weight)														
Mistral-Large-3	69.7	60.7	48.0	35.1	46.8	40.7	50.2	72.7 $^{+3.0}$	66.3 $^{+5.6}$	51.7 $^{+3.6}$	39.5 $^{+4.3}$	48.6 $^{+1.8}$	43.2 $^{+2.5}$	53.7 $^{+3.5}$
Llama-4-Scout-17B	58.2	47.2	49.8	31.4	46.8	46.6	46.7	61.8 $^{+3.6}$	51.7 $^{+4.5}$	54.0 $^{+4.2}$	35.1 $^{+3.8}$	50.5 $^{+3.7}$	49.2 $^{+2.5}$	50.4 $^{+3.7}$
Qwen3-Next-80B	53.9	47.2	45.1	52.4	35.8	39.8	45.7	62.4 $^{+8.5}$	51.7 $^{+4.5}$	49.6 $^{+4.5}$	56.8 $^{+4.3}$	39.9 $^{+4.1}$	47.5 $^{+7.6}$	51.3 $^{+5.6}$
GLM-5	61.8	39.3	47.9	40.5	39.9	37.3	44.5	65.5 $^{+3.6}$	43.8 $^{+4.5}$	51.7 $^{+3.8}$	46.5 $^{+5.9}$	44.0 $^{+4.1}$	42.4 $^{+5.1}$	49.0 $^{+4.5}$
(Mid open-weight)														
DeepSeek-R1	57.6	53.9	33.5	38.9	39.9	47.5	45.2	58.2 $^{+0.6}$	52.8 $^{-1.1}$	33.0 $^{-0.4}$	38.9 $^{+0.0}$	41.3 $^{+1.4}$	46.6 $^{-0.8}$	45.1 $^{-0.1}$
Llama-3.1-70B	46.7	43.8	38.1	41.6	34.9	43.2	41.4	46.7 $^{+0.0}$	43.8 $^{+0.0}$	38.1 $^{+0.0}$	40.0 $^{-1.6}$	35.3 $^{+0.5}$	44.1 $^{+0.8}$	41.3 $^{-0.1}$
Qwen3-32B	49.7	33.7	35.8	38.4	35.8	40.7	39.0	48.5 $^{-1.2}$	33.7 $^{+0.0}$	36.7 $^{+0.9}$	38.9 $^{+0.5}$	36.7 $^{+0.9}$	40.7 $^{+0.0}$	39.2 $^{+0.2}$
Llama-4-Mav-17B	56.4	24.7	37.0	31.4	25.7	27.1	33.7	56.4 $^{+0.0}$	24.7 $^{+0.0}$	36.8 $^{-0.1}$	31.9 $^{+0.5}$	25.7 $^{+0.0}$	27.1 $^{+0.0}$	33.8 $^{+0.1}$
(Small / mid-weak)														
Gemma-3-27B	50.9	43.8	30.1	35.1	38.5	34.7	38.9	49.7 $^{-1.2}$	42.7 $^{-1.1}$	29.4 $^{-0.7}$	34.6 $^{-0.5}$	37.6 $^{-0.9}$	33.1 $^{-1.7}$	37.8 $^{-1.0}$
GPT-OSS-120B	52.7	48.3	28.2	29.2	38.1	34.7	38.5	50.3 $^{-2.4}$	47.2 $^{-1.1}$	27.7 $^{-0.6}$	29.2 $^{+0.0}$	37.2 $^{-0.9}$	33.1 $^{-1.7}$	37.4 $^{-1.1}$
Llama-3.2-90B	47.9	38.2	32.8	31.9	28.4	31.4	35.1	45.5 $^{-2.4}$	32.6 $^{-5.6}$	31.1 $^{-1.6}$	31.4 $^{-0.5}$	28.4 $^{+0.0}$	27.1 $^{-4.2}$	32.7 $^{-2.4}$
GPT-OSS-20B	37.6	22.5	29.3	25.9	32.1	29.7	29.5	37.0 $^{-0.6}$	22.5 $^{+0.0}$	29.0 $^{-0.3}$	25.9 $^{+0.0}$	30.7 $^{-1.4}$	27.1 $^{-2.5}$	28.7 $^{-0.8}$
Mean (18)	59.1	46.3	40.6	41.2	40.6	39.2	44.5	61.0 $^{+1.9}$	49.4 $^{+3.1}$	44.0 $^{+3.4}$	44.8 $^{+3.6}$	43.7 $^{+3.1}$	42.6 $^{+3.4}$	47.6 $^{+3.1}$

from distributed thread evidence.

Model strength changes the overall accuracy, but the same local-to-structural drop appears across the model pool. Stronger frontier models are more stable across capabilities, while mid-tier and smaller models show larger capability-specific swings. Appendix F.3, Fig. 7, visualizes this pattern as a per-model heatmap, showing which capabilities each model handles consistently and where performance collapses; Figs. 8 and 9 provide the corresponding platform-level views.

5.3 CoT helps strong models, but does not solve the task

Reasoning prompts help unevenly across model tiers. Closed-frontier models gain between 6.5 and 8.3 percentage points when accuracy is averaged across the six capabilities. Their largest gains appear on harder capabilities, especially C3, C4, and C6, suggesting that CoT helps more on multi-comment reasoning than on already-local judgments. Strong open-weight models gain less, roughly 3.5 to 5.6 percentage points. Mid-tier and smaller models gain little or move backward: DeepSeek-R1, Llama-3.1-70B, GPT-OSS-120B/20B, Gemma-3-27B, and Llama-3.2-90B are flat or negative on the six-capability average (Appendix F.4, Table 17, Fig. 10).

Reasoning prompts therefore provide a limited and tier-dependent benefit. They improve stronger models and help more on the harder multi-comment capabilities, but the gains shrink or disappear for weaker models. CoT is a useful intervention, not a reliable solution: it raises the best rows, but the overall model pool remains far below the human reference.

5.4 Retrieval helps shallow evidence, not deeper threads

Retrieval helps only in limited cases. Figure 2 shows that raw BM25 and E5 retrieval often hurt performance on harder capabilities. Structured retrieval and LLM-as-retriever perform better than flat top-k retrieval, but their gains remain limited. This suggests that selecting isolated relevant-looking comments is often insufficient for questions that require the surrounding reply context.

Figure 3 shows where retrieval helps. All re-

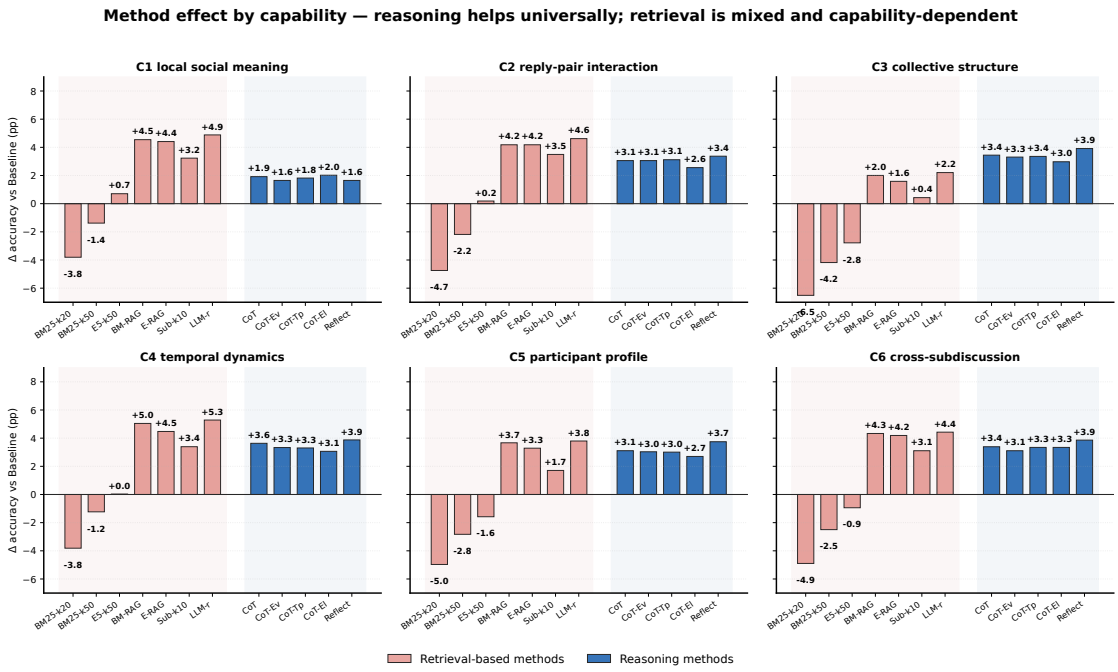


Figure 2: **Method effect by capability.** Mean Δ accuracy vs. Baseline across 18 models, per capability C1–C6. Reasoning helps universally; raw retrieval is weakest on hard capabilities; structured retrieval recovers part of the gap.

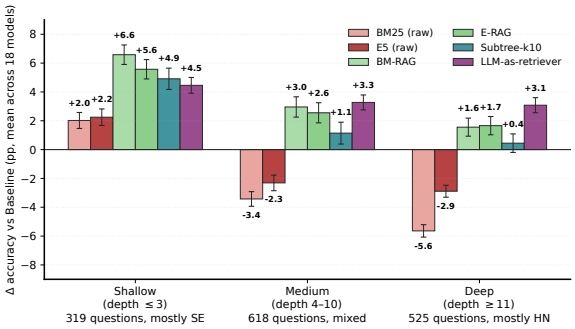


Figure 3: **Retrieval by thread depth.** Mean Δ accuracy vs. Baseline across 18 models. Raw retrieval helps shallow threads but hurts deep ones; structured retrieval partly stabilizes it.

trieval methods improve Shallow questions, where the relevant evidence is close to the root. The benefit shrinks or reverses for Medium-depth and Deep-depth questions, with raw retrieval becoming harmful on Deep questions. Thus, retrieval is useful when the evidence is near the surface of the thread, but becomes unreliable when the answer depends on evidence deeper in the reply structure.

An author-evidence coverage check explains why flat retrieval is weak. On average, BM25/E5 recover only 19.0%/19.2% of cited evidence nodes at k=20 and 30.4%/30.6% at k=50, only 3.2% of items reach $\geq$ 75% recall in the best of the four retrievers, and no item attains full coverage (Appendix F.5, Table 18).

5.5 Validity checks: shortcuts, item shape, and contamination

Reduced-input probes. Removing the thread substantially lowers accuracy: options-only, question-only, and title/root-only settings score 32.6%, 37.7%, and 36.5%, compared with 43.0% under the full-context Baseline. Keeping all comments but shuffling their order also lowers accuracy to 39.7%. These probes show that models do use information from the discussion itself, including reply order, rather than solving the benchmark from answer choices, question wording, or topic cues alone (Appendix F.6, Table 19).

The multiple-choice format gives stable rankings. We also test an open-answer format (no options shown). A response is correct if it gives the same content as the gold option: the same number for counting items, or the same social judgment for interpretive items. Open-answer accuracy is about 7.5 pp lower than multiple-choice, showing that the option set makes answering easier. At the same time, model rankings are almost unchanged, with Spearman ρ=0.96 on counting and ρ=0.99 on interpretive questions. Multiple-choice is therefore an objective, scalable evaluation target that pre-

serves open-answer model rankings (Appendix F.6, Tables 19, 20).

Counting vs. interpretive items expose different shortcuts. Reduced-input probes reveal a question-shape tradeoff (Appendix F.6, Table 21). Options-only counting scores 30.0%, 12.3 pp below the 42.3% counting Baseline. Interpretive items score 34.9%, 8.7 pp below Baseline (43.6%), indicating stronger option-text priors. The two shapes therefore stress complementary risks: interpretive items expose surface priors, while counting items resist leakage but require applying a social criterion across the reply graph.

Contamination check. Averaged over 18 models, post-cutoff Hacker News (2026 Q1) is 0.4 pp higher than pre-cutoff Stack Exchange (2008–2018), with per-tier signs split (-4.8 pp closed-frontier; $+7.7$ pp mid-tier; Appendix F.6, Table 23). Since Stack Exchange is also shallower (max-depth 2.0 vs. 12.9), the closed-frontier gap mixes memorization with a structure-depth advantage this probe cannot separate. Source-page memorization in the strongest tier is therefore bounded rather than ruled out, and is not the main driver of the headline gap.

Length is not the main driver. Model-facing prompts span 50K–153K cl100k_base tokens (median 96K), including the serialized episode and evaluation prompt. Across 94 episodes, the Spearman correlation between prompt length and per-episode mean Baseline accuracy is only $\rho{=}-0.23$, with linear slope -0.06 pp per 1K tokens. The effect is concentrated in mid-tier models ($\rho{=}-0.31$) and near zero for closed-frontier ($\rho{=}-0.01$), so prompt length alone cannot explain why the strongest models stay far below human accuracy (Appendix F.6.6, Fig. 12). Long inputs form the realistic setting in which the gap appears, not its source.

5.6 Where the Failure Happens

With shortcut, format, and contamination explanations ruled out or bounded, we localize the remaining errors to two stages: finding the relevant evidence scope and applying the social criterion within it. Supplying oracle scope yields tier-monotone lifts over Baseline: 15.3 pp for closed-frontier, 11.7 pp for strong open-weight, 9.6 pp for mid-tier, and 6.5 pp for smaller models (Appendix F.7, Table 25), so all tiers lose accuracy from failing to identify the relevant thread region.

A component oracle on Claude-Opus-4.7 separates scope recovery from criterion-specific interpretation. On a stratified 30-episode / 510-item subset, CoT accuracy is 59.5%; gold scope raises it to 69.0%, gold per-comment criterion labels to 79.5%, and both together to 98.6% (Appendix F.7, Table 26); a model-generated enumerate-label-aggregate pipeline reaches 67.0%. Models must therefore recover the evidence scope, then apply the item-specific social criterion; current methods improve parts of this process but none makes both stages reliable.

These results suggest that complete-thread social reasoning should be treated as a structured reading problem rather than a longer-context version of document QA. A model must identify the relevant social unit, such as a chain, subtree, participant trajectory, or comparison set, and then apply a criterion that may depend on how multiple comments jointly function. This explains why generic CoT and flat retrieval help only partially: they improve pieces of the pipeline, but do not enforce reliable scope selection and criterion application together.

6 Conclusion

We introduced LongSocialBench, 1,462 verified items over 94 real reply-structured discussion episodes. Across 18 models and 29 settings, Baseline, CoT, and retrieval variants remain far below human performance. Diagnostics trace the gap to two bottlenecks: recovering the relevant thread scope and applying the social criterion within it. These findings frame complete-thread social understanding as a structured reading problem. By making this failure measurable, LongSocialBench provides a diagnostic testbed for future work on long-context models that reason over reply chains, participant trajectories, and cross-branch contrasts.

Limitations

Scope of discussion data. LongSocialBench is limited to English, text-based, public reply-threaded discussions from Hacker News, Reddit r/ChangeMyView, and Stack Exchange. This design gives us complete episodes with explicit reply structure and stable public access, but it does not cover private chats, live conversations, multimodal posts, non-English discussions, or platforms with different interaction norms. Future work should test whether the same capability surfaces and failure patterns hold in broader conversational settings.

Benchmark format. LongSocialBench uses multiple-choice questions to make long-thread social reasoning objectively scoreable and auditable. We partially check format dependence with an open-answer probe, which preserves the main model ranking after answer options are removed. Still, the benchmark primarily evaluates evidence-grounded social judgments rather than free-form explanation quality, calibrated uncertainty, or faithful long-thread summarization. Future extensions can complement the current setup with richer answer formats and explanation-faithfulness checks.

References

Rob Abbott, Brian Ecker, Pranav Anand, and Marilyn Walker. Internet argument corpus 2.0: An SQL schema for dialogic social media and the corpora to go with it. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4445–4452, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1704/.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3119–3137, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.172. URL https://aclanthology.org/2024.acl-long.172/.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3639–3664, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.183. URL https://aclanthology.org/2025.acl-long.183/.

Amanda Bertsch, Adithya Pratapa, Teruko Mitamura, Graham Neubig, and Matthew R. Gormley. Oolong: Evaluating long context reasoning and aggregation capabilities. *arXiv preprint arXiv:2511.02817*, 2025. URL https://arxiv.org/abs/2511.02817.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. ToMBench: Benchmarking theory of mind in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15959–15983, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.847. URL https://aclanthology.org/2024.acl-long.847/.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. MuTual: A dataset for multi-turn dialogue reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1406–1416, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.130. URL https://aclanthology.org/2020.acl-main.130/.

Alexander R. Fabbri, Faiaz Rahman, Imad Rizvi, Borui Wang, Haoran Li, Yashar Mehdad, and Dragomir Radev. ConvoSumm: Conversation summarization benchmark and improved abstractive summarization with argument mining. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6866–6880, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.535. URL https://aclanthology.org/2021.acl-long.535/.

Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. URL https://arxiv.org/abs/2403.05530.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=kIoBbc76Sy.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. Revisiting the evaluation of theory of mind through question answering. 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)*, pages 5872–5877, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1598. URL https://aclanthology.org/D19-1598/.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. LooGLE: Can long-context language models understand long contexts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16304–16333, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.859. URL https://aclanthology.org/2024.acl-long.859/.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of

llm agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13851–13870, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.747. URL https://aclanthology.org/2024.acl-long.747/.

Fuqiang Niu, Min Yang, Ang Li, Baoquan Zhang, Xiaojiang Peng, and Bowen Zhang. A challenge dataset and effective models for conversational stance detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 122–132, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.11/.

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)*, pages 4463–4473, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1454. URL https://aclanthology.org/D19-1454/.

Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. ZeroSCROLLS: A zero-shot benchmark for long text understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7977–7989, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.536. URL https://aclanthology.org/2023.findings-emnlp.536/.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web*, pages 613–624, 2016. doi: 10.1145/2872427.2883081. URL https://doi.org/10.1145/2872427.2883081.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. HELMET: How to evaluate long-context models effectively and thoroughly. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/f5332c8273d02729730a9c24dec2135e-Abstract-Conference.html.

A Additional Analysis and Positioning

This appendix provides supporting positioning analysis for LongSocialBench.

A.1 Cited Area-Level Comparison

Table 4 expands the compact comparison in Table 1 with representative references. It summarizes how major evaluation areas differ in input unit, thread structure, social-reasoning target, and verification setup.

B Dataset Coverage and Episode Statistics

This appendix section supports the episode-sampling claims in §3.2. It reports the released snapshot, topic coverage, episode scale, and structural diversity of LongSocialBench.

B.1 Dataset Snapshot

All statistics in the paper are computed from **LongSocialBench v1.0**, the released benchmark snapshot containing 94 complete public discussion episodes and 1,462 verified items after verification and adjudication.

B.2 Topic Coverage

Table 5 summarizes the primary-topic coverage of the 94 released episodes. Because the three platforms do not share a single topic taxonomy, each episode is assigned one broad topic family based on its title and, for Stack Exchange, its site domain. These labels summarize dataset coverage and are not used as model-facing metadata or evaluation labels.

B.3 Episode Scale and Structural Diversity

Table 6 reports the scale and structural profile of the released episodes by source. The three sources provide distinct discussion shapes: Hacker News has the deepest reply chains on average, Stack Exchange is shallow but wide, and Reddit r/ChangeMyView contains deliberative multi-step exchanges. Fig. 4 visualizes the same coverage along serialized length, maximum depth, and platform-specific depth-frequency curves.

C Item Design and Construction Safeguards

This appendix section supports the task-design and item-authoring claims in §3.1 and §3.2. It expands the capability surfaces, gives a worked item example, and records the release criteria and shortcut controls used during authoring.

Table 4: **Cited area-level comparison with LongSocialBench.** ✓ = central in the typical setup; ▲ = partial; ✗ = generally absent.

Evaluation area	Representative work	Long	Threa	Social	Verif.	Positioning note
Long-context evaluation / aggregation	ZeroSCROLLS, LongBench, LooGLE, RULER, HELMET, LongBench v2, Oolong [Shaham et al., 2023, Bai et al., 2024, Li et al., 2024, Hsieh et al., 2024, Yen et al., 2025, Bai et al., 2025, Bertsch et al., 2025]	✓	▲	▲	✓	Long inputs and aggregation, with limited focus on complete reply-tree social QA.
Long-memory dialogue	LoCoMo [Maharana et al., 2024]	✓	▲	▲	▲	Evaluates memory over user–assistant histories rather than public multi-party discussions.
Social commonsense / ToM	SocialIQA, MuTual, ToMi, ToMBench [Sap et al., 2019, Cui et al., 2020, Le et al., 2019, Chen et al., 2024]	✗	✗	✓	✓	Focuses on localized social inference where the relevant context is already given.
Argument / stance / persuasion	IAC 2.0, ChangeMyView studies, MT-CSD [Abbott et al., 2016, Tan et al., 2016, Niu et al., 2024]	▲	✓	✓	▲	Provides labels or outcomes for real interactions, with limited verified complete-episode QA.
Discussion summarization	ConvoSumm [Fabbri et al., 2021]	▲	▲	✓	✗	Produces free-form summaries rather than exact-match social reasoning questions.
LongSocialBench	**This work**	✓	✓	✓	✓	Verifiable social QA over complete real reply-structured episodes.

Long = long-context input; **Thread** = naturally occurring reply-thread or branching structure; **Social** = social meaning, stance, persuasion, roles, or participant behavior; **Verif.** = objectively checkable QA or gold-answer verification.

Table 5: **Coarse topic coverage of released episodes.** Each episode is assigned one primary topic family based on its title and source domain.

Primary topic family	Total	HN	Reddit	SE
Computing / AI / internet	25	20	0	5
Economy / labor / business	21	10	10	1
Science / engineering / health	17	3	5	9
Public policy / law / governance	11	6	4	1
Culture / media / social life	7	3	4	0
Education / academia	6	0	4	2
Climate / energy / environment	4	1	3	0
Humanities / social thought	3	0	0	3
Total	**94**	**43**	**30**	**21**

Table 6: **Episode-level statistics by source.** Tokens are tiktoken `cl100k_base` counts over the full serialized thread. Max-depth is the deepest node distance from the root.

Source	# eps	Tokens med.	Tokens max	Tokens total	Mean max-depth	Mean nodes	Mean root replies
Hacker News	43	72,013	116,049	3.31M	12.9	815.7	137.2
Stack Exchange	21	72,776	84,454	1.66M	2.0	601.0	312.1
Reddit r/CMV	30	55,467	78,524	1.84M	8.9	485.5	66.2
Overall	**94**	**≈72K**	**116,049**	**≈6.81M**	**9.2**	**662.4**	**147.0**

C.1 Capability Surfaces in Detail

Table 7 expands the six capability surfaces introduced in §3.1. The table summarizes what each surface asks the model to decide and illustrates it with abridged examples adapted from human-authored items.

C.2 Worked Item Example

Fig. 5 shows how one LongSocialBench item is grounded in a real threaded discussion. The example compresses one scoped reply chain from a larger Reddit r/ChangeMyView episode and shows why the item requires following the chain rather than matching repeated keywords.

C.3 Release Criteria

Table 8 lists the release criteria used to decide whether a candidate item can enter the benchmark.

C.4 Shortcut and Contamination Controls

Table 9 summarizes the authoring controls used to reduce surface shortcuts, improve distractor quality,

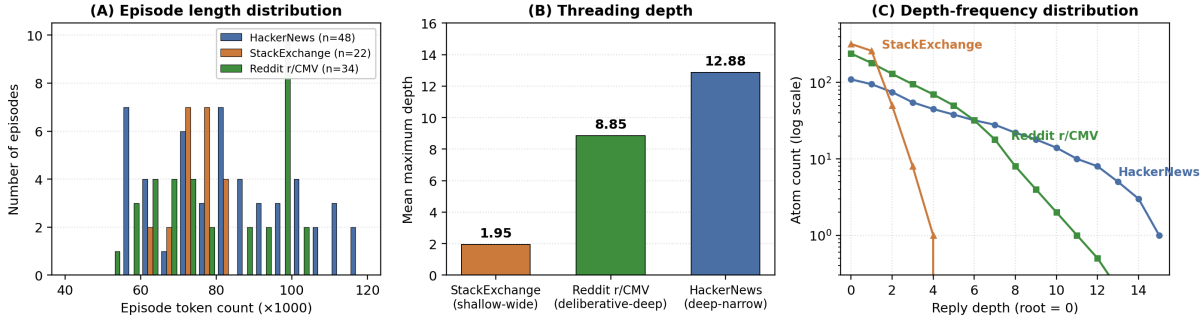


Figure 4: **Structural diversity of LongSocialBench episodes. (A)** Episode lengths span 37 to 116K tokens with platform-specific medians of 72K for Hacker News, 73K for Stack Exchange, and 55K for Reddit r/CMV. **(B)** Mean maximum depth differs substantially across sources: Stack Exchange 2.0, Reddit r/CMV 8.9, and Hacker News 12.9. **(C)** Depth-frequency curves show different threading regimes: Stack Exchange is concentrated near depths 1 to 2, Reddit has more mass at depths 1 to 6, and Hacker News retains substantial mass at depths 4 to 10.

Table 7: **Six capability surfaces in LongSocialBench.** Each surface is a recurring social-reading decision over a complete discussion episode.

Capability surface	Decision tested	Abridged example item
C1: Local social meaning	Decide what one comment is doing in context.	A target comment replies to a parent claim. Decide its **local role**: support, pushback, reframing, or clarification.
C2: Reply-pair interaction	Compare a reply with the comment it answers.	A parent uses one basis of comparison; the reply switches basis. Decide the **parent reply relation**: refute, qualify, redirect, or concede.
C3: Collective structure	Identify the positions organizing a scoped discussion region.	In a scoped subtree, choose the **position structure** that best organizes the replies: framing, consequences, responsibility, or evidence.
C4: Temporal dynamics	Locate where an ordered exchange changes direction.	In a long reply chain, identify the first **turning point** where one argumentative framing gives way to another.
C5: Participant profile	Track one participant across comments.	Across one author's comments, infer the **participant trajectory**: stable stance, narrowed claim, concession, escalation, or strategy shift.
C6: Cross-subdiscussion variation	Compare behavior across separate branches.	Across two branches, decide the **cross-branch change**: whether a role, stance, or strategy stays stable or shifts.

Table 8: **Release criteria for LongSocialBench items.** Each candidate item must satisfy all criteria before release; failed items are clarified and re-verified, or removed if the issue cannot be resolved.

Criterion	Requirement	Failure action
Answerability	The gold option is recoverable from the supplied episode alone, without external facts or author intent.	Rewrite or remove.
Option uniqueness	Exactly one option is supported under the stated scope and author-written criterion.	Revise options or remove.
Scope clarity	The stem specifies the relevant unit, such as a comment, direct-reply set, subtree, chain, participant trajectory, or cross-branch comparison.	Clarify stem and re-verify.
Evidence sufficiency	Evidence nodes support both the gold option and key premises such as branch boundary, counting unit, participant trajectory, or comparison set.	Add evidence, rewrite, or remove.
Observable social criterion	Interpretive questions are grounded in observable thread behavior rather than preference, speculation, or unstated norms.	Rewrite criterion or remove.
Option balance	Distractors are plausible wrong readings with comparable specificity, not fillers or wording artifacts.	Rewrite options and re-verify.

EPISODE **Reddit r/ChangeMyView** | 661 comments | longest reply chain (9 turns).

"I'm left-leaning, but the way the left treats the middle class is becoming alienating."

Released item. The released set contains an item that asks where this chain first shifts from the original class-and-privilege debate into a corporate wage-distribution argument.

STEPS 1–3 • class-and-privilege framing

1. Class-position framing: a participant argues that the frustration is about recognizing that people with more income and security experience the system differently from poorer people.

2. Baseline-affordability objection: the reply pushes back that formerly middle-class baselines such as housing and children now feel unaffordable.

3. Wealth-hoarding reframing: the original participant agrees the middle class is being squeezed, then redirects blame toward wealthy actors, investors, and institutions.

▼ **pivot** ▼

STEP 4 • CORPORATE-OUTPUT PIVOT (first turn that crosses into the wage-distribution debate)

A new participant introduces GDP output per worker, McDonald's revenue per employee, and worker take-home pay, shifting the discussion into corporate wage distribution.

STEPS 5–9 • wage-distribution dispute

The remaining turns debate whether revenue or profit is the right denominator, whether executive pay explains low worker wages, and whether company-level profits can be redistributed to employees.

QUESTION

Follow the episode's unique longest reply chain. At which position does the chain first move from the original class-and-privilege debate into a corporate wage-distribution argument? (Position 1 is the opening comment in the chain.)

A. Position 6 | **B.** Position 3 | **C.** Position 4 ★ | **D.** Position 5

GOLD ★ C | Position 4

The first three comments discuss middle-class treatment, privilege, housing, and wealth hoarding. The fourth comment introduces McDonald's revenue per employee, GDP per worker, and worker pay, which is the first move into a corporate wage-distribution argument.

Figure 5: **Worked item example.** A reader must first resolve the relevant conversational path, then interpret how each reply positions itself against its parent, and finally identify the first turn where the debate changes from class-and-privilege framing to corporate wage distribution. Colour bands separate the three trajectory phases (pre-pivot, pivot, post-pivot); the gold answer is the position where the chain crosses from one phase into the next.

and screen public-source contamination.

D Verification and Release Details

This appendix section supports the verification and release claims in §3.3. It records the independent verification protocol, release audit, capability-label reliability check, and failure taxonomy.

Table 9: **Question-design controls used during authoring.** The controls support shortcut-resistant multiple-choice evaluation over complete discussion episodes.

Control	Shortcut reduced	Authoring restriction
Minimal locators	The stem reveals the relevant node, chain, or subtree.	Use only the locator needed to define the scope; avoid exact quotes, usernames, redundant structural hints, or answer-implying descriptors.
Distributed evidence	A single salient comment solves a thread-level item.	Require evidence from the intended scope whenever the question asks for aggregation, comparison, trajectory, or cross-branch reasoning.
Controlled options	Option wording makes the answer obvious.	Use parallel option forms and close numeric or relational alternatives when appropriate; avoid answer-specific descriptive phrasing.
No hint vocabulary	Keyword search or hint words reveal the answer.	State the criterion without listing answer-bearing instances or repeating distinctive phrases from the gold evidence.
Diagnostic distractors	Distractors are implausible or too easy.	Tie distractors to plausible wrong scopes, wrong criteria, lexical matches, boundary errors, or aggregation mistakes.
Open-web contamination	The item is recoverable from open-web memory rather than the supplied episode.	Check distinctive stem/option phrases and named events; rewrite or exclude items whose answers can be recovered from open-web summaries or third-party transcripts.

D.1 Verification Protocol

Every released LongSocialBench item passes independent verification or adjudication. In Phase 1, the author writes the question stem, answer options, gold option, plain-language criterion, and evidence nodes. In Phase 2, a different trained contributor answers from the same model-facing input used in evaluation: the serialized thread, question stem, and four answer options, without the author gold option, criterion note, or evidence-node list. In Phase 3, verifier disagreements and author-flagged edge cases are checked by a third reviewer, who confirms the item, requires revision and a new Phase-2 verification pass, or removes it.

For the parseable candidate records, independent Phase-2 verification matches the author gold option in $1{,}419/1{,}553$ cases, yielding a candidate-level first-pass agreement of 91.37% and Cohen's κ=$\mathbf{0.88}$ against a four-option chance baseline (P_e=0.25). Of the 134 mismatches, 43 are resolved by adjudication and kept, while 91 unresolved items are excluded. Phase 3 also reviews 24 author-flagged edge cases that passed first-pass agreement. The total Phase-3 record count is 67, with verdicts 21 confirm, 40 clarify, and 6 rewrite.

D.2 Stratified Release Audit

The release audit estimates the residual error rate after verification and adjudication. We draw a 100-item stratified audit set balanced across capability surfaces and platforms, then check each item against structural criteria and a human re-read.

Table 10: **100-item stratified audit allocation.** Per-cell counts are proportional to released distribution with a floor of 2 per non-empty cap$\times$platform cell.

Domain	C1	C2	C3	C4	C5	C6	Total
hackernews	6	3	21	5	7	4	46
reddit	2	2	16	5	5	2	32
stackexchange	4	2	8	3	3	2	22
Total	12	7	45	13	15	8	100

Table 11: **Stratified release audit.** CORRECTED means a wording-level fix without changing the gold option. RESIDUAL means the adjudicator could not recover the gold option from the model-facing input.

Domain	Pass	Corrected	Residual
hackernews	44 / 46	2 / 46	0 / 46
reddit	32 / 32	0 / 32	0 / 32
stackexchange	22 / 22	0 / 22	0 / 22
All	**98 / 100**	**2 / 100**	**0 / 100**

Every capability is at $\geq 92\%$ pass; the only CORRECTED verdicts fall in C2 ($1/7$) and C3 ($1/45$).

The audit checks whether evidence nodes resolve, criteria match the intended scope, stems and options are well formed, infrastructure leaks are absent, option forms are balanced, and duplicate options are absent. An adjudicator then re-reads each audited item and assigns a PASS, CORRECTED, or RESIDUAL verdict.

Table 12: **Models used in the evaluation grid.** CW = configured context window in tokens.

Model	Source	CW
Closed-source frontier systems		
Claude-Opus-4.7	Anthropic	200K
GPT-5	OpenAI	400K
Claude-Sonnet-4.5	Anthropic	200K
Claude-4	Anthropic	200K
Gemini-2.5-Pro	Google	1M
Claude-3.7-Sonnet	Anthropic	200K
Open-weight systems		
Mistral-Large-3	HuggingFace	128K
Llama-4-Scout-17B	HuggingFace	128K
GLM-5	HuggingFace	128K
Qwen3-Next-80B	HuggingFace	256K
Llama-3.1-70B	HuggingFace	128K
DeepSeek-R1	HuggingFace	128K
Qwen3-32B	HuggingFace	128K
Gemma-3-27B	HuggingFace	128K
GPT-OSS-120B	HuggingFace	128K
Llama-4-Maverick-17B	HuggingFace	128K
Llama-3.2-90B	HuggingFace	128K
GPT-OSS-20B	HuggingFace	128K

D.3 Capability-Label Agreement

The C1 to C6 capability labels are author-assigned during construction. To estimate label reliability, a second trained contributor independently re-codes a stratified 100-item sample using the definitions in Table 7. The two coders agree on 86 items, yielding raw agreement 0.86 and Cohen's κ=**0.82** against the six-class chance baseline.

E Model and Workflow Details

This appendix provides the model list and workflow settings referenced in §4. It complements the main text by specifying the evaluated model set, configured context windows, workflow construction, and workflow-specific prompts.

E.1 Models

Table 12 summarizes the models used in the evaluation grid. The set covers closed-source frontier systems and open-weight systems across large, mid-tier, and smaller model families. We report the configured context window because full-context workflows require the serialized episode prompt to fit within the available window.

E.2 Workflow Settings and Prompts

Table 13 summarizes the 29 evaluation settings used in the paper, grouped as 22 deployable workflows (Baseline, CoT variants, retrieval variants) and 7 diagnostic settings (validity probes, scope diagnostics). The table records how each setting constructs the model input and guides the answer. All settings use benchmark model-facing inputs only; author-side gold labels, criteria, evidence nodes, verifier notes, and adjudication records are excluded from prompting.

F Additional Evaluation Results

This appendix supplements the Results section with diagnostic tables and figures for each major result claim. The organization follows the main Results sequence: headline robustness, capability profiles, reasoning prompts, retrieval diagnostics, validity and item-shape ablations, and component diagnostics.

F.1 Headline Robustness and Human Reference

This subsection supports the headline comparison in §5.1. Table 15 and Fig. 6 report the accuracy band of the deployable workflows together with the two scope diagnostics for reference, while Table 16 reports the human answerability reference used in the main comparison.

Representative workflow results.

F.2 Human Reference Protocol

The human reference reported in the main text is collected from five graduate-student readers with no prior exposure to LongSocialBench. They answer the same stratified 150-item sample under the same information condition as models: serialized thread, reply structure, question stem, and four answer options, with no criterion notes, author evidence records, or verifier annotations. The mean accuracy is 72.4%. These five readers are a separate cohort from the construction-time verifier pool, which involved the full 20-contributor team during Phase 2 cross-verification and reached 91.37% first-pass agreement (Cohen's κ=0.88) on the released items. We use this as an answerability and model-comparison reference.

F.3 Capability and Platform Profiles

This subsection supports the capability-level analysis in §5.2. Fig. 7 reports per-model capability accuracy under Baseline. Figs. 8 and 9 show platform-level patterns that refine the capability story.

Table 13: **Evaluation settings on LongSocialBench.** Deployable workflows are used for the headline accuracy comparison. Diagnostic settings test shortcut artifacts, multiple-choice format effects, or evidence localization and are reported separately.

Setting group	Variants	Input construction and prompt
Deployable workflows		
Full-context prompting	Full	The model receives the complete serialized discussion, reply-structure information, question stem, and four options; the prompt asks it to locate relevant evidence, compare options, and output one answer letter.
Reasoning prompts	CoT, CoT-Ev, CoT-Tp, CoT-El, Reflect	Same input as Full; the prompt adds one reasoning strategy before the final answer: free-form reasoning, evidence enumeration, temporal tracking, option elimination, or self-reflection.
Sparse retrieval	BM25-$k20$, BM25-$k50$	Comments are ranked by BM25 using the question stem as query; the top-k comments are passed in original thread order and answered with the shared answer format.
Dense retrieval	E5-$k20$, E5-$k50$	Comments and question stems are embedded with E5; the top-k comments by cosine similarity are passed in original thread order and answered with the shared answer format.
RAG with anchor notes	BM25/E5 $\times$ $k \in \{20, 50\}$ $\times$ anchors $\in \{1, 2\}$	Retrieved comments are supplemented with brief model-generated anchor notes that identify what each retrieved region contributes before final answering.
Structure-aware retrieval	Subtree-$k1$, Subtree-$k5$, Subtree-$k10$	The model receives the largest connected reply subtrees by node count, preserving local reply structure without using gold evidence, and answers from those subtrees.
LLM-as-retriever	LLM-r	The model first selects relevant comments or reply regions; the final answer prompt uses only the selected context. The selection prompt asks for evidence needed to answer the question, not merely lexical overlap.
Diagnostic settings		
Validity probes	Options-only, question-only, title/root-only, shuffled-thread, open-answer	The model receives a reduced or perturbed input; for open-answer, the four options are removed and the response is content-scored against the gold answer content. These probes test answer-option artifacts, stem leakage, topic familiarity, thread-order dependence, and multiple-choice format effects.
Scope diagnostics	Gold-scope, infer-scope	The model receives diagnostic scope information, either from gold scope metadata or from a separate inferred-scope prompt. These settings estimate the role of evidence localization rather than serving as standard deployment workflows.

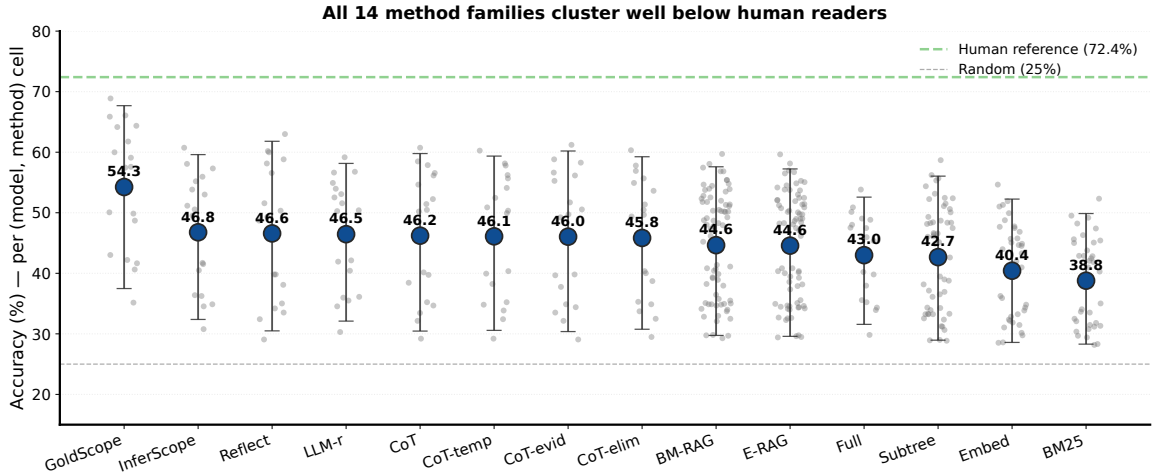


Figure 6: **Method-family accuracy band on LongSocialBench.** Each grey dot is one model-workflow cell from the 18×29 evaluation grid; markers and error bars show the family mean and ± 1 standard deviation across cells. All deployable workflows cluster below the human answerability reference of 72.4%; the Gold-Scope diagnostic reaches 54.3%.

Table 14: **Per-model accuracy on three representative workflows.** Full = full-context prompting; CoT = chain-of-thought; LLM-r = LLM-as-retriever. Best† is the per-model maximum across all evaluated non-oracle workflows, reported as a diagnostic only. C-mac = macro mean over the six capability surfaces.

Model	Full	CoT	LLM-r	Best†	C-mac
(Closed-source frontier)					
Claude-Opus-4.7	53.8	60.7	59.2	63.0	55.7
Claude-Sonnet-4.5	50.1	58.5	56.8	60.2	52.4
GPT-5	50.9	57.9	56.6	60.0	54.2
Gemini-2.5-Pro	49.0	56.6	54.9	58.1	49.6
Claude-4	48.0	56.2	54.0	58.8	50.8
Claude-3.7-Sonnet	47.5	54.7	53.1	56.6	50.0
(Strong open-weight)					
Mistral-Large-3	48.8	52.3	52.5	52.8	50.2
Llama-4-Scout-17B	47.5	51.4	51.6	51.8	46.7
Qwen3-Next-80B	45.3	50.5	50.4	50.5	45.7
GLM-5	46.0	50.2	50.3	50.3	44.5
(Mid open-weight)					
DeepSeek-R1	40.2	40.2	41.9	41.9	45.2
Llama-3.1-70B	39.8	39.7	42.1	42.1	41.4
Qwen3-32B	38.0	38.4	40.4	40.4	39.0
Llama-4-Mav-17B	35.2	35.2	36.1	36.1	33.7
(Small / mid-weak)					
Gemma-3-27B	35.6	34.7	36.0	36.0	38.9
GPT-OSS-120B	34.3	33.4	35.5	35.5	38.5
Llama-3.2-90B	33.9	32.1	34.5	34.5	35.1
GPT-OSS-20B	29.8	29.2	30.3	30.3	29.5
Mean (18)	43.0	46.2	46.5	47.7	44.5

Table 15: **Accuracy band by setting group, with cell-level bootstrap 95% CIs.** Values are full-benchmark micro accuracy on the 1,462 released items, averaged over the per-cell accuracies inside each group. ΔFull is the lift over Full-context (Baseline). Cells = number of (model, setting) cells in the group; the eight rows above the divider are deployable workflows, and the two rows below are scope diagnostics reported for reference.

Family	Cells	Mean	95% CI	ΔFull
LLM-as-retriever	18	46.5	[42.3, 50.5]	+3.5
Chain-of-thought	90	46.2	[44.0, 48.2]	+3.2
RAG k=50	72	45.5	[43.4, 47.7]	+2.5
RAG k=20	72	43.7	[41.8, 45.5]	+0.7
Full-context	18	43.0	[39.7, 46.2]	—
Structure-aware retrieval	54	42.7	[40.3, 44.9]	−0.3
Retrieval k=50	36	40.7	[38.1, 43.4]	−2.3
Retrieval k=20	36	38.5	[36.3, 40.7]	−4.5
Infer-Scope diagnostic	18	46.8	[42.4, 51.1]	+3.8
Gold-Scope diagnostic	18	54.3	[49.5, 58.8]	+11.3

F.4 Reasoning Prompt Effects

This subsection supports the reasoning-prompt analysis in §5.3. Table 17 summarizes lift across reasoning-style workflows, and Fig. 10 shows per-model reasoning lift over Baseline.

F.5 Retrieval Diagnostics

This subsection supports the retrieval analysis in §5.4. Table 18 measures whether standard comment-level retrievers recover the author-side evidence nodes used to justify gold answers. For each item, recall is the fraction of cited evidence nodes that appear in the retrieved comment set; mean recall is averaged over items and over the 18 evaluated models.

F.6 Validity and Item-Shape Ablations

This subsection supports the validity and item-shape analyses in §5.5. It reports shortcut probes, input-content gradients, non-counting subset results, contamination checks, and the matched 150-

Table 16: **Human answerability reference.** Five readers answer the same stratified 150-item sample without LongSocialBench-specific training.

Reader	Accuracy	# correct / 150
R1	74.7	112
R2	73.3	110
R3	72.7	109
R4	71.3	107
R5	70.0	105
Mean	**72.4**	**108.6**

95% bootstrap CI on the reader mean: $[71.1, 73.7]$. Random multiple-choice baseline: 25.0%.

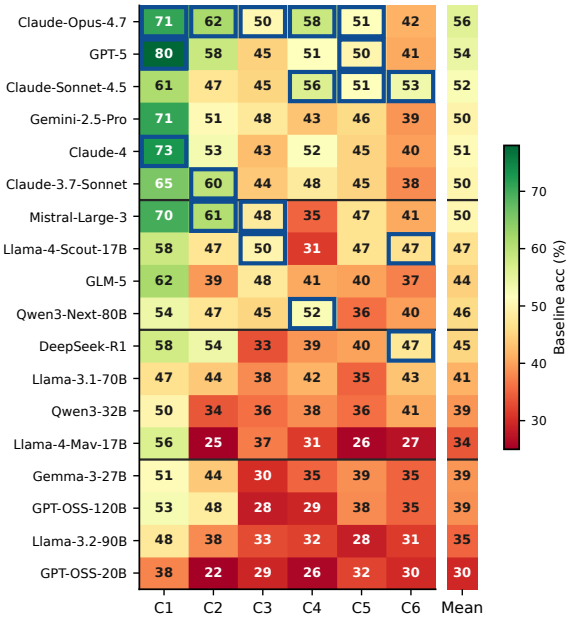


Figure 7: **Per-model accuracy on each social capability under Baseline.** Cells are mean accuracy on the 1,462-item benchmark restricted to that capability; the rightmost strip is the per-model mean.

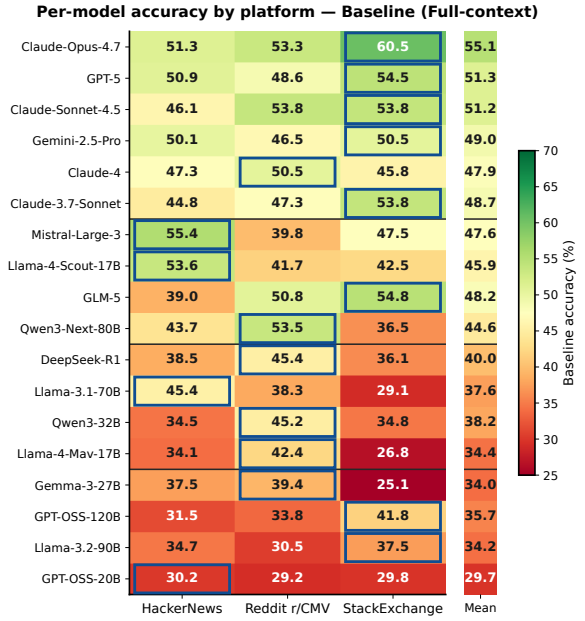


Figure 8: **Per-model accuracy by platform under Baseline.** Hacker News, Reddit r/CMV, Stack Exchange, and per-model mean are shown for each model.

item model-human comparison.

F.6.1 Shortcut Robustness

We run shortcut probes to check whether LongSocialBench can be solved from answer-option artifacts, question wording, topic familiarity, or bag-of-comments cues. All probes use stratified subsets by capability and source. Table 19 summarizes the shortcut tested and the observed result.

All probes are computed on the released set. For the open-answer probe, each model receives the same thread, title, and question stem as the multiple-choice setting but *without* the four answer options, and is asked to produce a short free-text answer. The free-text response is then content-scored against the gold-option text rather than parsed to a letter: a counting-style gold option such as "C. 5 answers" is credited if the response asserts the same count, and a textual gold option such as "A. real-world example, formal framework, philosophical argument" is credited if the response names the same items in the same ordering. This evaluates whether the model can recover the correct content unaided, rather than whether it can pick one of A–D. None of the surface conditions clears the Baseline accuracy of 43.0%.

Table 20 reports the per-model multiple-choice and open-answer accuracies, separated into counting items (numeric gold options, scored by exact match on the count) and textual items (free-text gold options, scored by content match on the named entities). Open-answer accuracy is uniformly lower than multiple-choice accuracy by roughly 7 pp, because removing the four-option pool removes the elimination shortcut and increases the cost of any single content-recall slip. The model ranking, however, is preserved on both subsets: Spearman ρ=0.987 on the full set, ρ=0.962 on the counting subset, and ρ=0.987 on the textual subset. The slightly lower counting correlation reflects a few mid-tier ties under exact-count scoring rather than a different ordering of strong vs. weak models.

F.6.2 Input-Content Gradient

Fig. 11 traces accuracy as the prompt moves from surface-only probes to stronger content and diag-

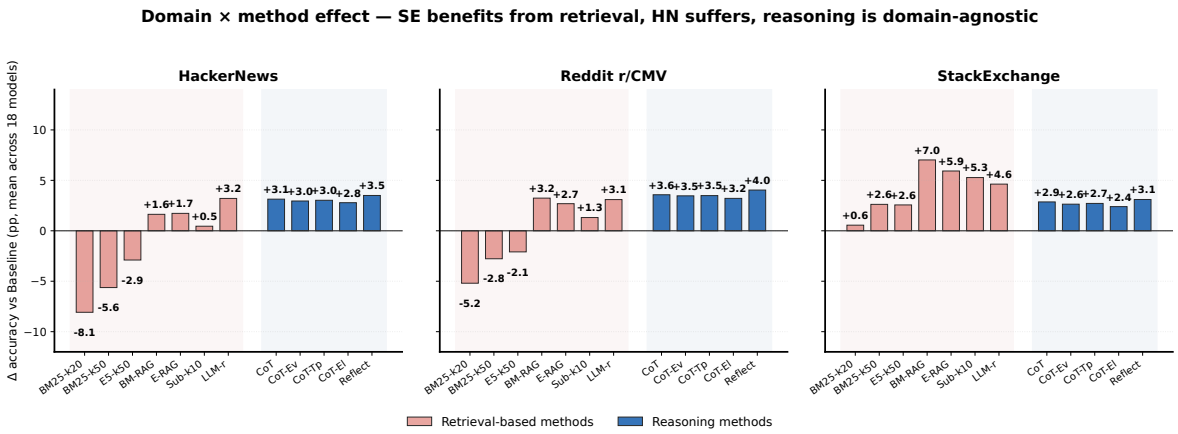


Figure 9: **Domain × method effect.** Each panel is one platform; bars are mean Δ accuracy vs. Baseline averaged across 18 models.

Table 17: **Reasoning workflow lift over Baseline, per model.** Each cell is accuracy under the named workflow minus accuracy under Baseline (pp). Negative entries (red) indicate the workflow hurts the model. Models are sorted by mean lift descending. Reflect = self-reflection; CoT-Evid / CoT-Temp / CoT-Elim are chain-of-thought variants prompting evidence enumeration, temporal reasoning, or elimination of distractors, respectively.

Model	CoT	CoT-Evid	CoT-Temp	CoT-Elim	Reflect	**Mean**
Claude-Sonnet-4.5	+8.3	+8.7	+8.0	+7.7	+10.1	+8.5
Claude-4	+8.1	+8.1	+7.7	+6.9	+10.8	+8.3
Gemini-2.5-Pro	+7.5	+7.6	+7.2	+6.6	+9.1	+7.6
Claude-3.7-Sonnet	+7.2	+7.8	+6.6	+6.2	+9.1	+7.4
GPT-5	+7.0	+7.4	+6.9	+6.0	+9.1	+7.3
Claude-Opus-4.7	+6.9	+7.4	+6.4	+6.5	+9.2	+7.3
Qwen3-Next-80B	+5.1	+3.9	+4.6	+4.0	+3.4	+4.2
GLM-5	+4.2	+3.8	+4.4	+3.7	+3.4	+3.9
Llama-4-Scout-17B	+3.9	+3.0	+3.4	+2.9	+2.8	+3.2
Mistral-Large-3	+3.4	+2.9	+3.6	+2.5	+2.7	+3.0
Qwen3-32B	+0.5	-0.2	+0.3	+0.8	+0.1	+0.3
Llama-3.1-70B	-0.1	-0.1	+0.1	+0.2	+0.0	+0.0
Llama-4-Mav-17B	+0.0	-0.3	+0.0	+0.1	-0.2	-0.1
DeepSeek-R1	-0.1	-0.4	+0.1	+0.1	-0.4	-0.1
GPT-OSS-20B	-0.6	-0.8	-0.6	-0.3	-0.8	-0.6
GPT-OSS-120B	-0.9	-0.8	-0.5	-0.6	-0.8	-0.7
Gemma-3-27B	-0.9	-1.2	-0.8	-0.7	-1.4	-1.0
Llama-3.2-90B	-1.8	-1.8	-1.5	-1.4	-1.5	-1.6
Mean (18)	+3.2	+3.1	+3.1	+2.8	+3.6	+3.2

Table 18: **Evidence-node recall for standard comment-level retrievers.** Values are mean recall against author-side evidence-node lists, averaged over items and over the 18 evaluated models on the released 1,462 items. No item achieves full evidence coverage in any of the four retrievers; only 3.2% reach ≥ 75% recall in the best of the four.

Retriever	**Budget**	**Mean evidence recall**
BM25	k=20	0.190
E5	k=20	0.192
BM25	k=50	0.304
E5	k=50	0.306

nostic support. It supports the validity claim in §5.5: surface probes are not enough to solve the benchmark, while extra content helps only when the model can use the provided evidence.

F.6.3 Surface Probes Split by Counting vs. Textual Items

To support the counting–textual tradeoff in §5.5, Table 21 reports each surface probe split into counting items (numeric gold options) and textual items (free-text gold options). The difference between the two columns is the headline finding: textual items retain meaningful priors from the question stem and option text alone, while counting items

Table 19: **Shortcut robustness summary.** Each probe removes or perturbs information that would be unnecessary if a shortcut were sufficient. Probe accuracies are means across 18 models on the full set; Baseline reaches 43.0% for reference.

Probe	Shortcut tested	Observed	Interpretation
Option-only	option prior	32.6%	below Baseline; option text alone is insufficient
Question-only	stem leakage	37.7%	question wording alone falls 5.3 pp short of Baseline
Title/root-only	topic familiarity	36.5%	topic-level information is insufficient without replies
Shuffled thread	bag-of-comments inference	39.7%	perturbing local thread order reduces performance
Open-answer subset	multiple-choice artifact	$\rho = 0.99$	open-answer ranking remains strongly correlated with multiple-choice ranking

Table 20: **Per-model multiple-choice (MC) and open-answer (Open) accuracy on the 1,462 released items.** Counting items have numeric gold options scored by exact match on the count; textual items have free-text gold options scored by content match on the named entities or ordering. Δ is Open$-$MC. Across all 18 models, model rankings are preserved between the two formats (Spearman ρ=0.987 overall, 0.962 counting, 0.987 textual).

	Full set			Counting		Textual	
Model	MC	Open	Δ	MC	Open	MC	Open
Claude-Opus-4.7	53.8	45.6	−8.3	50.6	42.9	56.7	48.0
GPT-5	50.9	43.7	−7.2	48.4	41.9	53.1	45.2
Claude-Sonnet-4.5	50.1	43.7	−6.4	49.8	43.3	50.5	44.1
Gemini-2.5-Pro	49.0	42.1	−6.9	48.7	41.2	49.6	43.2
Mistral-Large-3	48.8	40.9	−7.9	48.8	41.4	48.9	40.4
Claude-4	48.0	41.0	−7.0	44.6	38.9	51.2	42.9
Llama-4-Scout-17B	47.5	39.7	−7.9	48.4	40.5	46.9	39.1
Claude-3.7-Sonnet	47.5	39.7	−7.7	44.0	37.9	50.5	41.3
GLM-5	46.0	36.7	−9.2	45.7	36.1	46.3	37.3
Qwen3-Next-80B	45.3	36.5	−8.8	45.4	37.6	45.1	35.3
DeepSeek-R1	40.2	31.1	−9.2	40.2	30.3	40.0	31.6
Llama-3.1-70B	39.8	31.5	−8.3	39.9	31.5	39.6	31.2
Qwen3-32B	38.0	29.6	−8.3	37.3	28.4	38.3	30.5
Gemma-3-27B	35.6	27.5	−8.1	36.2	28.1	34.9	27.1
Llama-4-Maverick-17B	35.2	29.0	−6.2	37.9	31.5	32.5	26.5
GPT-OSS-120B	34.3	28.2	−6.2	33.8	27.4	35.1	29.1
Llama-3.2-90B	33.9	26.5	−7.4	31.0	23.8	36.7	29.1
GPT-OSS-20B	29.8	25.6	−4.2	30.1	25.0	29.5	26.0
Mean	43.0	35.5	−7.5	42.3	34.9	43.6	36.0

collapse toward 4-option chance once the thread is removed. Question-only is the strongest illustration: textual accuracy reaches the Baseline level (43.1% vs. 43.6%), while counting stays close to chance (31.9%, only $\sim$ 7 pp above 25%). On the open-answer probe the counting–textual gap closes to $\sim$ 1 pp because content scoring already removes the elimination shortcut on textual items. The pattern reflects benchmark design rather than a flaw: textual items must be distractor-resistant against semantic priors, while counting items must be enumeration-correct against an explicit social criterion, and the two question shapes constrain different families of shortcut.

Table 21: **Surface probes split by counting vs. textual items.** Mean accuracy across the 18 models; counting items have numeric gold options, textual items have free-text gold options. The Baseline row is included as a reference for the within-subset ceiling on the same items.

Setting	Overall	Counting	Textual
Baseline	43.0	42.3	43.6
Options-only	32.6	30.0	34.9
Question-only	37.7	31.9	43.1
Title/root-only	36.5	34.1	38.8
Shuffled thread	39.7	37.9	41.3
Open-answer	35.5	34.9	36.0

4-option chance is 25%. Counting items collapse to 30–32% once the thread is removed (options-only, question-only); textual items reach 43.1% from question + options alone, essentially matching the within-subset Baseline of 43.6%.

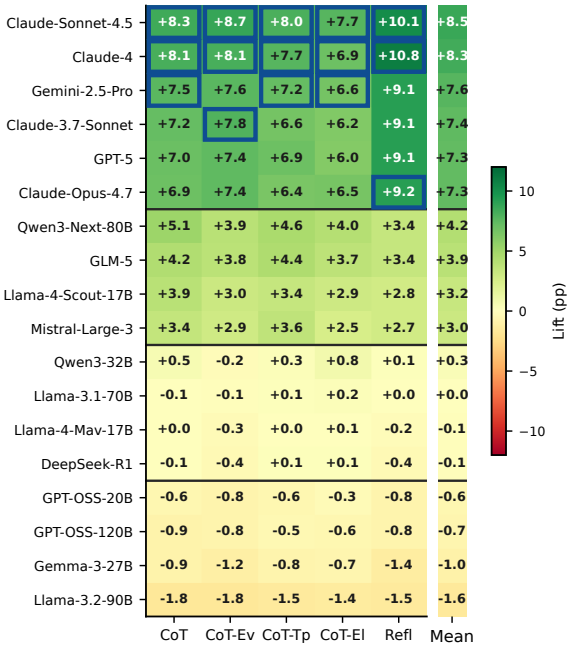


Figure 10: **Reasoning workflow lift over Baseline, per model.** Cells are accuracy under the named workflow minus Baseline in percentage points; the rightmost strip is the per-model mean lift.

F.6.4 Non-Counting Subset

Counting items in LongSocialBench are socially conditioned counts, such as direct replies that push back on a parent, comments that cross a turning point, or the fraction of a participant's contributions that defend a position. To check whether the headline gap is driven by an arithmetic-counting confound, we restrict to the 529 non-counting items and recompute the per-cell accuracy grid. Table 22 reports the result: the top single deployable cell rises by only 1.6 pp, from 63.0% to 64.6%, and the 18-model best-cell mean rises by 0.7 pp, from 47.7% to 48.4%. The gap to the human reference therefore persists across question types.

Table 22: **Non-counting subset headline.** Best-cell statistics on the 529 non-counting items vs. the full set.

Aggregation	Full	Non-cnt.	Δ
Top single cell	63.0	64.6	+1.6
18-model best-cell mean	47.7	48.4	+0.7

Top cell on both subsets is Claude-Opus-4.7 with self-reflection. Counting items are classified by stem and numeric option signature.

F.6.5 Pre- and Post-Cutoff Contamination Probe

We test whether memorization of source pages at pretraining time inflates accuracy on platforms that overlap pretraining corpora. The probe restricts to two time-disjoint cohorts whose dates are unambiguous from public metadata: Stack Exchange, collected from 2008 to 2018 archives, and Hacker News, collected from 2026 Q1. Reddit r/CMV episodes span 2016 to 2025 and are reported separately for context. If pretraining-time recall were a primary advantage, accuracy on the Stack Exchange cohort should be systematically higher than on the Hacker News cohort. Table 23 reports the per-tier Baseline means.

Table 23: **Pre- and post-cutoff contamination probe.** Mean Baseline accuracy by tier on the two time-disjoint cohorts; Reddit is reported for context. $\Delta_{\text{HN}-\text{SE}}$ is the post-cutoff minus pre-cutoff gap.

Tier	HN	Reddit	SE	$\Delta_{\text{HN}-\text{SE}}$
Closed frontier (6)	48.4	50.0	53.2	−4.8
Strong open (4)	47.9	46.5	45.3	+2.6
Mid open (4)	39.0	42.0	31.3	+7.7
Small (4)	32.6	34.0	33.9	−1.3
All 18	**42.7**	**43.9**	**42.3**	**+0.4**

HN = Hacker News post-cutoff cohort, 2026 Q1. SE = Stack Exchange pre-cutoff cohort, 2008 to 2018. Reddit r/CMV span is mixed and excluded from the formal probe.

Averaged over all 18 models, the post-cutoff platform is 0.4 pp higher than the pre-cutoff platform, but the per-tier signs alternate: closed-frontier −4.8 pp, strong-open +2.6 pp, mid-open +7.7 pp, small −1.3 pp. Stack Exchange is also structurally shallower (mean max-depth 2.0 vs. 12.9 for Hacker News), so the closed-frontier −4.8 pp gap is jointly consistent with partial memorization and a structure-depth advantage that this probe cannot separate. We therefore bound rather than rule out source-page memorization in the strongest tier, and conclude only that it is not the main driver of the fleet-level headline gap.

F.6.6 Episode-Length Effect

LongSocialBench episodes vary substantially in size: the 94 released episodes span 50,475 to 152,528 cl100k_base tokens, with median 96,312 and inter-quartile range 75,720 to 99,972. Because all 18 models in our pool report context windows of at least 128K tokens, the benchmark stays inside their advertised long-context range, but the variation is large enough that length could in principle

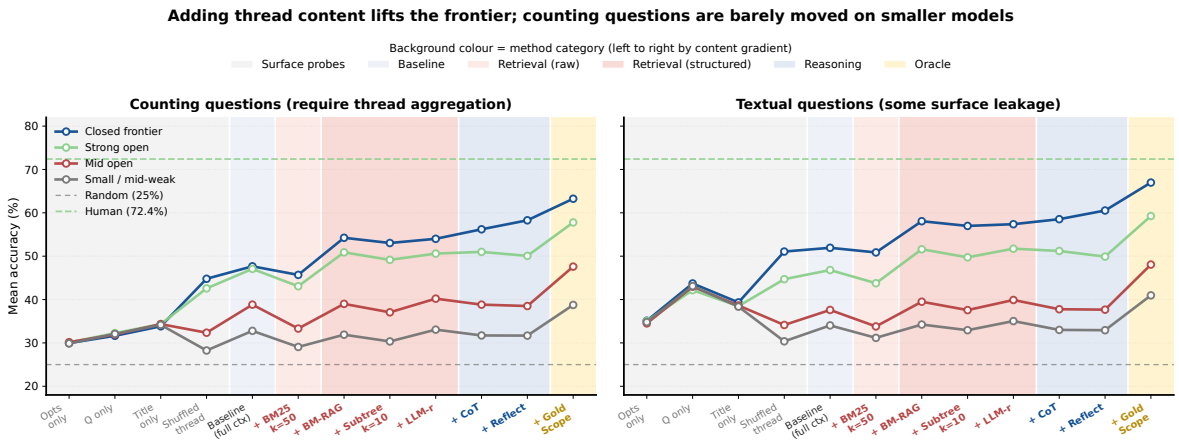


Figure 11: **Accuracy as a function of thread content, by tier and question type.** Conditions are ordered from surface probes to diagnostic support. Counting questions sit well below Baseline on surface probes; textual questions retain weak topic or wording priors; stronger content lifts frontier models more than smaller models.

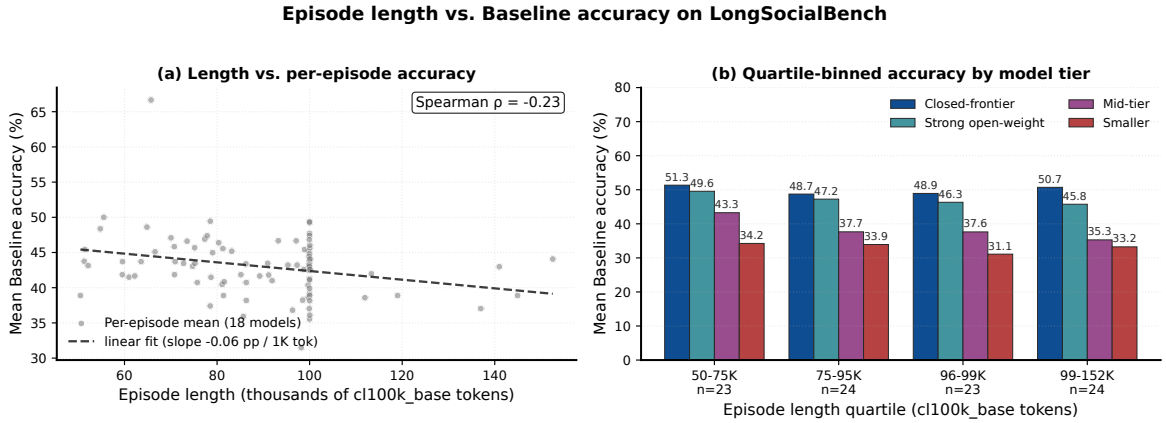


Figure 12: **Prompt length vs. Baseline accuracy.** (a) Per-episode mean accuracy across 18 models against model-facing prompt length in `cl100k_base` tokens, with linear fit and pooled Spearman ρ. (b) Prompt-length quartile means by model tier; quartiles are computed over the 94 released episodes. Prompt length is only weakly anti-correlated with accuracy and the per-tier rankings are unchanged across quartiles, so input length does not explain the headline gap.

dominate the headline gap. We therefore check whether per-episode Baseline accuracy degrades with episode length.

Figure 12(a) plots the per-episode mean Baseline accuracy (averaged across the 18 models) against episode length in tokens, together with a least-squares fit. The fit has slope -0.06 pp per 1K tokens, and the Spearman rank correlation between length and accuracy is $\rho{=}{-}0.23$. Length is therefore weakly anti-correlated with accuracy: doubling the input from 50K to 100K tokens predicts roughly a 3 pp drop, much smaller than the 25 to 30 pp gap to the human reference.

The effect is also unevenly distributed across tiers. The per-tier Spearman correlations are -0.01 for closed-frontier, -0.08 for strong open-weight, -0.31 for mid-tier, and -0.11 for smaller models. Length therefore matters most for mid-tier models – the same tier that gains least from CoT and from oracle scope – while the strongest models are essentially flat in length.

Figure 12(b) makes the same point at the bucket level: across the four length quartiles, the closed-frontier mean moves only between 48.7% and 51.3%, the strong-open mean stays inside 45.8% to 49.6%, and the rank order of the four tiers is unchanged in every quartile. The longest quartile is harder than the shortest by roughly 3 pp on average, but no tier collapses in long episodes and no tier breaks out in short ones. Together with the contamination probe and the shortcut/format probes, this rules out raw episode length as the dominant explanation for the headline social-reasoning gap.

F.6.7 Stratified 150-Item Matched-Aggregation Subset

The headline mean-vs-mean comparison computes the model number on the full benchmark and the human number on a stratified 150-item subsample. We re-run the model evaluation on a stratified 150-item subset whose capability-platform proportions match the human reference sample, bootstrap-resampled 1000 times. Table 24 reports best-cell statistics matched to the human comparator.

Table 24: **Stratified 150-item matched-aggregation subset.** Bootstrap (B=1000) over draws of 150 items in the natural capability-platform proportions. Each row reports best-cell mean and the 95% CI across draws; human-reader numbers are for reference.

Cohort	Mean	95% CI
Top single cell	65.2	[60.7, 71.3]
Claude-Opus-4.7	62.9	[54.7, 70.0]
Closed-frontier mean (n=6)	59.3	—
Strong-open mean (n=4)	52.0	—
18-model fleet mean	47.8	—
Human readers (n=5)	**72.4**	[**71.1**, **73.7**]
Worst human reader (R5)	70.0	—

Mean = best-cell mean across 1000 bootstrap draws of 150 stratified items. Best-cell = each model's highest-accuracy deployable workflow on the draw.

F.7 Scope and Component Diagnostics

This subsection supports the component-diagnostic analysis in §5.6. Table 25 reports Gold-Scope lift across model tiers, and Table 26 reports the component oracle case study on the stratified subset.

Table 25: **Per-tier Gold-Scope diagnostic lift over Baseline.** Mean accuracy across the models in each tier on the full 1,462 items.

Tier	Base	Gold-Sc	Lift
Closed frontier (6)	49.9	65.2	+15.3
Strong open (4)	46.9	58.6	+11.7
Mid open (4)	38.3	47.9	+9.6
Small (4)	33.4	39.9	+6.5
All 18	43.0	54.3	**+11.3**

Table 26: **Component oracle case study.** Claude-Opus-4.7 on the stratified 30-episode / 510-item subset.

Condition	Accuracy
CoT baseline	59.5
Gold scope	69.0
Gold per-comment labels	79.5
Gold scope + gold labels	98.6
Structured enumerate-label-aggregate pipeline	67.0
Scope Jaccard in generated pipeline	0.71
Label macro-F1 in generated pipeline	0.68

G Representative Questions

This appendix shows representative LongSocialBench items that illustrate the kinds of social-reading decisions tested by the benchmark. The examples are selected for readability and diversity across reasoning styles rather than to enumerate every capability surface. To keep the paper aligned with the released benchmark, the question stems and answer options below are copied from released LongSocialBench v1.0 items; the rationales are shortened summaries of the released criteria. Usernames are replaced with structural references where needed.

G.1 Example 1: Local Social Meaning

Local Social Meaning *HackerNews;* `hn_46419970/q9_v6`

Episode topic. A HackerNews discussion of a news story about a passenger being unable to leave a Deutsche Bahn train. One sub-discussion defends German rail by citing positive personal travel experiences.

Question. *"One top-level comment defends DB as Germany's most comfortable transportation mode and cites two on-time Berlin-to-Konstanz Christmas trips. Among its direct replies, how many push back against that praise rather than endorse it?"*

Options. A. Two replies B. Three replies **C. Four replies** D. All five replies

Gold and rationale. C. The released criterion counts only direct replies to the DB-praise comment. Four replies push back against the praise, while one reply endorses the praise or calls for a sense of proportion. This item tests whether a model can interpret the local social role of replies relative to their parent comment rather than only recognizing the topic "Deutsche Bahn."

G.2 Example 2: Temporal Dynamics Along a Reply Chain

Temporal Dynamics *HackerNews;* `hn_46488576_q04`

Episode topic. A HackerNews discussion of a blog post titled "Web development is fun again," where a long reply chain starts with whether LLM output helps users learn and later shifts into the legality and ethics of training data.

Question. *"Walk down the longest reply chain that passes through the comment saying an LLM teaches you nothing. At which position does the discussion permanently shift from whether LLM output teaches the user anything to the legality and ethics of LLM training data? Count the chain's first comment as position one."*

Options. A. Position five B. Position eleven C. Position seven **D. Position nine**

Gold and rationale. D. The released criterion notes that the episode has two maximum-length chains, so the item explicitly scopes the chain by requiring it to pass through the comment saying an LLM teaches the user nothing. In that scoped 13-comment chain, positions 5–8 debate whether LLM-generated output is learnable; position 9 introduces the training-data legality and ethics pivot. This item tests ordered path following and topic-shift detection.

G.3 Example 3: Collective Discussion Structure

Collective Discussion Structure *Reddit*
r/ChangeMyView; `reddit_kz6qz2/q1_v6`

Episode topic. A ChangeMyView discussion argues whether a minimum wage indexed to local cost of living would solve problems caused by minimum-wage increases. The largest top-level thread debates whether higher wages would drive higher rents.

Question. *"In the largest top-level reply thread, look only at direct replies to its opening comment whose text is visible. How many either ask for evidence, cite a source, or use a specific place to challenge the claim that higher wages would cause higher rents? Do not count bare denials or purely theoretical replies."*

Options. A. Four replies B. Eight replies **C. Six replies** D. Three replies

Gold and rationale. C. The released criterion scopes the item to visible direct replies to the opening comment of the largest top-level thread. Six replies meet the inclusion rule by explicitly asking for empirical evidence, citing a study or article, or using a named real-world place as an empirical counterexample. This item tests aggregation over a local reply set with a non-trivial inclusion rule.

H Construction Workflow

This appendix documents the portal-based workflow used to author, verify, and adjudicate LongSocialBench items. The key design is role separation: contributors cannot verify or adjudicate items they authored, and human-reference readers do not participate in item construction.

H.1 Annotation and Verification Protocol

Table 27 summarizes the contributor-facing protocol. The protocol standardizes item authoring, artifact control, independent verification, and adjudication while keeping model-facing inputs separate from internal construction records.

H.2 Audit Portal

Figs. 13–15 show the three portal stages: authoring, independent verification, and adjudication. The portal records construction and review status for each item while preserving a clean separation between model-facing inputs and internal annotation records.

I Ethics, Licensing, and Contributor Protocol

This appendix documents the data-handling, contributor, and intended-use considerations behind LongSocialBench.

Source data and licensing. The released episodes are drawn from three public reply-threaded sources with established research-use norms: Hacker News, Reddit r/ChangeMyView, and Stack Exchange. Stack Exchange content is governed by the platform's standard content-license terms with attribution; Hacker News public threads are accessed in line with the platform's research-use guidance; and Reddit r/ChangeMyView episodes are processed under the platform's then-current public-research terms. We redistribute only what is needed for benchmark evaluation: the serialized thread text with reply structure used by the items, with usernames replaced by structural references and personally identifying spans removed where surfaced during verification.

Contributors and readers. Contributor recruitment and tasking follow a written protocol with role separation between authoring, independent verification, and adjudication, as summarized in Appendix H. Contributors completed a short calibration step before authoring or verifying released items. Contributor and human-reader participation was compensated at rates consistent with institutional norms for annotation work of comparable difficulty, and the study was conducted under the

Table 27: **LongSocialBench annotation and verification protocol.** Authoring records define the social criterion and supporting evidence; verification records check whether the gold option is recoverable by an independent reader under the same input condition given to models.

Stage	Contributor task	Quality control
Authoring	Read the complete discussion, choose a clear scope, write the question stem, four plausible options, verified gold option, plain-language criterion, and evidence nodes.	The item must require reasoning over the specified scope rather than topic familiarity or one salient comment. Distractors should be plausible wrong readings, such as wrong scope, nearby count, boundary error, or near-synonym social label.
Artifact control	Remove answer-bearing wording from stems and options, including exact quotes, usernames, redundant structural hints, and distinctive phrases.	This reduces locator leakage, option paraphrase matching, keyword shortcuts, and open-web contamination. Capability and item-form tags are used only as metadata.
Independent verification	A different contributor re-answers the item from the same model-facing input used in evaluation: serialized thread, reply structure, question stem, and four answer options.	The verifier does not see the author's gold option, criterion notes, evidence-node list, or rationale. Agreement checks both answer stability and scope stability.
Adjudication	Phase-2 mismatches and author-flagged borderline cases receive a third-pass review.	The item is confirmed, clarified, rewritten and reverified, or removed. A released item must have independent agreement or a stable adjudicated answer.

institutional process applicable to public-text annotation at our home institution.

I.1 Contributor Pool and Authoring Guidelines

The item-construction pool consisted of 20 invited contributors with prior experience reading and analyzing long-form English discussions, including graduate students and researchers in NLP, social computing, computational social science, HCI, or related areas. All contributors were fluent English readers and were instructed to author items from the supplied discussion episode alone, without using external knowledge or platform-specific assumptions. Before contributing released items,

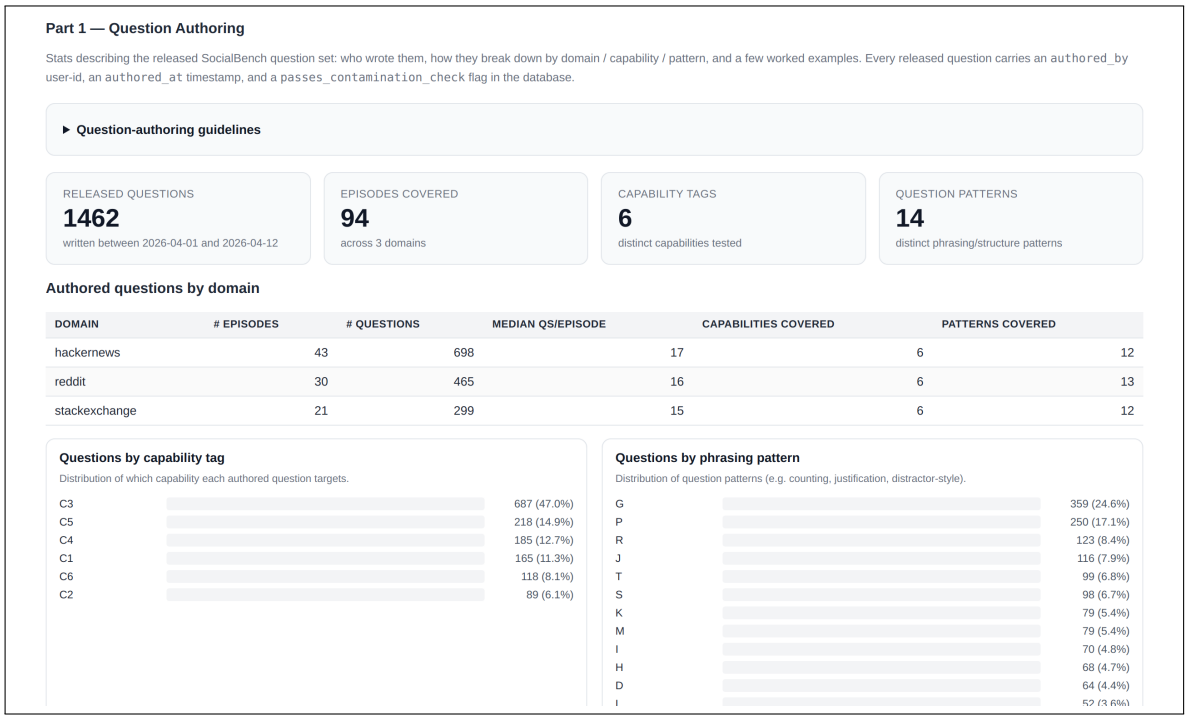
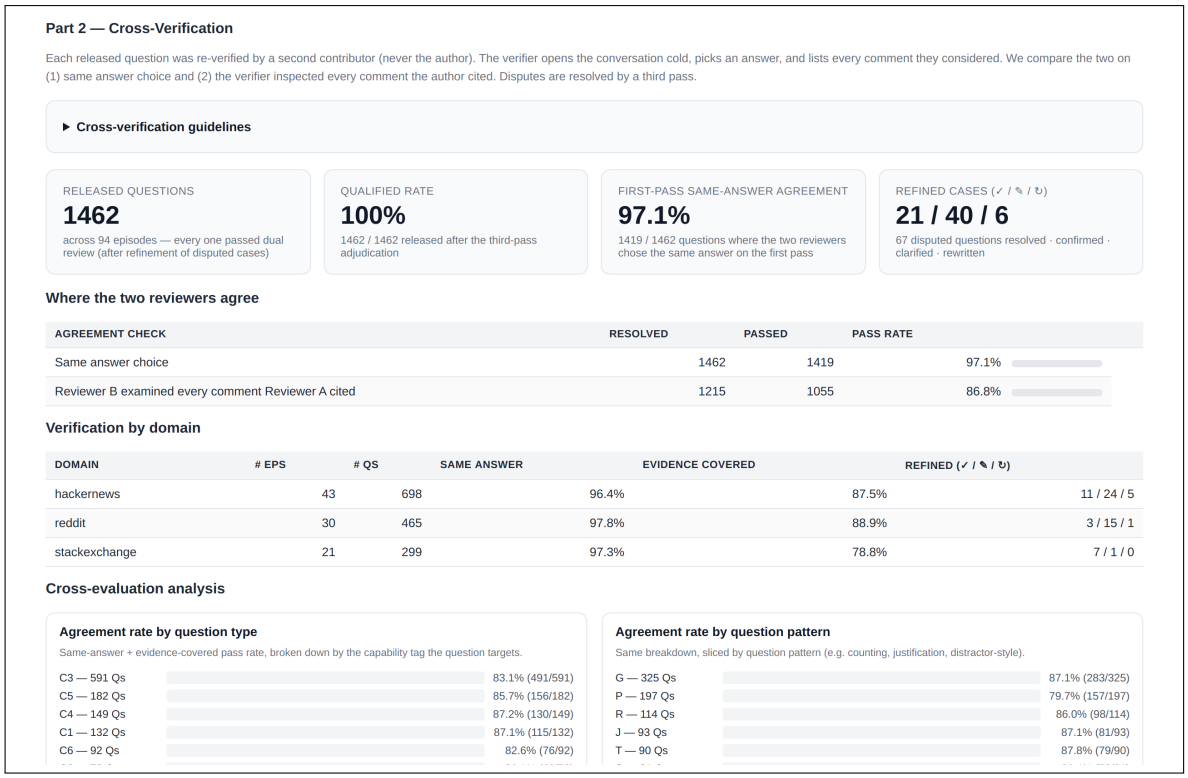


Figure 13: **LongSocialBench portal: authoring view.** Authors create the question stem, four answer options, verified gold option, plain-language criterion, and node-level evidence notes. Evidence notes support verification and audit but are not shown to evaluated models.



Part 2 — Cross-Verification

Each released question was re-verified by a second contributor (never the author). The verifier opens the conversation cold, picks an answer, and lists every comment they considered. We compare the two on (1) same answer choice and (2) the verifier inspected every comment the author cited. Disputes are resolved by a third pass.

▸ **Cross-verification guidelines**

RELEASED QUESTIONS	QUALIFIED RATE	FIRST-PASS SAME-ANSWER AGREEMENT	REFINED CASES (✓ / ✎ / ↻)
1462	**100%**	**97.1%**	**21 / 40 / 6**
across 94 episodes — every one passed dual review (after refinement of disputed cases)	1462 / 1462 released after the third-pass adjudication	1419 / 1462 questions where the two reviewers chose the same answer on the first pass	67 disputed questions resolved · confirmed · clarified · rewritten

Where the two reviewers agree

AGREEMENT CHECK	RESOLVED	PASSED	PASS RATE
Same answer choice	1462	1419	97.1%
Reviewer B examined every comment Reviewer A cited	1215	1055	86.8%

Verification by domain

DOMAIN	# EPS	# QS	SAME ANSWER	EVIDENCE COVERED	REFINED (✓ / ✎ / ↻)
hackernews	43	698	96.4%	87.5%	11 / 24 / 5
reddit	30	465	97.8%	88.9%	3 / 15 / 1
stackexchange	21	299	97.3%	78.8%	7 / 1 / 0

Cross-evaluation analysis

Agreement rate by question type

Same-answer + evidence-covered pass rate, broken down by the capability tag the question targets.

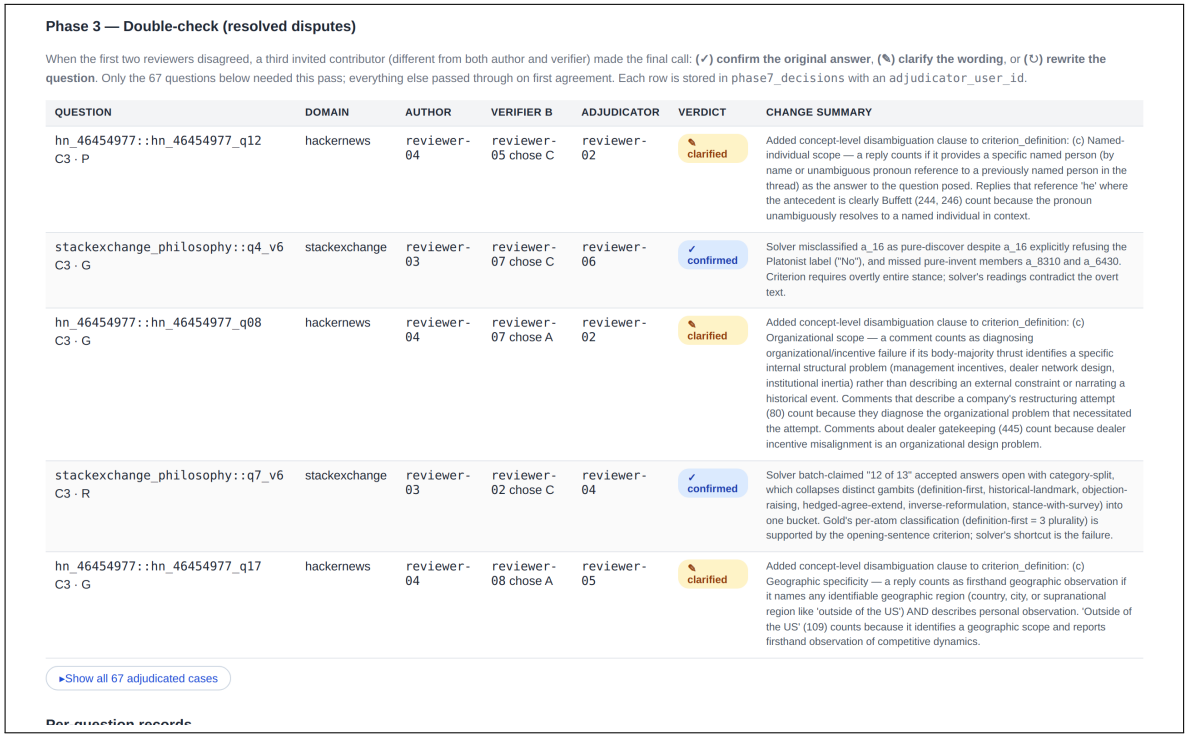
C3 — 591 Qs	83.1% (491/591)
C5 — 182 Qs	85.7% (156/182)
C4 — 149 Qs	87.2% (130/149)
C1 — 132 Qs	87.1% (115/132)
C6 — 92 Qs	82.6% (76/92)

Agreement rate by question pattern

Same breakdown, sliced by question pattern (e.g. counting, justification, distractor-style).

G — 325 Qs	87.1% (283/325)
P — 197 Qs	79.7% (157/197)
R — 114 Qs	86.0% (98/114)
J — 93 Qs	87.1% (81/93)
T — 90 Qs	87.8% (79/90)

Figure 14: **LongSocialBench portal: independent verification view.** A different contributor re-answers the item from the model-facing input only: serialized thread, reply structure, question stem, and four answer options.



Phase 3 — Double-check (resolved disputes)

When the first two reviewers disagreed, a third invited contributor (different from both author and verifier) made the final call: **(✓) confirm the original answer**, **(✎) clarify the wording**, or **(↻) rewrite the question**. Only the 67 questions below needed this pass; everything else passed through on first agreement. Each row is stored in `phase7_decisions` with an `adjudicator_user_id`.

QUESTION	DOMAIN	AUTHOR	VERIFIER B	ADJUDICATOR	VERDICT	CHANGE SUMMARY
`hn_46454977::hn_46454977_q12` C3 · P	hackernews	reviewer-04	reviewer-05 chose C	reviewer-02	✎ clarified	Added concept-level disambiguation clause to criterion_definition: (c) Named-individual scope — a reply counts if it provides a specific named person (by name or unambiguous pronoun reference to a previously named person in the thread) as the answer to the question posed. Replies that reference 'he' where the antecedent is clearly Buffett (244, 246) count because the pronoun unambiguously resolves to a named individual in context.
`stackexchange_philosophy::q4_v6` C3 · G	stackexchange	reviewer-03	reviewer-07 chose C	reviewer-06	✓ confirmed	Solver misclassified a_16 as pure-discover despite a_16 explicitly refusing the Platonist label ("No"), and missed pure-invent members a_8310 and a_6430. Criterion requires overtly entire stance; solver's readings contradict the overt text.
`hn_46454977::hn_46454977_q08` C3 · G	hackernews	reviewer-04	reviewer-07 chose A	reviewer-02	✎ clarified	Added concept-level disambiguation clause to criterion_definition: (c) Organizational scope — a comment counts as diagnosing organizational/incentive failure if its body-majority thrust identifies a specific internal structural problem (management incentives, dealer network design, institutional inertia) rather than describing an external constraint or narrating a historical event. Comments that describe a company's restructuring attempt (80) count because they diagnose the organizational problem that necessitated the attempt. Comments about dealer gatekeeping (445) count because dealer incentive misalignment is an organizational design problem.
`stackexchange_philosophy::q7_v6` C3 · R	stackexchange	reviewer-03	reviewer-02 chose C	reviewer-04	✓ confirmed	Solver batch-claimed "12 of 13" accepted answers open with category-split, which collapses distinct gambits (definition-first, historical-landmark, objection-raising, hedged-agree-extend, inverse-reformulation, stance-with-survey) into one bucket. Gold's per-atom classification (definition-first = 3 plurality) is supported by the opening-sentence criterion; solver's shortcut is the failure.
`hn_46454977::hn_46454977_q17` C3 · G	hackernews	reviewer-04	reviewer-08 chose A	reviewer-05	✎ clarified	Added concept-level disambiguation clause to criterion_definition: (c) Geographic specificity — a reply counts as firsthand geographic observation if it names any identifiable geographic region (country, city, or supranational region like 'outside of the US') AND describes personal observation. 'Outside of the US' (109) counts because it identifies a geographic scope and reports firsthand observation of competitive dynamics.

▸Show all 67 adjudicated cases

Per-question records

Figure 15: **LongSocialBench portal: adjudication view.** Phase-3 adjudication records the verdict, release action, and change summary for Phase-2 mismatches or author-flagged borderline cases.

contributors completed a calibration step covering the six capability surfaces, evidence-node selection, option construction, and common shortcut risks. Human-reference readers used for the final

answerability comparison were kept separate from item authors, verifiers, and adjudicators.

Contributors received written guidelines requiring each item to specify: (i) the relevant discussion scope, such as a comment, direct-reply set, subtree, chain, participant trajectory, or cross-branch comparison; (ii) one of the six capability surfaces; (iii) a four-option question with exactly one supported answer; (iv) a plain-language decision criterion; and (v) supporting evidence nodes. The guidelines instructed authors to avoid answer-revealing locators, exact quotes, usernames, distinctive phrases, and distractors that could be eliminated by wording alone. They also required distractors to reflect plausible wrong readings, such as wrong scope, wrong criterion, boundary errors, lexical matches, or aggregation mistakes.

Intended use and dual-use considerations. LongSocialBench is intended for the evaluation of long-context language models on reply-structured discussion understanding. The capability surfaces (e.g., reply role, turning point, participant trajectory) are scoped to behavior observable in the released thread; they are not designed for identity inference, off-platform attribution, or individual-level profiling of participants outside the released episodes. Released items, evidence references, and prompts are distributed for research use only, and we discourage applications that target individual users in deployed settings.

J Release and Reproducibility

We release LongSocialBench v1.0 to support reproducible evaluation.

Released artifacts. The release includes: (i) the $1{,}462$ verified items with question stem, four options, gold option, scope locator, and capability label; (ii) the 94 serialized episodes with reply-structure metadata used by the items; (iii) author-side evidence-node references used during verification, distributed for diagnostic purposes only and excluded from model-facing input; (iv) the workflow construction and prompt templates corresponding to the 29 evaluation settings reported in the paper; and (v) per-cell raw accuracy records over the 18×29 model-by-workflow grid sufficient to reproduce the headline and capability tables.

Evaluation reproducibility. All model evaluations were run under the workflow settings documented in Appendix E.2, using publicly accessible model endpoints or open-weight checkpoints. The evaluation harness, scoring scripts, and aggregation utilities used to produce the reported tables and figures are released alongside the benchmark. Decoding parameters, prompt versions, and random seeds where applicable are recorded per setting in the released configuration files.

Versioning and updates. The release snapshot is identified as LongSocialBench v1.0. Subsequent revisions follow semantic versioning: PATCH for non-content fixes (wording, formatting), MINOR for additive items or capability annotations, and MAJOR for changes that affect the gold option of any released item. A changelog accompanies each released version so that downstream comparisons remain anchored to a specific snapshot.