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

Xinyi Liu
University of Illinois
Urbana-Champaign
liu323@illinois.edu

Rinat Khaziev
Semantic Code Lab
rinat@semantic.sh

Dilek Hakkani-Tür
University of Illinois
Urbana-Champaign
dilek@illinois.edu

Tarek F. Abdelzaher
University of Illinois
Urbana-Champaign
zaher@illinois.edu

Abstract

Long-context LLMs can now ingest entire online discussion threads, but understanding their social discourse requires more than reading a long document: models must track parent–reply relations, turning points, scoped subtrees, cross-branch contrasts, and participant trajectories. To test this structure-aware social reasoning, we introduce **LongSocialBench**, a benchmark of 1,462 verified human-authored multiple-choice items drawn from 94 complete Hacker News, Stack Exchange, and Reddit r/ChangeMyView episodes, with a median length of ≈73K tokens. Each item pairs a complete serialized discussion and reply structure with a four-option question, requiring models to recover structured social evidence. Released items are verified for answerability, option uniqueness, and evidence grounding.

Across 18 models and 29 evaluation settings, current long-context workflows remain far below human performance. The best individual result comes from Claude-Opus-4.7, which reaches 63.0% when prompted to eliminate incorrect options before answering, compared with 72.4% for independent human readers. Averaged across all 18 models, the full-context Baseline scores 43.9%. Supplying the gold evidence scope raises this to 55.0%, showing that substantial errors remain even after the relevant thread region is identified. 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 a discussion therefore requires more than summarizing a long document. Relevant evidence is distributed across reply links, participant identities, branch boundaries, and temporal turning points. A model must identify the relevant part of the thread, interpret how comments relate, and integrate evidence across the episode. 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 such discussions as structured social interaction. 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?*

One LongSocialBench item illustrates the challenge. In a r/ChangeMyView discussion about middle-class alienation, the item asks where the episode's unique longest reply chain first shifts from a class-and-privilege debate to a corporate wage-distribution argument. Repeated words such as *class*, *privilege*, or *wages* appear throughout the discussion and do not locate this transition. The model must isolate the specified chain and identify where the exchange shifts from housing and wealth-hoarding claims to concrete claims about company revenue, GDP per worker, and worker pay. The item therefore requires both structural localization and reasoning over how the exchange develops across turns.

Existing benchmarks cover parts of this setting, but not this combination. Long-context benchmarks evaluate large inputs organized as docu-

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 20638–20660
October 24-29, 2026 ©2026 Association for Computational Linguistics

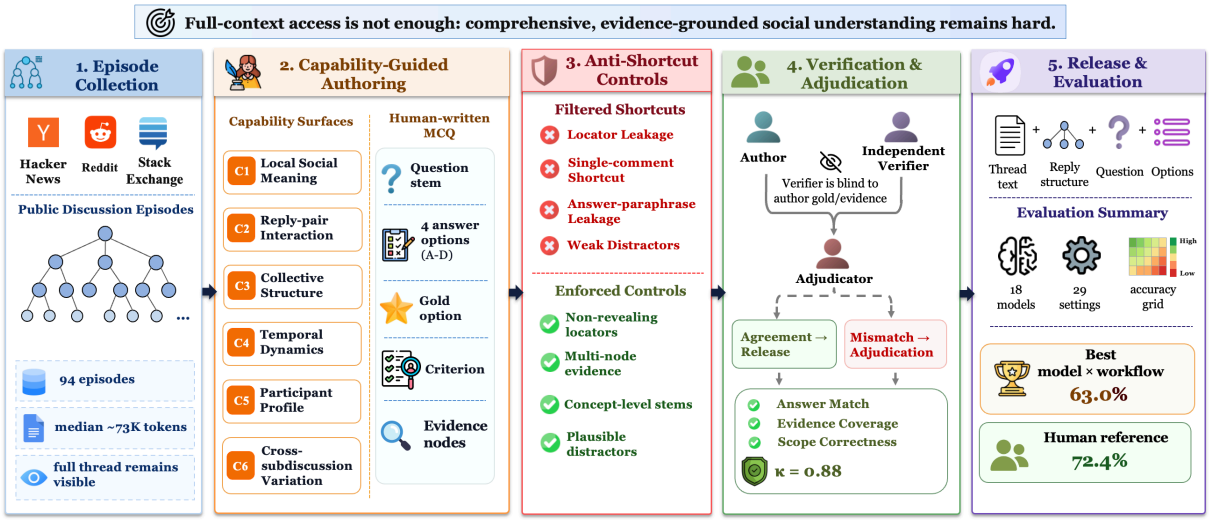

Figure 1: **LongSocialBench construction and verification pipeline (§3.1).** We sample long public discussions, author items guided by capability definitions, apply anti-shortcut controls, independently verify answers and evidence, and adjudicate residual disputes.

ments, 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). 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 evaluate information retention and reasoning over extended dialogue histories, but typically within linear dialogue rather than branching, multi-party reply structures (Maharana et al., 2024; Wu et al., 2025). Online-discussion resources capture real threaded interaction, but typically formulate it as stance classification, persuasion prediction, or summarization rather than verifiable questions that require locating and reasoning over evidence in a complete reply tree (Abbott et al., 2016; Tan et al., 2016; Fabbri et al., 2021; Niu et al., 2024). What remains missing is evidence-grounded social reasoning over a full reply tree, where a model must recover the relevant branch, chain, participant trajectory, or comparison set and determine which social interpretation is supported.

LongSocialBench addresses this gap with 1,462 human-authored multiple-choice items drawn from 94 Hacker News, Stack Exchange, and Reddit r/ChangeMyView discussions, with a median serialized episode length of ≈73K tokens. We use complete public discussions that fit within the evaluation context limits and span complementary forum formats and topic ranges. Hacker News contributes many technical and policy threads, Reddit r/ChangeMyView contributes social and policy debates, and Stack Exchange contributes expert-domain Q&A. To define capability coverage, we draw on prior work on social inference, stance, persuasion, conversational stance, and discussion summarization (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 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. For each item, the human author specifies a decision rule for determining the correct answer from the thread and provides node-level supporting evidence. Figure 1 summarizes the construction and verification process.

Our evaluation reveals a substantial gap between full-context access and reliable social understanding. We evaluate 18 models across 29 settings: 22 deployable workflows that use no gold annotations, spanning full-context, reasoning, and retrieval methods, and seven diagnostic settings probing benchmark validity and evidence scope. Across the deployable workflows, model-average accuracy ranges from 37.2% to 44.7%, far below the 72.4% mean of five independent human readers given the same full discussions. The strongest individual model-workflow combination, Claude-Opus-4.7 prompted to eliminate incorrect options before

answering, reaches 63.0%, still 9.4 pp below the human mean. As a diagnostic, we also give models the gold evidence scope to separate evidence localization from interpretation. This raises the average across all 18 models from 43.9% under the full-context Baseline to 55.0%, leaving substantial error even when the relevant thread region is identified. These results indicate difficulties in both locating relevant social evidence and interpreting it once found.

Our contributions are:

- **A human-authored benchmark for long-episode social reasoning.** LongSocialBench contains carefully verified items written over complete real discussions and organized by six capability surfaces spanning local interaction, collective structure, participant trajectories, temporal dynamics, and cross-subdiscussion variation.
- **A broad evaluation of long-context workflows.** We evaluate 18 models across 29 settings, spanning deployable full-context, reasoning, and retrieval workflows together with validity and scope diagnostics to assess robustness across methods.
- **A diagnosis of failure sources.** We use gold-scope diagnostics to separate errors in locating relevant evidence from errors in interpreting it correctly, showing whether failures arise before or after the relevant evidence is identified.

Together, LongSocialBench provides a reliable and diagnostic testbed for long-context models that must understand real online discussions as structured social interaction.

2 Related Work

LongSocialBench sits at the intersection of long-context evaluation and social-discourse understanding. Prior work covers several parts of this setting, including long-input reasoning, conversational memory, localized social inference, and real online discussions, but rarely combines long contexts, reply-tree structure, social interpretation, and verifiable judgments over complete episodes. Table 1 summarizes this gap.

Table 1: **Area-level comparison of related evaluation settings** across long context, reply-tree structure, social interpretation, and verifiable evaluation.

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 is central to the task; Social = social interpretation; Verif. = item-level gold-answer verification.

2.1 Long-context evaluation and conversational memory.

Long-context benchmarks test retrieval, aggregation, multi-hop reasoning, code, structured data, and synthetic context use over large 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). Long-conversation benchmarks such as LoCoMo and LongMemEval instead evaluate memory and reasoning over extended dialogue histories (Maharana et al., 2024; Wu et al., 2025). Together, these benchmarks establish the difficulty of using information distributed across long contexts. Their inputs, however, are typically documents, structured sequences, or linear dialogue histories rather than branching multi-party reply trees. LongSocialBench adds the requirement to recover the relevant interaction structure and reason about how participants and positions develop within it.

2.2 Social reasoning and online discussions.

SocialIQA (Sap et al., 2019), MuTual (Cui et al., 2020), ToMi (Le et al., 2019), and ToMBench (Chen et al., 2024) evaluate social commonsense, dialogue, or theory-of-mind reasoning, but generally use short or localized interactions. A complementary line studies real online discussions. IAC 2.0 (Abbott et al., 2016), ChangeMyView persuasion studies (Tan et al., 2016), RumourEval (Gorrell et al., 2019), and MT-CSD (Niu et al., 2024) annotate stance, persuasion, or related interaction outcomes, while ConvoSumm (Fabbri et al., 2021) targets discussion summarization. These resources motivate our capability design, including reply relations, collective structure, participant trajectories, and changes across subdiscussions. LongSocialBench turns these forms of social evidence into verified questions over complete episodes, requiring models to locate the relevant thread scope and reason over the evidence within it.

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 makes whole-episode social comprehension measurable through observable interaction. Throughout, we use *social* to describe how participants respond to one another, what stances and roles they take, and how these patterns persist or change across replies, participant trajectories, and sub-discussions. Items are grounded in observable thread behavior and do not require inferences about participants' private mental states.

Drawing on prior work on social inference, stance, persuasion, conversational stance, and discussion summarization (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 recurring interaction patterns into six capability surfaces, summarized in Table 2. Together, these surfaces provide systematic coverage by linking each item to a specific evidence pattern and social-reading decision. Appendix B.1, Table 7 provides full decision descriptions, and Appendix F gives a worked item for each surface.

Table 2: **Overview of the six capability surfaces.** Each surface captures a recurring evidence pattern in reply-structured discussions.

Capability surface	Evidence focus
C1: Local social meaning	A comment's social role or meaning in its local context.
C2: Reply-pair interaction	The relation between a reply and the parent comment it addresses.
C3: Collective structure	How positions or roles are distributed across a discussion region.
C4: Temporal dynamics	How an exchange changes over time, including shifts and turning points.
C5: Participant profile	A participant's stance or interaction pattern across multiple comments.
C6: Cross-subdiscussion variation	How social patterns differ or remain consistent across branches or sub-discussions.

3.2 Data Collection

Episode collection and coverage. We collect 94 complete public discussion episodes from Hacker News, Reddit r/ChangeMyView, and Stack Exchange. The sources provide complementary discussion settings: Hacker News contributes many computing, AI, economy, and policy discussions, Reddit r/ChangeMyView contributes social and policy debates, and Stack Exchange contributes expert-domain Q&A across a broad range of fields (Appendix A, Table 5). The released set contains ≈6.63M serialized tokens, with a median episode length of ≈73K tokens and a maximum of 119K. Per-source length and structural statistics are reported in Appendix A, Table 6.

Item design and safeguards. Twenty invited contributors authored items under written guidelines designed to ensure systematic capability coverage and reliable evaluation. For each item, the author selected one of the six capability surfaces and wrote the question stem, four options, gold option, item-specific decision criterion, and supporting evidence nodes grounded in the episode. Appendix H.1 summarizes the authoring protocol, contributor training, and role separation.

To make each answer independently verifiable, a candidate item is released only if its gold option is recoverable from the episode alone, uniquely supported under the decision criterion, and grounded in cited evidence. The evidence must also establish any structural premises needed to answer the question, such as the relevant scope, reply relation, counting unit, participant trajectory, or comparison set. Appendix B.2, Table 8, details the release criteria.

To prevent shortcut solving, the guidelines control wording, option balance, and distractor design. Stems avoid answer-revealing locators, options use parallel wording, and distractors reflect plausible errors such as selecting the wrong scope, misreading boundaries, relying on lexical overlap, or aggregating evidence incorrectly. To reduce external answer leakage, authors also search distinctive stem and option phrases and named events on the open web, following contamination concerns in long-context benchmark construction such as LooGLE (Li et al., 2024). Items whose answers can be recovered without the supplied episode are rewritten or excluded. Appendix B.3, Table 9, details these controls.

3.3 Verification and Release

Each candidate item goes through a three-phase verification process before release. A second contributor independently answers the item, while disagreements and author-flagged items requiring additional review are examined by a third contributor. Appendix G documents the annotation workflow and role separation.

Phase 1 is authoring: a contributor writes the item, decision criterion, gold option, and evidence nodes. **Phase 2** is independent verification: a different trained contributor sees only the model-facing input, answers the question from scratch, and checks whether the authored gold option is supported. **Phase 3** is adjudication: a third reviewer examines verifier disagreements and author-flagged items requiring additional review, re-checking the criterion, evidence, and option uniqueness. The reviewer then confirms the item, requires revision followed by a new Phase-2 verification pass, or removes it.

Among 1,553 parseable Phase-2 candidate records, 1,419 match the author gold option on the first pass, yielding 91.37% first-pass agreement before adjudication. Of the 134 mismatches, 43 are retained after adjudication and 91 unresolved cases are removed. Phase 3 also reviews 24 author-flagged items among the 1,419 first-pass matches. The final benchmark contains 1,462 items. A stratified audit of 100 released items finds 98 passes and two items requiring minor wording revisions, with no residual gold-option errors (Appendix C.2, Table 10). Additional release details, including public-data licensing, contributor protocol, and released artifacts, are documented in Appendix H and Appendix I.

4 Experimental Setup

Models. We evaluate 18 contemporary LLMs spanning closed-source frontier and strong open-weight systems, including multiple versions within several model families. Closed-source models include Claude Opus 4.7, Claude Sonnet 4.6, 4.5, and 4, GPT-5, and Gemini 2.5 Pro. Open-weight models include Mistral Large 3, Llama 4 Scout and Maverick, Llama 3.1 and 3.2, Qwen3-Next and Qwen3-32B, GLM-5, DeepSeek-R1, Gemma 3, and OpenAI's gpt-oss models. This selection provides broad coverage of frontier, open-weight, and reasoning-oriented systems. Appendix D.1 reports the full model list and context-window limits, while Appendix D.2 details workflow construction and prompt instructions.

Workflows. We evaluate 29 settings: 22 deployable workflows that use no gold annotations and 7 diagnostic settings.

Reasoning prompts. These settings use the same complete discussion, question, and answer options as the full-context Baseline, but add an instruction for how to reason before selecting the final answer. Standard chain-of-thought (CoT) prompting asks for free-form step-by-step reasoning. Evidence enumeration first identifies the relevant evidence, temporal tracking reconstructs how the exchange develops across turns, and option elimination compares the choices and rules out unsupported ones. Self-reflection asks the model to re-check its reasoning before committing to the final answer.

Retrieval workflows. These methods first select or organize a smaller portion of the discussion for answering. BM25 and E5 retrieve comments by lexical and semantic relevance, respectively. Anchor-note RAG supplements retrieved regions with brief model-generated notes, subtree retrieval selects connected reply subtrees while preserving their structure, and LLM-based retrieval first asks a model to select relevant comments or regions.

Diagnostics. The seven diagnostic settings alter the information available to the model to test potential shortcuts, answer-format effects, and evidence localization. Five validity probes isolate specific cues: options-only provides only the four answer choices, question-only provides only the question stem, title-and-root-only provides the discussion title and root post without replies, shuffled-thread keeps all comments but shuffles their order, and open-answer keeps the thread and question but removes the answer choices and requires a short free-text response. Two scope diagnostics test whether errors come from failing to find the relevant evidence: gold-scope provides the author-annotated evidence scope, while infer-scope provides a scope predicted by a separate prompt.

Table 12 in Appendix D.2 gives the exact settings, input construction, and prompt instructions.

Metrics. For multiple-choice settings, responses are deterministically parsed to an A–D choice and scored by exact match. Accuracy is computed over successfully parsed responses only, excluding outputs for which no unique answer choice can

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	**74.5**	65.2	49.1	**67.0**	54.1	50.0	**60.0**	**79.4** $^{+4.8}$	71.9 $^{+6.7}$	53.1 $^{+4.1}$	**70.8** $^{+3.8}$	**65.6** $^{+11.5}$	**59.3** $^{+9.3}$	**66.7** $^{+6.7}$
Gemini-2.5-Pro	65.5	58.4	**52.1**	49.7	45.9	36.4	51.3	71.5 $^{+6.1}$	69.7 $^{+11.2}$	**55.9** $^{+3.8}$	56.8 $^{+7.0}$	56.9 $^{+11.0}$	56.8 $^{+20.3}$	61.2 $^{+9.9}$
GPT-5	61.8	68.5	43.7	58.9	**63.8**	**53.4**	58.3	67.9 $^{+6.1}$	**80.9** $^{+12.4}$	44.3 $^{+0.6}$	49.7 $^{-9.2}$	63.3 $^{-0.5}$	58.5 $^{+5.1}$	60.8 $^{+2.4}$
Claude-Sonnet-4.5	73.3	59.6	46.6	49.7	50.5	46.6	54.4	71.5 $^{-1.8}$	67.4 $^{+7.9}$	46.1 $^{-0.4}$	53.0 $^{+3.2}$	56.0 $^{+5.5}$	48.3 $^{+1.7}$	57.1 $^{+2.7}$
Claude-4	72.7	**71.9**	45.0	54.6	50.5	47.5	57.0	72.1 $^{-0.6}$	71.9 $^{+0.0}$	46.9 $^{+1.9}$	51.9 $^{-2.7}$	50.5 $^{+0.0}$	48.3 $^{+0.8}$	56.9 $^{-0.1}$
Claude-Sonnet-4.6	56.4	49.4	41.5	43.2	47.2	36.4	45.7	65.5 $^{+9.1}$	58.4 $^{+9.0}$	47.6 $^{+6.1}$	51.4 $^{+8.1}$	53.7 $^{+6.4}$	47.5 $^{+11.0}$	54.0 $^{+8.3}$
(Strong open-weight)														
GLM-5	54.5	36.0	41.0	35.1	41.3	39.8	41.3	72.1 $^{+17.6}$	64.0 $^{+28.1}$	47.0 $^{+6.0}$	52.4 $^{+17.3}$	59.2 $^{+17.9}$	45.8 $^{+5.9}$	56.8 $^{+15.5}$
DeepSeek-R1	64.8	60.7	42.5	51.4	45.0	39.0	50.6	65.5 $^{+0.6}$	61.8 $^{+1.1}$	42.4 $^{-0.1}$	46.5 $^{-4.9}$	46.8 $^{+1.8}$	39.8 $^{+0.8}$	50.5 $^{-0.1}$
Qwen3-Next-80B	56.4	49.4	40.2	44.9	38.5	38.1	44.6	64.2 $^{+7.9}$	52.8 $^{+3.4}$	37.4 $^{-2.8}$	42.7 $^{-2.2}$	42.7 $^{+4.1}$	37.3 $^{-0.8}$	46.2 $^{+1.6}$
Mistral-Large-3	55.8	43.8	40.2	42.7	36.7	40.7	43.3	64.8 $^{+9.1}$	55.1 $^{+11.2}$	37.8 $^{-2.3}$	34.6 $^{-8.1}$	45.9 $^{+9.2}$	38.1 $^{-2.5}$	46.1 $^{+2.8}$
(Mid open-weight)														
Llama-4-Mav-17B	55.8	42.7	40.9	37.8	37.6	33.9	41.5	58.8 $^{+3.0}$	62.9 $^{+20.2}$	41.0 $^{+0.1}$	38.9 $^{+1.1}$	44.5 $^{+6.9}$	30.5 $^{-3.4}$	46.1 $^{+4.7}$
Llama-4-Scout-17B	55.2	47.2	38.7	33.5	39.9	33.1	41.3	53.9 $^{-1.2}$	49.4 $^{+2.2}$	39.7 $^{+1.0}$	30.3 $^{-3.2}$	39.0 $^{-0.9}$	32.2 $^{-0.8}$	40.8 $^{-0.5}$
Llama-3.1-70B	53.9	41.6	36.1	35.1	33.9	36.4	39.5	54.5 $^{+0.6}$	52.8 $^{+11.2}$	33.9 $^{-2.2}$	36.2 $^{+1.1}$	39.9 $^{+6.0}$	40.7 $^{+4.2}$	43.0 $^{+3.5}$
Qwen3-32B	60.6	51.7	38.9	37.8	41.7	32.2	43.8	51.5 $^{-9.1}$	39.3 $^{-12.4}$	26.8 $^{-12.1}$	31.9 $^{-5.9}$	28.4 $^{-13.3}$	29.7 $^{-2.5}$	34.6 $^{-9.2}$
(Small / mid-weak)														
GPT-OSS-120B	58.8	48.3	36.2	34.1	36.7	38.1	42.0	54.5 $^{-4.2}$	49.4 $^{+1.1}$	33.2 $^{-3.1}$	33.0 $^{-1.1}$	37.2 $^{+0.5}$	32.2 $^{-5.9}$	39.9 $^{-2.1}$
GPT-OSS-20B	49.7	46.1	31.6	36.8	38.1	33.1	39.2	49.1 $^{-0.6}$	48.3 $^{+2.2}$	31.7 $^{+0.1}$	34.1 $^{-2.7}$	37.6 $^{-0.5}$	33.1 $^{+0.0}$	39.0 $^{-0.2}$
Llama-3.2-90B	50.3	32.6	35.8	28.6	33.9	30.5	35.3	46.7 $^{-3.6}$	32.6 $^{+0.0}$	31.3 $^{-4.5}$	25.4 $^{-3.2}$	42.2 $^{+8.3}$	31.4 $^{+0.8}$	34.9 $^{-0.4}$
Gemma-3-27B	57.0	40.4	33.9	33.5	33.9	29.7	38.1	47.3 $^{-9.7}$	44.9 $^{+4.5}$	25.9 $^{-8.0}$	31.4 $^{-2.2}$	35.3 $^{+1.4}$	31.4 $^{+1.7}$	36.0 $^{-2.0}$
Mean (18)	59.8	50.7	40.8	43.0	42.7	38.6	46.0	61.7 $^{+1.9}$	57.4 $^{+6.7}$	40.1 $^{-0.7}$	42.8 $^{-0.2}$	46.9 $^{+4.2}$	41.1 $^{+2.5}$	48.4 $^{+2.4}$

be recovered. Benchmark-level results report micro accuracy, while capability summaries report the unweighted macro mean over C1–C6. For the open-answer validity probe, free-text responses are content-scored against the gold option.

5 Results

We first establish the overall performance gap, then examine how it varies across capabilities and workflows, and finally use diagnostic analyses to identify where the remaining errors arise.

5.1 Current models remain well below human performance

Table 3 shows that even the strongest models remain limited on complete-thread social reasoning. Claude-Opus-4.7 performs best under the full-context Baseline, with a C1–C6 macro average of 60.0%. Standard chain-of-thought prompting raises its average to 66.7%, while the C1–C6 macro average across all 18 models increases modestly from 46.0% to 48.4%. For reference, five independent human readers average 72.4% when answering from the complete discussions.

Overall, the strongest models perform better, while chain-of-thought provides only a modest aggregate gain and does not close the gap to human readers.

5.2 Failures emerge as reasoning spans the thread

Table 3 reveals a clear progression in difficulty. Under the full-context Baseline, average accuracy is 59.8% on **C1 local social meaning** and drops to 50.7% on **C2 reply-pair interaction**. For **C3–C6**, which require reasoning over larger discussion structures, accuracy falls further to 38.6%–43.0%. The sharpest failures appear on **C6 cross-subdiscussion variation** at 38.6% and **C3 collective structure** at 40.8%, where models must compare behavior across branches or recover the positions organizing a discussion region. The main difficulty therefore emerges when evidence must be integrated across the thread into a higher-level social judgment.

Standard chain-of-thought prompting does not remove this weakness. It improves **C2** by 6.7 pp, but leaves **C3 collective structure** and **C4 temporal dynamics** essentially unchanged, from 40.8% to 40.1% and from 43.0% to 42.8%. Cross-

subdiscussion reasoning also remains low at 41.1%. Thus, additional reasoning helps some local interaction decisions but does little to resolve the capabilities that require organizing evidence across larger thread structures, a pattern that also appears across individual models and platforms in Appendix E.3, Figs. 5–7.

5.3 Reasoning benefits are strongly model-dependent

Chain-of-thought prompting produces substantial gains for some models, but its effects vary considerably across the model pool. Gemini-2.5-Pro and Claude-Sonnet-4.6 improve by 9.9 and 8.3 pp, respectively, while several other models show little improvement or decline under CoT.

This heterogeneity limits the aggregate benefit of additional reasoning. Across all 18 models, CoT raises the C1–C6 macro average by only 2.4 pp, despite much larger gains for several individual models. The results therefore show that the benefit of additional reasoning depends strongly on the model. Appendix E.4, Table 17 and Fig. 8, show that this model dependence extends beyond standard CoT to evidence enumeration, temporal tracking, option elimination, and self-reflection.

5.4 Flat retrieval loses both evidence and thread structure

Figure 2 shows that standard comment-level retrieval is often counterproductive. Both lexical BM25 and dense E5 retrieval reduce accuracy on several capabilities, with the largest losses on **C3 collective structure**. Methods that preserve reply structure or let an LLM select coherent regions perform better, although they still provide only limited gains over the full-context Baseline. The pattern suggests that relevance at the level of individual comments is a poor match for questions whose evidence is distributed across connected parts of the thread.

The evidence-coverage diagnostic reveals a complementary failure. Standard comment-level retrievers recover only a small fraction of the author-cited evidence, with high-coverage retrieval remaining rare even as more comments are retrieved. Thus, flat retrieval can fail before the final reasoning step by omitting necessary evidence, and it can further weaken the remaining evidence by separating comments from the reply structure in which they are meaningful. Appendix E.5, Table 18, quantifies this gap through evidence recall at different retrieval budgets and the frequency of high-coverage retrieval.

5.5 Validity checks support genuine thread-level difficulty

We first test whether simpler cues can substitute for the full discussion. The options-only, question-only, and title-and-root-only probes retain only the answer choices, the question stem, or the discussion title and root post, respectively. Their accuracies remain limited at 31.5%, 38.0%, and 35.4%. Retaining all comments but shuffling their order yields 40.4%. Together, these probes show that answer-choice cues, question wording, and topic-level information are insufficient, while the organization of the discussion also contributes to successful reasoning. Appendix E.6, Table 19, gives the full shortcut analysis.

We next examine whether the multiple-choice format drives the model comparison. In the open-answer probe, models receive the same discussion and question without the answer choices and instead produce a short free-text response. Accuracy drops by 7.6 pp on the same items, but the model ranking is preserved almost exactly, with Spearman ρ=1.00. Answer choices therefore make the task easier in absolute terms without materially changing which models perform better.

Finally, we test whether the remaining difficulty is primarily explained by contamination or input length. A time-disjoint comparison finds no systematic advantage for the older Stack Exchange cohort over the 2026 Hacker News cohort across model tiers, providing no indication that source-page memorization is a major driver. Separately, episode length has only a weak negative association with Baseline accuracy, with Spearman ρ=−0.18. Longer discussions are somewhat harder, but length accounts for little of the observed performance variation. Neither analysis supports memorization or episode length as a primary explanation for the benchmark difficulty. Appendix E.6.2, Table 20, reports the time-disjoint comparison, while Appendix E.6.3, Fig. 9, shows the episode-level length relationship.

5.6 Failures arise in both evidence localization and criterion application

We next ask where the remaining errors arise. Providing the gold evidence scope improves every model tier by roughly 10–12 pp in Table 4, raising the overall average from 43.9% to 55.0%. Evi-

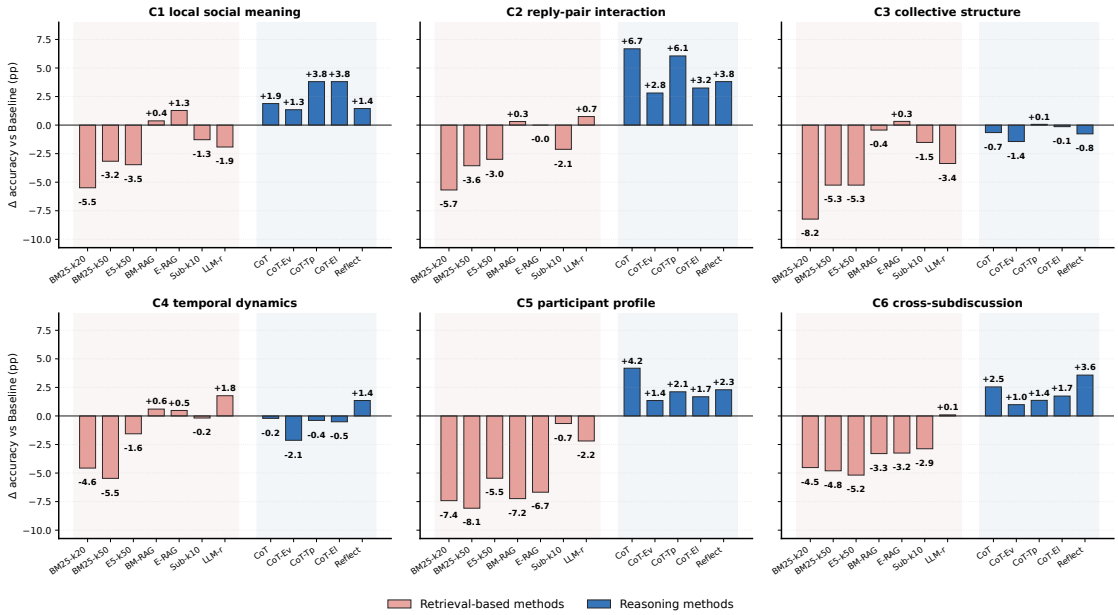

Figure 2: **Method effect by capability.** Mean Δ accuracy relative to Baseline across 18 models for C1–C6. Reasoning methods are generally beneficial, while flat retrieval often reduces accuracy, especially on C3. Structure-aware and LLM-based retrieval recover part of these losses.

Table 4: **Gold evidence scope improves accuracy across all model tiers, but substantial error remains.** Mean accuracy across models in each tier on all 1,462 items.

Model tier	Full	Gold-Scope	Δ
Closed frontier (6)	51.4	62.3	+10.9
Strong open (4)	43.4	53.9	+10.5
Mid open (4)	40.3	51.2	+10.9
Small (4)	37.0	48.9	+11.9
All 18	43.9	55.0	**+11.1**

dence localization is therefore a substantial source of error, yet substantial difficulty remains even when the relevant thread region is known.

To understand what remains after evidence localization, we run a separate component oracle on Claude-Opus-4.7 using a stratified subset of 510 items from 30 episodes. The CoT baseline reaches 63.1%. Gold scope raises accuracy to 66.7%, while gold per-comment criterion labels raise it to 81.4%. Providing both raises accuracy to 98.6%. On this subset, the much larger improvement from criterion labels shows that knowing where to look is not enough: models must also determine correctly how each comment should be treated under the question-specific criterion. Appendix E.7, Table 21, gives the full component comparison, including the model-generated enumerate-label-aggregate pipeline.

A manual audit of 50 Claude-Opus-4.7 Gold-Scope errors supports the same diagnosis. In 60% of the audited errors, the model sees the relevant thread region but applies the question-specific criterion too broadly, too narrowly, or otherwise incorrectly. The remaining errors involve thread structure, counting rules, or surface cues, with the full breakdown reported in Appendix E.7.1.

Together, these results identify two distinct sources of failure. Models must first locate the relevant part of the discussion and then apply the correct criterion to the evidence within it. Current reasoning and retrieval methods can improve parts of this process, but reliable performance requires both stages to succeed.

6 Conclusion

We introduced LongSocialBench, a benchmark of 1,462 verified human-authored items over 94 complete reply-structured discussions. Across 18 models and 29 evaluation settings, current long-context workflows remain substantially below human readers, while additional reasoning and retrieval provide only partial and model-dependent improvements. Our diagnostics identify two distinct sources of error: models must first locate the relevant part of the discussion and then correctly apply the question-specific social criterion to the evidence within it. Even when the relevant region

is supplied, criterion application remains a major source of failure.

These findings frame complete-thread social understanding as a structured reading problem over reply trees. LongSocialBench provides a diagnostic testbed for developing models that can reliably reason across reply chains, participant trajectories, and cross-branch comparisons.

Limitations

Scope of discussion data. LongSocialBench focuses on English, text-based, public reply-threaded discussions from Hacker News, Reddit r/ChangeMyView, and Stack Exchange. These sources provide complete discussions with explicit reply structure, but they do not represent private or synchronous conversations, multimodal interaction, non-English discourse, or platforms with different interaction norms. Future work should examine whether the capability surfaces and failure patterns identified here generalize to these broader conversational settings.

Benchmark format. LongSocialBench uses multiple-choice questions to support verifiable and reproducible evaluation of long-thread social reasoning. Our open-answer probe shows that removing the answer choices lowers absolute accuracy while largely preserving the model ranking, suggesting that the main comparison is not specific to the multiple-choice format. The benchmark nevertheless focuses on evidence-grounded social judgments and does not directly evaluate free-form explanation quality, calibrated uncertainty, or faithful long-thread summarization. Future work can extend the benchmark with richer response formats and corresponding evaluation protocols.

Ethical Considerations

LongSocialBench is constructed from public discussion data from Hacker News, Reddit r/ChangeMyView, and Stack Exchange under the applicable platform and licensing terms. To reduce privacy risk, released discussions replace usernames with structural references and remove personally identifying spans surfaced during verification. Human contributors and readers follow a written annotation protocol and are compensated in accordance with institutional norms. The benchmark evaluates observable interaction within public discussions and is not intended for identity inference, off-platform attribution, or individual-level profiling. Further details on data handling, contributor procedures, and intended use are provided in Appendix H.

Acknowledgments

We thank the contributors and human readers who supported the construction and evaluation of LongSocialBench. This work was supported in part by the Defense Advanced Research Projects Agency (DARPA) under the Methodological Advancements for Generalizable Insights into Complex Systems (MAGICS) program, Agreement No. HR0011263E079.

References

Rob Abbott, Brian Ecker, Pranav Anand, and Marilyn Walker. 2016. 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. European Language Resources Association (ELRA).

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. 2024. 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. Association for Computational Linguistics.

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. 2025. 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. Association for Computational Linguistics.

Amanda Bertsch, Adithya Pratapa, Teruko Mitamura, Graham Neubig, and Matthew R. Gormley. 2025. Oolong: Evaluating long context reasoning and aggregation capabilities. *arXiv preprint arXiv: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. 2024. 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. Association for Computational Linguistics.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. 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. Association for Computational Linguistics.

Alexander R. Fabbri, Faiaz Rahman, Imad Rizvi, Borui Wang, Haoran Li, Yashar Mehdad, and Dragomir Radev. 2021. 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. Association for Computational Linguistics.

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. RULER: What's the real context size of your long-context language models? In *First Conference on Language Modeling*.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. 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. Association for Computational Linguistics.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2024. 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. Association for Computational Linguistics.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. 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. Association for Computational Linguistics.

Fuqiang Niu, Min Yang, Ang Li, Baoquan Zhang, Xiaojiang Peng, and Bowen Zhang. 2024. 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. ELRA and ICCL.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. 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. Association for Computational Linguistics.

Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. 2023. ZeroSCROLLS: A zero-shot benchmark for long text understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7977–7989, Singapore. Association for Computational Linguistics.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. 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.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. 2025. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations*.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. 2025. HELMET: How to evaluate long-context models effectively and thoroughly. In *International Conference on Learning Representations*.

Appendix overview. Appendix A characterizes the discussion corpus, while Appendices B and C detail item construction, safeguards, and verification. Appendix D specifies the evaluated models and workflows, and Appendix E provides the additional analyses supporting the main results. Appendix F gives representative questions across the six capability surfaces. Finally, Appendices G, H, and I document the construction protocol, ethical and licensing considerations, and release and reproducibility details.

A 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.

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**

A.1 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.

A.2 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. 3 visualizes the same coverage along serialized length, maximum depth, and platform-specific depth-frequency curves.

B Item Design and Construction Safeguards

This appendix expands the capability surfaces introduced in §3.1 and documents the release criteria and shortcut controls used during item construction, complementing the data-collection details in §3.2.

B.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.

B.2 Release Criteria

Table 8 summarizes the criteria that every candidate item must satisfy before release.

B.3 Shortcut and Contamination Controls

Table 9 summarizes the authoring controls used to reduce surface shortcuts, improve distractor quality, and screen public-source contamination.

C Verification and Release Details

This appendix section supports the verification and release claims in §3.3. It records the independent verification protocol and release audit.

C.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.

Among the 1,553 parseable Phase-2 candidate records, independent verification matches the author gold option in 1,419 cases, yielding a candidate-level first-pass agreement of 91.37% before adjudication and Cohen's κ=**0.88** against a four-option chance baseline (P_e=0.25). Of the 134 first-pass mismatches, 43 are resolved by adjudication and retained, while 91 unresolved cases are excluded. Phase 3 also reviews 24 author-flagged edge cases that had passed first-pass agreement.

Table 6: **Episode-level scale and structure by source.** Token counts use `cl100k_base` over the complete serialized discussion. Maximum depth is measured from the root, and root replies are comments at depth 1.

Source	# eps	Tokens med.	Tokens max	Tokens total	Mean max-depth	Mean nodes	Mean root replies
Hacker News	43	73,397	118,981	3.29M	12.9	815.7	109.9
Stack Exchange	21	75,982	85,184	1.59M	2.0	601.0	385.2
Reddit r/CMV	30	56,322	79,597	1.75M	8.9	485.5	258.8
Overall	**94**	**≈73K**	**118,981**	**≈6.63M**	**9.2**	**662.4**	**218.9**

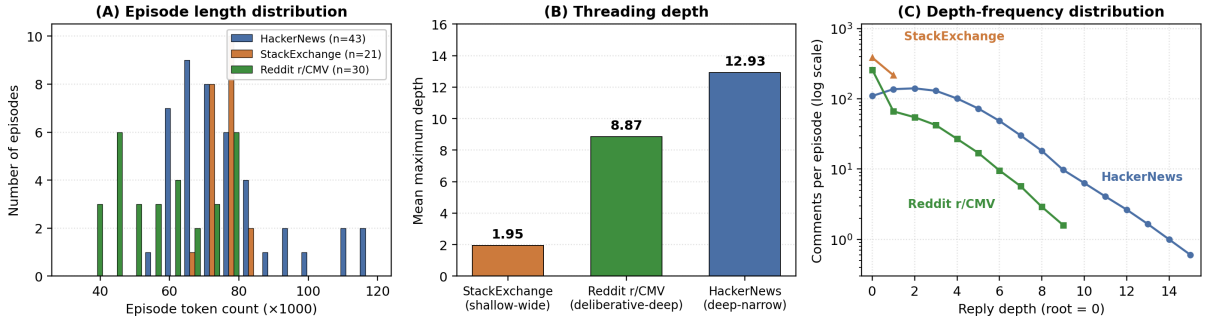

Figure 3: **Structural diversity of LongSocialBench episodes. (A)** Episode lengths span 37 to 119K tokens with platform-specific medians of 73K for Hacker News, 76K for Stack Exchange, and 56K for Reddit r/CMV. **(B)** Mean maximum depth differs substantially across sources: Stack Exchange 2.0, Reddit 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.

The 43 retained mismatch cases and 24 author-flagged cases yield 67 Phase-3 records carried forward in the release, with verdicts of 21 confirm, 40 clarify, and 6 rewrite.

C.2 Stratified Release Audit

The release audit estimates residual item-level errors after verification and adjudication. We sample 100 released items stratified by capability surface and platform, with allocation approximately proportional to the benchmark distribution and coverage of every non-empty capability–platform combination.

For each item, the audit checks that the cited evidence resolves, the criterion matches the intended scope, the stem and options are well formed, and no infrastructure or option-design artifacts reveal the answer. An adjudicator then re-reads the item from the model-facing input and assigns one of three verdicts. PASS requires no correction, CORRECTED

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.

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.

indicates a wording-level revision that leaves the gold option unchanged, and RESIDUAL indicates that the gold option cannot be recovered from the supplied input.

Table 10: **Stratified release audit.** Among 100 audited items, 98 require no correction, two require wording-level revisions without changing the gold option, and none contains a residual gold-answer error.

Domain	Pass	Corrected	Residual
Hacker News	44 / 46	2 / 46	0 / 46
Reddit r/CMV	32 / 32	0 / 32	0 / 32
Stack Exchange	22 / 22	0 / 22	0 / 22
All	**98 / 100**	**2 / 100**	**0 / 100**

No residual gold-option errors were found in any capability. The two CORRECTED items occurred in C2 and C3, one each.

D 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, evaluation context caps, workflow construction, and workflow-specific prompts.

Table 11: **Models used in the evaluation grid.** CW = evaluation context cap 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-Sonnet-4.6	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.1 Models

Table 11 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 evaluation context cap because full-context workflows require the serialized episode prompt to fit within the window configured for the run.

D.2 Workflow Settings and Prompts

Table 12 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. Deployable workflows and validity probes exclude author-side gold labels, criteria, evidence nodes, verifier notes, and adjudication records from prompting. Scope diagnostics additionally provide either gold or inferred scope information, as specified below. For multiple-choice settings, responses use the same deterministic answer parser and exact-match scoring rule; open-answer responses are content-scored separately. Per-response parse status is retained for reproducibility.

E Additional Evaluation Results

This appendix reports additional results in the same sequence as the main Results section: headline robustness and the human comparison, capability and platform profiles, reasoning prompts, retrieval diagnostics, validity checks, and component diagnostics.

E.1 Headline Robustness

Table 14 and Fig. 4 summarize the accuracy band of the deployable workflows together with the two scope diagnostics for reference. Table 15 reports the human answerability reference used in the main comparison.

E.2 Human Reference and Matched Comparison

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% candidate-level first-pass agreement (Cohen's κ=0.88) on the 1,553 parseable Phase-2 candidate records before adjudication. The five-reader results provide the human answerability and model-comparison reference.

E.2.1 Matched Model–Human Comparison

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 therefore 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 16 reports model statistics matched to the human comparator.

E.3 Capability and Platform Profiles

Fig. 5 reports per-model capability accuracy under Baseline. Figs. 6 and 7 provide the corresponding platform-level views.

Table 12: **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.

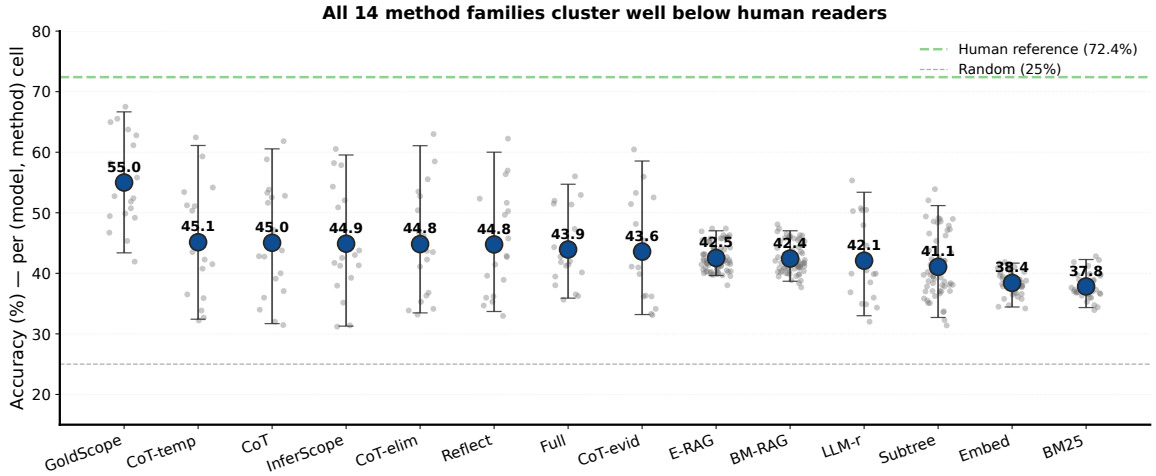

Figure 4: **Method-family accuracy band on LongSocialBench.** Each grey dot is one model-workflow cell from the 18×29 evaluation grid; markers and error bars show the family mean and ± 1 standard deviation across cells. All deployable workflows cluster below the human answerability reference of 72.4%; the Gold-Scope diagnostic reaches 55.0%.

Table 13: **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	56.0	61.8	55.3	63.0	60.0
Gemini-2.5-Pro	51.5	58.8	50.4	59.3	51.3
GPT-5	52.9	53.8	50.3	56.4	58.3
Claude-Sonnet-4.5	51.4	52.8	50.8	53.4	54.4
Claude-4	52.0	52.5	50.5	53.5	57.0
Claude-Sonnet-4.6	44.3	51.6	39.9	51.6	45.7
(Strong open-weight)					
GLM-5	41.5	53.3	44.6	53.3	41.3
DeepSeek-R1	47.3	47.1	47.9	49.0	50.6
Qwen3-Next-80B	42.7	42.8	39.9	43.7	44.6
Mistral-Large-3	42.0	42.7	38.5	43.5	43.3
(Mid open-weight)					
Llama-4-Mav-17B	41.2	43.8	40.5	45.3	41.5
Llama-4-Scout-17B	40.2	40.0	36.9	41.7	41.3
Llama-3.1-70B	38.0	39.1	36.0	41.1	39.5
Qwen3-32B	41.9	31.5	34.9	43.3	43.8
(Small / mid-weak)					
GPT-OSS-120B	39.5	37.1	34.3	42.2	42.0
GPT-OSS-20B	36.3	36.0	32.0	40.2	39.2
Llama-3.2-90B	35.6	34.0	39.9	40.0	35.3
Gemma-3-27B	36.5	32.0	35.0	40.4	38.1
Mean (18)	43.9	45.0	42.1	47.8	46.0

Table 14: **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
Chain-of-thought	90	44.7	[43.0, 46.5]	+0.7
Full-context	18	43.9	[41.3, 46.9]	—
RAG k=50	72	42.7	[42.1, 43.2]	−1.3
RAG k=20	72	42.3	[41.8, 42.8]	−1.6
LLM-as-retriever	18	42.1	[39.1, 45.4]	−1.8
Structure-aware retrieval	54	41.1	[39.6, 42.6]	−2.9
Retrieval k=50	36	39.0	[38.4, 39.7]	−4.9
Retrieval k=20	36	37.2	[36.6, 37.9]	−6.7
Infer-Scope diagnostic	18	44.9	[40.8, 49.2]	+1.0
Gold-Scope diagnostic	18	55.0	[51.8, 58.4]	+11.1

E.4 Reasoning Prompt Effects

Table 17 summarizes per-model lift across reasoning-style workflows, while Fig. 8 provides the corresponding visual comparison.

E.5 Retrieval Diagnostics

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 the released items.

E.6 Validity and Robustness Checks

This subsection reports three complementary checks discussed in §5.5: shortcut robustness, contamination, and the relationship between episode length and accuracy.

Table 15: **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%.

Table 16: **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	66.1	$[61.1, 71.8]$
Claude-Opus-4.7	65.4	$[59.7, 71.8]$
Closed-frontier mean (n=6)	59.3	—
Strong-open mean (n=4)	51.1	—
18-model fleet mean	51.6	—
Human readers (n=5)	**72.4**	$[\mathbf{71.1}, \mathbf{73.7}]$
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.

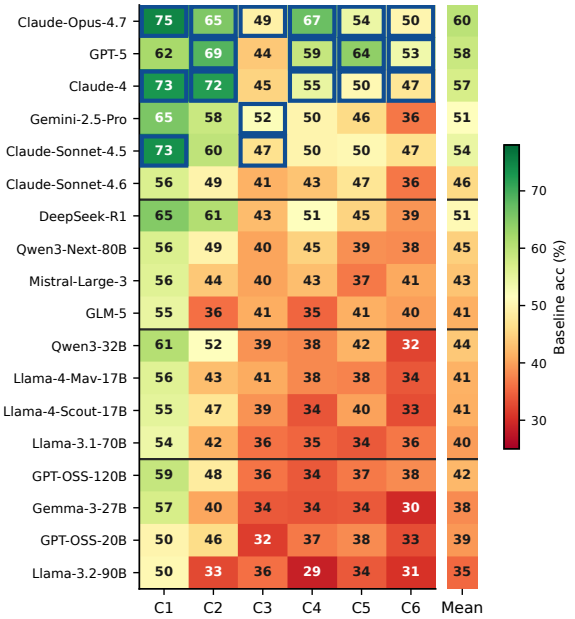

Figure 5: **Per-model accuracy on each social capability under Baseline.** Cells are mean accuracy on the 1,462-item benchmark restricted to that capability; the rightmost strip is the per-model mean.

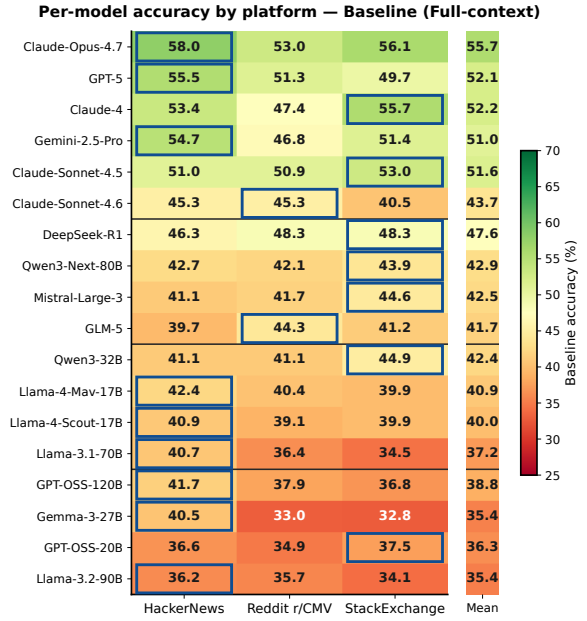

Figure 6: **Per-model accuracy by platform under Baseline.** Hacker News, Reddit r/CMV, Stack Exchange, and per-model mean are shown for each model.

E.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 are evaluated on stratified subsets by capability and source. Table 19 summarizes the shortcut tested and the observed result.

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 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 an interpretive 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 without selecting among A–D.

On the same items, open-answer accuracy is 7.6 pp lower than multiple-choice, while the model ranking is preserved almost exactly (Spearman ρ=1.000 across the 18 models). Together, the reduced-input, shuffled-thread, and open-answer probes indicate that the benchmark is not solvable from answer options, question wording, topic cues,

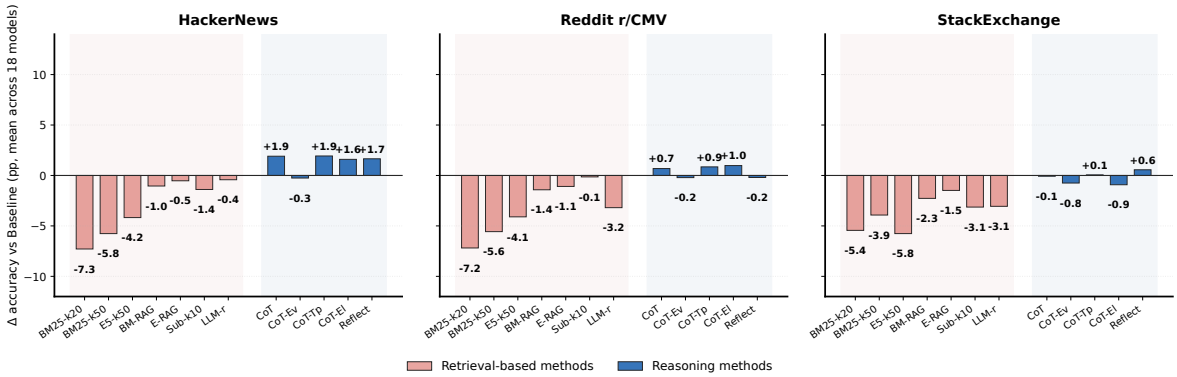

Figure 7: **Domain × method effect.** Each panel is one platform; bars are mean Δ accuracy vs. Baseline averaged across 18 models.

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**
GLM-5	+11.8	+2.3	+9.6	+5.3	+8.8	+7.6
Gemini-2.5-Pro	+7.3	+4.4	+7.8	+7.0	+5.5	+6.4
Claude-Opus-4.7	+5.8	+4.4	+6.4	+7.0	+6.2	+6.0
Claude-Sonnet-4.6	+7.3	+3.8	+6.9	+6.1	+5.3	+5.9
Llama-4-Mav-17B	+2.5	-0.3	+4.1	+1.0	+1.6	+1.8
GPT-5	+0.9	+0.3	+1.2	+2.6	+3.4	+1.7
Claude-Sonnet-4.5	+1.4	+0.1	+2.1	+1.4	+1.0	+1.2
Mistral-Large-3	+0.8	+0.5	+0.3	+1.5	-0.5	+0.5
GPT-OSS-120B	-2.4	+0.4	+2.1	+1.6	-0.5	+0.2
Claude-4	+0.5	+0.5	-1.6	+1.5	-0.3	+0.1
Qwen3-Next-80B	+0.1	-1.6	+0.8	+1.0	-0.1	+0.0
Llama-3.1-70B	+1.1	-1.7	+2.7	-1.2	-2.1	-0.2
GPT-OSS-20B	-0.3	-0.1	-0.3	-1.0	-0.1	-0.3
DeepSeek-R1	-0.2	-1.9	-1.4	-1.1	-1.6	-1.2
Llama-3.2-90B	-1.6	-1.5	-2.9	-2.5	-2.7	-2.2
Llama-4-Scout-17B	-0.1	-4.0	-3.6	-3.8	-0.5	-2.4
Gemma-3-27B	-4.5	-3.2	-2.7	-2.4	-1.8	-2.9
Qwen3-32B	-10.4	-8.8	-9.6	-8.0	-6.6	-8.7
Mean (18)	+1.1	-0.3	+1.2	+0.9	+0.8	+0.7

Table 18: **Evidence-node recall for standard comment-level retrievers.** Values are mean recall against author-side evidence-node lists, averaged over the 1,462 released items. Only 8.8% of items reach $\geq 75\%$ recall in the best of the four retrievers, and 3.5% attain full coverage at the largest budget.

Retriever	**Budget**	**Mean evidence recall**
BM25	k=20	0.201
E5	k=20	0.203
BM25	k=50	0.308
E5	k=50	0.310

or bag-of-comments cues alone, and that the model ranking is not an artifact of the multiple-choice format.

E.6.2 Pre- and Post-Cutoff Contamination Probe

We test whether memorization of source pages inflates accuracy on platforms that overlap pre-training corpora, comparing two time-disjoint cohorts: Stack Exchange, collected from 2008 to 2018 archives, and Hacker News, collected from 2026 Q1. Table 20 reports the per-tier Baseline means. The pre-cutoff cohort shows no systematic advantage on any tier, providing no indication that source memorization explains the headline gap. The probe cannot rule contamination out entirely.

Table 19: **Shortcut robustness summary.** Each probe removes or perturbs information that would be unnecessary if the corresponding shortcut were sufficient. Accuracy values are means across 18 models on the probe subsets; the full-benchmark Baseline is 43.9% for context. For open-answer, the table reports the same-item accuracy change and model-ranking correlation.

Probe	Shortcut tested	Observed	Interpretation
Option-only	option prior	31.5%	option text alone is insufficient
Question-only	stem leakage	38.0%	question wording alone is insufficient
Title/root-only	topic familiarity	35.4%	topic-level information is insufficient without replies
Shuffled thread	bag-of-comments inference	40.4%	performance remains limited when local thread order is perturbed
Open-answer subset	multiple-choice artifact	-7.6 pp; ρ=1.00	absolute accuracy drops, but the model ranking is preserved

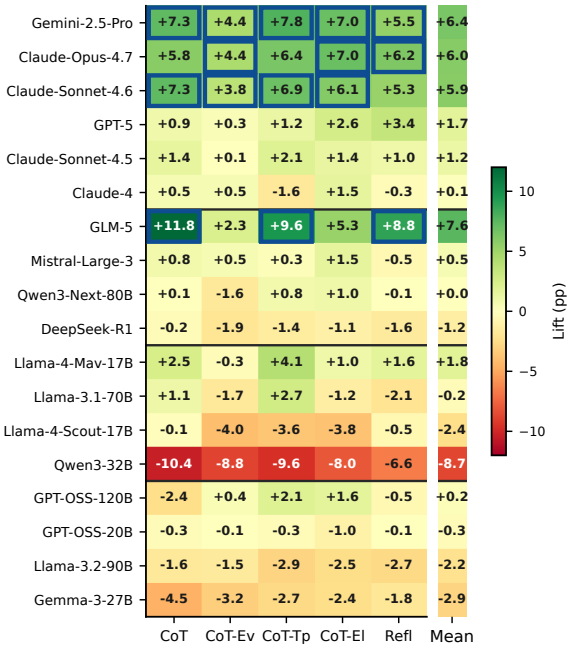

Figure 8: **Reasoning workflow lift over Baseline, per model.** Cells are accuracy under the named workflow minus Baseline in percentage points; the rightmost strip is the per-model mean lift.

Table 20: **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)	53.0	49.1	51.1	+1.9
Strong open (4)	42.4	44.1	44.5	−2.1
Mid open (4)	41.3	39.3	39.8	+1.5
Small (4)	38.8	35.4	35.3	+3.5
All 18	**44.9**	**42.8**	**43.6**	**+1.3**

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

E.6.3 Episode-Length Effect

Episodes vary in size, so we check whether accuracy simply tracks input length. Figure 9 plots per-episode mean Baseline accuracy against episode length; the least-squares fit has slope -0.07 pp per 1K tokens and the Spearman rank correlation is ρ=-0.18. Length is therefore only weakly anti-correlated with accuracy, far too weak to account for the roughly 28 pp gap to the human reference.

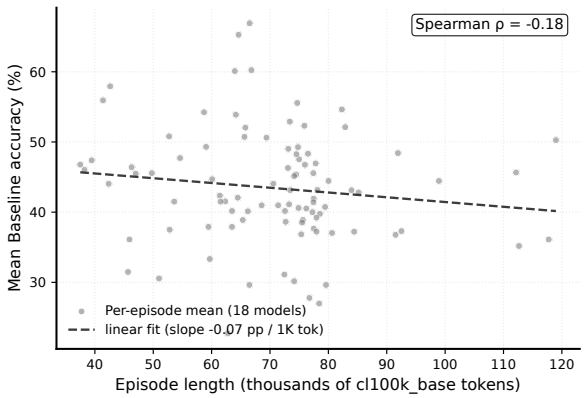

Figure 9: **Episode length vs. Baseline accuracy.** Each point is one of the 94 released episodes: mean Baseline accuracy across the 18 models against episode length in `cl100k_base` tokens, with the least-squares fit and the pooled Spearman ρ.

E.7 Scope and Component Diagnostics

Per-tier Gold-Scope lift is reported in the main text (Table 4). Table 21 reports the component oracle case study on the stratified subset.

E.7.1 Residual Errors under Gold-Scope

We manually inspect 50 Claude-Opus-4.7 errors under Gold-Scope and categorize the immediate source of each failure (Table 22).

Table 21: **Component oracle case study.** Claude-Opus-4.7 on the stratified 30-episode / 510-item subset.

Condition	Accuracy
CoT baseline	63.1
Gold scope	66.7
Gold per-comment labels	81.4
Gold scope + gold labels	98.6
Structured enumerate-label-aggregate pipeline	70.0

Table 22: **Manual audit of** 50 **Claude-Opus-**4.7 **Gold-Scope errors.** Each error is assigned the immediate source of the failure; percentages describe the audited sample.

Error type	Share (%)
Criterion use	**60**
Thread structure	16
Counting rule	14
Surface cue	10

Criterion-use errors dominate: the model has access to the relevant region but applies the question-specific inclusion or interpretation rule too broadly or too narrowly.

F Representative Questions

This appendix presents one representative LongSocialBench item for each of the six capability surfaces, providing a worked example for every surface in Table 2. The examples are chosen for readability rather than difficulty and illustrate the social-reading decision each capability is designed to test. Question stems and answer options follow the released benchmark items, with usernames replaced by structural references where they are not essential; rationales are shortened summaries of the released criteria.

F.1 Example 1: Local social meaning

C1: Local social meaning

Episode topic. A Hacker News thread on IPv6's thirtieth birthday and its still-incomplete adoption. One top-level comment blames the confusion on display in the comment section itself—in a forum whose users are supposedly more technically savvy than most—and draws seven direct replies.

Question. *"One top-level comment argues that the confusion on show in this very discussion is why IPv6 has not been adopted. Which of its direct replies accepts that observation but aims it back at the people making it?"*

Options. A. A request for concrete examples of the confusion being alleged

B. A joke about how the commenters rate their own expertise

C. A link to adoption figures set against the closing prediction

D. A claim that the real obstacle was the absence of any incentive

Gold and rationale. B. The winning reply quotes the "yawning chasm of gaps in knowledge" line and answers only that there is a Dunning–Kruger joke to be made here: it grants the diagnosis and turns it on the commenters, the author of the diagnosis included. The other options name real replies in the same set, which ask which misconceptions are meant, answer the closing prediction with an adoption figure, or reject the causal story outright. Every option shares the thread's vocabulary, so the item turns on reading a one-line joke as agreement plus turnabout against the comment it sits under.

F.2 Example 2: Reply-pair interaction

C2: Reply-pair interaction

Episode topic. A Hacker News thread on a post titled "Google is dead. Where do we go now?" One commenter reports that they and the other small software-business owners they know have all watched paid search become less profitable, and that most have given up on it. Four people reply directly.

Question. *"A comment reports that the small software-business owners its author knows have all found paid search less and less profitable. Which of its direct replies answers with the same kind of evidence but the opposite conclusion?"*

Options. A. The reply describing the commenter's own slow decline and shrinking budget

B. The reply asking what other approaches the commenter uses instead

C. The reply describing many owners in other lines of business whose results have held up

D. The reply describing twenty years of running such campaigns and rising complexity

Gold and rationale. C. The parent argues from a circle of acquaintances who all run the same kind of business and all report the same decline; the gold reply answers in exactly that currency—many non-software owners whose results have held up—matching the form of the evidence while reversing the verdict. Two of the remaining replies corroborate the parent from personal experience and the third asks a question instead of offering evidence at all. Choosing correctly requires reading each reply against what its parent claimed and how it

claimed it, not judging any reply on its own.

F.3 Example 3: Collective discussion structure

C3: Collective discussion structure

Episode topic. A Hacker News thread about a macOS redesign that made window resizing harder. Mid-discussion, one top-level comment proposes crowdfunding billboards in Cupertino to tell Apple how badly it missed; ten replies land underneath it.

Question. *"A top-level comment proposes crowdfunding a Cupertino billboard criticizing Tahoe. How do its direct replies split among endorsing the idea, pushing back, and moving to a different topic?"*

Options. **A. Five endorse; three push back; two move elsewhere**

B. Six endorse; two push back; two move elsewhere

C. Three endorse; five push back; two move elsewhere

D. Four endorse; four push back; two move elsewhere

Gold and rationale. **A.** Five replies join in—pledging $25, quoting real rates for an electronic billboard on Highway 101, and riffing on what it should say—while three deflate the idea, arguing the money belongs with open-source projects, that outside pressure will not turn the ship around, or that Cupertino may not even have billboards. Two more give a general opinion of the redesign without taking any position on the proposal. The item turns on reading a joke as endorsement and a rhetorical question as refusal, then holding the whole reply set in view at once.

F.4 Example 4: Temporal dynamics

C4: Temporal dynamics

Episode topic. A Hacker News thread on a television showrunner telling viewers to switch off their sets' "garbage" picture and sound processing. The episode's longest reply chain opens with a viewer who needs subtitles to follow anything and runs several rounds on whether built-in speakers are to blame.

Question. *"Walk down this episode's longest reply chain. At which position does the conversation permanently move away from the opening topic of audio intelligibility? Count the starting comment as position one."*

Options. A. Position four B. Position six C. Position ten **D. Position eight**

Gold and rationale. **D.** The first seven comments stay on one subject—subtitles, dialogue mixing, and whether a thin panel can produce intelligible sound—each turn answering the last. The eighth is a two-word correction of which service a show streams on, the first comment in the chain that says nothing about audio, and every later turn follows it into brand names becoming generic words rather than back. Answering requires walking the chain in order and separating the true point of no return from the later turn at which the new topic merely becomes obvious.

F.5 Example 5: Participant profile

C5: Participant profile

Episode topic. A ChangeMyView discussion arguing that nuclear energy is good, where the leading replies press on cost, waste, and build times. The pro-nuclear participant profiled in the question appears in five separate parts of the thread.

Question. *"Among this participant's comments in the episode, how many use a concrete number, statistic, or calculation to rebut another comment?"*

Options. A. Six comments B. Five comments **C. Three comments** D. Four comments

Gold and rationale. **C.** This participant writes seven comments spread across five sub-discussions, and the recognizable move is answering objections with hard figures: Germany's spending on renewables against a grid still far dirtier than France's, fifty gigawatt-hours of storage covering roughly seven minutes of average US load, and no recorded injuries from spent fuel alongside the eight-day half-life of iodine-131. The remaining four comments contain no figures at all. Reaching the answer means collecting one participant's comments from across a long thread and judging what each one does, rather than reading any single exchange.

F.6 Example 6: Cross-subdiscussion variation

C6: Cross-subdiscussion variation

Episode topic. A Hacker News thread on a blog post arguing that macOS Tahoe's new icons are indefensible. The episode's second-most prolific commenter appears in eight separate top-level threads—panning the system icon set in one, cataloguing broken permissions in another, and opening a third one themselves.

Question. *"The episode's second-most prolific user writes comments across eight distinct top-level*

threads. In which thread does that user adopt a notably different stance from their pattern elsewhere—offering a conciliatory defence of macOS against a Linux comparison rather than criticizing Apple's design choices?"

Options. A. The thread where the opening comment tells people to click the snow buttons

B. The thread whose opening comment calls Apple's designer worship the root problem

C. The thread about peak desktop UI contenders from the pre-2000 era

D. The thread discussing the article's hamburger menu and user agreement joke

Gold and rationale. B. In seven of the eight threads this commenter is among the harshest Apple critics in the episode, dismissing the system icons as "cheap, lame, boring, flat little hieroglyphics" and reporting that the permissions panel has been broken for several releases. Only under the comment blaming Apple's worship of its own designers does that commenter change sides, replying that no Linux desktop environment comes close to the macOS experience despite its flaws and regressions. The baseline has to be assembled from several unconnected sub-discussions before the exception is visible; no single thread contains enough to decide it.

G Construction Workflow

This appendix summarizes the contributor-facing construction workflow and the portal controls that enforce role separation. Contributors cannot verify or adjudicate items they authored, and human-reference readers do not participate in item construction.

G.1 Annotation and Verification Protocol

Table 23 summarizes the four stages of item construction, verification, and adjudication while keeping model-facing inputs separate from internal construction records.

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

H.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, 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

Table 23: **LongSocialBench annotation and verification protocol.** Authoring records define the social criterion and supporting evidence; verification records test whether an independent reader can recover the gold option under the same input condition given to models.

Stage	Contributor task	Quality control
Authoring	Read the complete discussion, choose a clear scope, and write the question stem, four plausible options, 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 reflect 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.	These controls reduce locator leakage, option-paraphrase matching, keyword shortcuts, and open-web contamination. Capability and item-form tags are retained 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 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 re-verified, or removed. A released item must have independent agreement or a stable adjudicated answer.

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.

I 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-response model outputs, parsed choices and parse status for multiple-choice settings, open-answer content scores, and per-cell accuracy records over the 18×29 model-by-workflow grid used to produce the reported results.

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