SCM-BENCH: AN AUDITABLE BENCHMARK FOR CONTEXT EFFICIENCY IN MULTI-TURN DIALOGUE

Anonymous authors
Paper under double-blind review

ABSTRACT

Context efficiency in multi-turn dialogue concerns the answer produced from a particular history, not only the amount of history retained. SCM-Bench defines a versioned target-turn task and an auditable framework for studying history retention, selection, compression, and model routing. Deployment decisions use the current request and strictly preceding legal history; evaluation uses frozen requirements, anonymous answers, and complete legal history. The framework distinguishes reference-history budgets from provider-native generation usage and shared-matrix construction from policy deployment. Its historical profile contains 3,893 deployment-eligible inputs from 261 sessions, but candidate selection is not a completed formal evaluation. Eight context interfaces and independent versus selected-context routing support controlled comparisons without implying learned joint optimization. The proposed analysis preserves failures, missing outcomes, and constraint applicability. Available development re-evaluations do not supply a qualified quality–efficiency release.

1 INTRODUCTION

Multi-turn assistance requires more than answering the latest message in isolation. A useful response may need to preserve an earlier constraint, build on a previous decision, or recognize that a later instruction supersedes an older one. Keeping less history is therefore not, by itself, evidence of better efficiency; neither is choosing a model with a lower price. The testable question is whether a specified combination of history retention, selection or compression, and model routing changes the quality and constraint adherence of the next answer, and what resources the complete generation process consumes. Answering this question requires measuring outputs, not merely their input sizes.

Related work offers complementary perspectives. LongMemEval studies memory over conversation histories, SCBench examines long-context methods through the key–value cache lifecycle, and RouteLLM studies model selection from preference data (Wu et al., 2025; Li et al., 2025; Ong et al., 2025). Memory accuracy, cache behavior, and routing performance are not interchangeable measurements. Our focus is a target-turn comparison: how the history actually supplied to a generator and the choice of generator relate to a newly produced answer under explicit input and accounting boundaries.

We introduce SCM-Bench as a benchmark task and auditable evaluation framework for this question. Each target consists of a current query and its strictly preceding, permitted history. Provenance, branch verification, session and lineage relationships, and prior exposure distinguish development and retrospective analyses from future confirmation on genuinely unexposed conversations. Structural eligibility is not human-validated task quality; a deployment-visible view does not imply public-release permission.

Figure 1 separates deployment decisions from post-answer evaluation. The study asks how context strategies compare under controlled models and budgets, how model choice changes that comparison, and whether giving a router the actual selected context changes outcomes relative to independent decisions. The sequential interface is an information-flow intervention, not a claim of learned joint optimization. Matched target–replicate observations and a valid evaluation release are necessary to answer these questions.

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

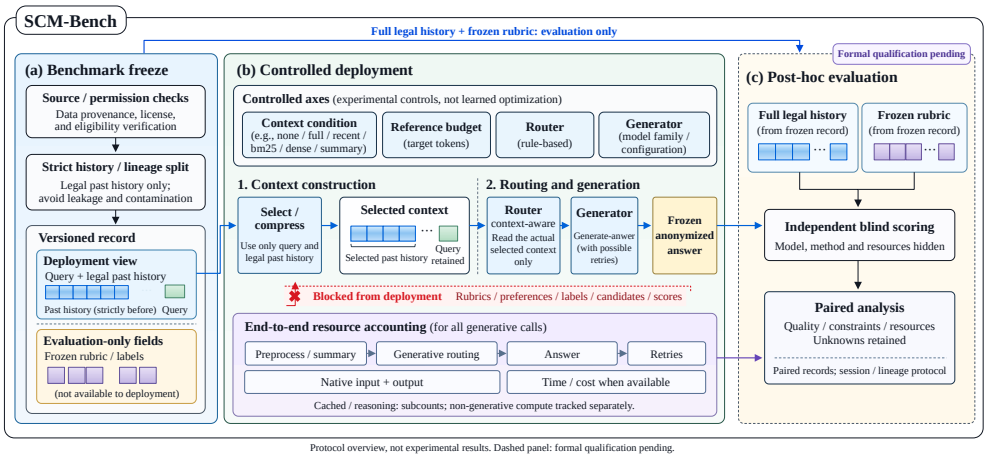


Figure 1: SCM-Bench protocol overview. Deployment uses only the query and legal prior history; frozen evaluation information stays outside deployment. This is a protocol schematic, not an empirical result.

Our contributions are (1) a versioned target-turn task and data contract with provenance, branch-history, and exposure controls; (2) a framework distinguishing deployment inputs from evaluation evidence, and history budgets from policy-generation usage and construction costs. Historical profiles and development diagnostics support the first two descriptions, not quality preservation, unseen-conversation generalization, or end-to-end savings.

2 Related Work

Conversational capability. Memory benchmarks already evaluate generated outputs. LoCoMo combines conversational question answering, event summarization, and multimodal generation; LongMemEval evaluates several forms of long-term recall and reasoning (Maharana et al., 2024; Wu et al., 2025). MultiChallenge is particularly close to target-turn assessment: a final response must satisfy earlier instructions, use relevant user information, handle versioned edits, and maintain coherence (Sirdeshmukh et al., 2025). MemoryArena studies interdependent multi-session tasks, while ContextWeave reconstructs workflows from real documents and evaluates task outcomes and preferences (He et al., 2026; Wang et al., 2026). ContextWeave also holds the acting model or memory system fixed in separate comparisons. Neither evaluating useful memory nor controlling model and memory choices alone distinguishes SCM-Bench. Our narrower object is a versioned historical target request with invariant current query and evaluation requirements. Historical provenance is not independently validated natural-conversation sampling.

Intervention layer. Retrieval and summarization alter text before generation; LLMLingua-2 learns extractive prompt compression through token classification (Pan et al., 2024). SCBench covers KV-cache generation, compression, retrieval, and loading, including prompt compression and cache reuse (Li et al., 2025). Its controlled multi-turn protocol is not equivalent to replaying a fixed historical transcript without persistent cache state. RouteLLM learns strong-versus-weak model selection (Ong et al., 2025); prompt-based routing does not guarantee access to complete conversation history. SCM-Bench's selected-context routing is an information-flow comparison, not a reproduction of RouteLLM.

Quality and resources. Budget-aware evaluation and preprocessing accounting are established: LongMemEval varies retrieval budgets, RouterBench compares routing using precomputed outcomes, RouteLLM reports routing overhead, and LLMLingua-2 includes compression latency (Wu et al., 2025; Hu et al., 2024; Ong et al., 2025; Pan et al., 2024). Our protocol specifies a particular boundary: each target–replicate–policy includes all required generative calls and retries, while non-generative work and evaluation costs remain separately observable. Native tokens are not a common

compute unit. Anonymous, full-history assessment retains unknown judgments; JudgeBench motivates checking judges against objective evidence rather than agreement alone, but does not validate our panel (Tan et al., 2025). Supplementary Appendix G compares task and measurement scopes without ranking unlike scores.

3 Task Definition and Data

3.1 Target queries, legal history, and eligibility

The unit is a target record, not an entire conversation or an independent user. Write

$$x_i = (q_i, H_i), \quad H_i = (h_{i,1}, \ldots, h_{i,n_i}), \qquad o_{i,1} < \cdots < o_{i,n_i} < o_i^{\text{target}}. \tag{1}$$

The query is nonempty and unchanged; historical identifiers are unique, belong to the same session, and follow the target's actual ancestral branch. The target is excluded. Structural checks do not establish branch authenticity or authorization. Source adapters reconstruct explicit lineage and selected historical continuations, not branches inferred from similarity. Missing attachments, tools, or modalities must be flagged rather than silently reduced to text.

This is *static target-turn replay*: each new answer is evaluated against a fixed historical transcript, not inserted into later targets' histories. It does not measure online state evolution, cumulative errors, or subsequent user behavior. Shared sessions, copies, and derived tasks remain dependent.

Input eligibility requires verified history, no missing modalities, and a permission status other than unknown. Generation uses this mask; original-choice, dependency, reviewed-evidence, and rubric masks are separate. An old preference is not correctness for a new answer. The inherited *human* and *silver* labels describe source/original-label provenance, not new-answer validation. Complete-input exact duplicates and reviewed duplicate groups connect sessions before splits; the near-duplicate proposer only normalizes whitespace over complete inputs. Semantic leakage remains unverified.

3.2 Historical population and candidate plan

Table 1 distinguishes source inputs, eligible inputs, selected candidates, and formal evaluation. The B0 snapshot (22 September 2026) contains 4,009 source records and 3,893 eligible inputs from 261 sessions. The difference of 116 is subtraction, not an exclusion-cause audit. Legacy train/dev/diagnostic contain 2,578/655/660 inputs. These are not distinct-user counts.

Table 1: Four denominators in the historical snapshot. Candidate rows overlap where noted; none is a formally scored cohort. A dash is unavailable, not zero.

Population	Records	Sessions	Status
Source pool	4,009	—	Raw session count missing
Eligible inputs	3,893	261	Historically exposed
Development candidates	120	37	Includes pilot (40; 34)
Historical candidates	143	43	Target 500; gap 357
Fresh candidates	0	0	Target 500; gap 500
Validated Challenge	0	0	Target 200; gap 200
Formal evaluated cohort	—	—	Manifest unavailable

The score-blind B0 planner preserves splits, uses seed 20260921, caps selection at four targets per session, and never backfills from train. Reserving Challenge-parent clusters leaves a Development planning pool of 634, not all 655 dev inputs. Development includes the 40-target, 34-session pilot. Historical has a 357-target shortfall. Deep–Historical nests the same 143 Historical targets; it adds no observations. All B0 sources are historically exposed. Fresh has no materialized candidates and Challenge no validated variants; two reserved parents are not Challenge examples. The 263 Development-plus-Historical candidates are not a formal benchmark denominator. A formal evaluation population and bound manifest are still required.

3.3 LENGTH AND ANNOTATION SCOPE

Reference lengths in B0 use `tiktoken:cl100k_base` (reported revision 0.14.0). The profiler counts newline-joined role–text history, excluding the query. Reported P50/P95/maximum are 7,641/34,785/122,428 tokens; quantiles use sorted index $\lfloor (n-1)p \rfloor$ without interpolation. This profiler is not the generation renderer and does not measure native usage. Aggregate identity and totals were checked, not private raw-history tokenization. Figure 2 shows the complete supplied bucket counts.

The B0 rubric, reviewed-turn evidence, reviewed-span evidence, and confirmation masks are all zero; span eligibility is hard-coded false in this release. Dependency is required/not required/unknown: helpful history need not be necessary. Validated dependency-category counts are unavailable. Later rubric work does not retroactively alter these B0 counters. Supplementary Appendix A gives all source, split, length, mask, candidate, and shortfall aggregates, with access limitations.

3.4 VERSION FREEZING AND ACCESS

The release design freezes candidate selection, then the final input/rubric set and exclusions, experimental configuration and generated answers, and finally evaluation labels. Inputs, rubrics, protocols, answers, and quality versions are separately bound; revising one requires a new version and explicit mapping. The historical B0 data/framework layer and later B1/B2 runtime/offline layers are not a single experiment. A source-reading revision identifies the inspected artifact, not the code that executed every reported result. Later rubric construction and evaluator code cannot silently replace old annotation or qualification status.

This distinction matters operationally: incomplete collection is not a zero-sized scored cohort, and an implementation smoke test is not model evaluation. The present material permits aggregate reconciliation and document reconstruction, not an independent replay of private branches or historical calls. A final release needs both the named manifests and the actual objects they bind. Data rights, leakage review, and unexposed confirmation must be established separately from structural eligibility.

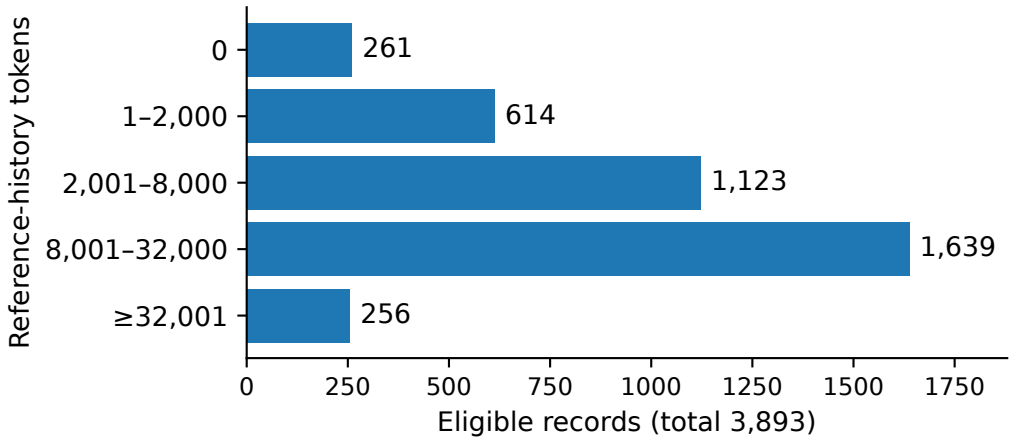


Figure 2: Historical input profile, not a model-performance result. Counts cover all 3,893 eligible B0 inputs; categories use the reference-history profiler and exclude the query. Bucket widths are unequal, so these are categorical counts rather than a density histogram.

4 FRAMEWORK AND METHODS

4.1 INFORMATION BOUNDARY AND CONTEXT BUDGET

For budget b, a selector returns $c = S(q, H, b)$ and a router chooses $m = R(q, c)$; the generator then produces $a = G_m(q, c)$ under fixed system and decoding settings. Deployment receives only the query and legal history, not original candidates, preferences, rubric, evidence labels, or quality. After answers are sealed, evaluation receives the complete legal history and frozen requirements. Full-history generation is not a gold answer.

The generation renderer writes each fragment as `[message_id] role:`, a newline, then text, joining fragments with two newlines in chronological order. With reference counter N_{ref} and renderer τ, define

$$F = N_{\text{ref}}(\tau(H)), \qquad b = \lfloor rF \rfloor, \qquad \rho = \frac{N_{\text{ref}}(c)}{F} \quad (F > 0). \tag{2}$$

Headers and separators count toward history budgets. Query, system prompt, and outer wrappers remain intact outside this budget. Empty history makes ρ not applicable. Realized retention can undershoot the requested ratio; no padding is added. Executed tokenizer/renderer bindings remain unavailable.

4.2 EIGHT CONTEXT INTERFACES AND FOUR POLICIES

The interfaces are `none`, `full`, `last1`, `recent`, `bm25`, `dense`, `llmlingua2`, and `summary`. None is empty; Full retains all legal history. Last1 retains the latest complete user block containing an assistant reply, ending before the next user, and is not ratio-clipped. Recent packs newest-first suffix fragments; BM25 and Dense rank relevance and pack prefix fragments. Ties favor later messages; retained fragments are rendered chronologically with traceable spans.

BM25 uses BM25Okapi with CJK-unigram/word tokenization. Dense uses a pinned local SentenceTransformer, normalized embeddings and cosine similarity; over-window complete queries/messages fail rather than being silently truncated. LLMLingua2 compresses history without receiving the query, forces newline handling, and then applies reference-budget prefix clipping. Its upstream long-input handling is not established by the Dense checks. Summary sees the query and full rendered history, has a separately configured output limit, and is also reference-clipped. All summary calls and retries count. These are implemented interfaces, not evidence that all eight were evaluated.

Table 2 defines the policy comparison. A query-only router is $R_q(q)$; a selected-context router is $R_c(q, c_S)$. "Independent" means separate information inputs, not statistical independence or concurrent execution. Sequential does not imply a learned joint policy. The controlled contrast holds selected context, budget, model pool, rule threshold, and decoding fixed and changes only the router's view. Matched independent/sequential results are unavailable.

Table 2: Policy interfaces. $c_S = S(q, H, b)$; m_0 is a fixed generator. These are controlled designs, not completed comparisons.

Policy	Generator history	Model decision
Context only	c_S	m_0
Model only	$\tau(H)$	$R_q(q)$
Independent	c_S	$R_q(q)$
Sequential	c_S	$R_c(q, c_S)$

For the transparent rules, model choice can be written explicitly as

$$f_\theta(z) = \begin{cases} m_{\text{last}}, & \text{Hard}(z) \ \vee \ N_{\text{ref}}(z) \geq \theta, \\ m_{\text{first}}, & \text{otherwise.} \end{cases} \tag{3}$$

The query-only policy uses $z = q$. The selected-context policy uses the actual query–header–context concatenation, not a hidden reference to the full history. A context that changes the threshold

decision is therefore an information-flow effect; a claim that it improves quality still needs matched answers and a qualified evaluator.

Available routers include fixed, seeded random, train-prior, and transparent rule-based interfaces. The rule routes to the last configured model when a keyword or reference-length threshold fires, otherwise the first; slot order is not a capability guarantee. The context-aware view contains the query, a fixed selected-context header, and c_S. Even that header contributes to the threshold. Adaptive budgets are explicit rules, not learned optima. Formal calibration and model identities remain unresolved; fixture settings are not experimental hyperparameters. Appendix B supplies exact selection behavior, pseudocode, prompt templates, and configuration requirements.

4.3 Policy resources, sharing, and failures

For assigned unit $u = (i, p, k)$, let $\mathcal{C}_u$ contain all required generative call keys and $\mathcal{A}(c)$ their attempts. Complete deployment-native usage is

$$T_u^{\text{native}} = \sum_{c \in \mathcal{C}_u} \sum_{a \in \mathcal{A}(c)} \left(t_a^{\text{in}} + t_a^{\text{out}} \right), \tag{4}$$

only when every required attempt has known usage. Otherwise the total is unavailable, with known subtotals and coverage retained. This includes summaries, generative routing when used, answers, and retries; the current rule routers are non-generative. Cached-input and reasoning tokens are subsets, not additions to an already inclusive total. A pre-generation failure may have zero attempts without a complete answer-usage observation. Non-generative embedding/compression work is reported separately, not called free.

Shared-matrix construction counts unique calls once. Each logical policy still owns all calls it would require in deployment, so summing logical totals can double-count construction. Cache hits are not free deployment. Online timing requires no-reuse traces with a per-unit scope; elapsed stage spans are not sums of overlapping intervals. Matrix timing does not establish deployment latency. Native tokens across tokenizers are accounting units, not interchangeable FLOPs, dollars, energy, or latency. Historical price snapshots and full measured traces are missing (E07).

Decisions and started attempts are written before requests. Failures stay in assigned units; missing usage is not zero. Dependency, identity, quota, or capacity failures cannot silently invoke an easier proxy. Evaluation and rubric-construction costs are separate from policy deployment. Appendix D states the attribution and statistical contracts.

Artifact graph and completion. A run records the plan, configuration, assigned jobs and inputs, selected-context identities, router decisions, final prompts, required call keys, and response status. A call receipt binds all attempts, returned model identity, finish state, native usage, and interruption/retry information. A completed response artifact is not the same as a complete run manifest. Quality labels subsequently bind the sealed answers, rubric, prompt/schema versions, and judging protocol; analysis must bind that same population and quality version. Missing receipts cannot be repaired by finding a later report with a similar label. This graph makes intended attribution inspectable, but the complete historical and formal runtime objects remain missing (E11).

5 Evaluation Protocol

5.1 Prospective freeze and answer assessment

This section specifies a proposed protocol, not a completed preregistration or an executed formal evaluation. Freeze inputs, exclusions, rubrics, condition matrix, replicas, random order, model/tokenizer identities, prompts, windows, output reservations, retry limits, and statistics before generation. The complete formal catalogue and statistical freeze are not supplied. Complete native prompts plus output reserves must fit every applicable model. Common capacity eligibility is determined before assignment, not by later deleting difficult failures.

Preflight checks frozen configuration and dependencies; it does not certify remote model immutability or scientific validity. Native windows differ among answer generators, summary models, encoders, and judges, and each must check its own actual input. A small retained history does not

make an encoder's full-message input short. Output limits, deterministic settings or seed support, attempt classes, and stopping rules must be bound with the executed requests rather than inferred from a method name.

Post-assignment generation failures remain assigned; the proposed zero task-success treatment requires the signed freeze. Delivered but truncated answers are assessed as delivered. Judge transport/parsing/evidence failures are distinct from generation failure and from legitimate semantic unknown. Unknown is not retried until favorable. Appendix C defines the rubric, prompt, schema, and unknown states.

Requirements before answers. In the v1 design, A/B independently extract query and full-history requirements, source anchors, active/superseded constraints, and pass/fail/unknown rules. Both proposals are sealed before C sees them with the original input for one merge, even when they agree. Critical ambiguity blocks readiness; noncritical uncertainty remains explicit. The rubric must contain a critical task-completion item and may not use a generated answer as its source.

Blind scoring and unresolved quorum. A/B assess the same full legal history, frozen rubric, and anonymous answer without model, method, routing, cost, or old preference information. Any disagreement, unknown, or protocol failure triggers C; otherwise a precommitted 10% audit of unique review items can trigger C. C scores all criteria, blind to A/B votes. Audit identity binds input, answer, rubric, and protocol, not method. Duplicate review sharing requires identical review-relevant identities; repeatability probes bypass sharing.

This design is not an integrated V5 deployment: the offline V5 semantic aggregator requires exactly three distinct role, model, and family identities whenever semantic criteria exist, and cannot consume an untriggered two-vote item. A new production protocol must resolve this incompatibility, not invent a third vote. Different prompts on the same model do not provide independent judges. Reusing one panel for requirements and assessment can preserve shared errors.

For frozen critical criteria $\mathcal{K}_i$, task success is

$$Y_{ipk} = \begin{cases} 0, & \text{some critical criterion fails,} \\ 1, & \text{all critical criteria pass,} \\ ?, & \text{otherwise.} \end{cases} \tag{5}$$

Missing critical task-completion or an empty criterion set invalidates the rubric rather than creating vacuous success. A definite failure can determine the outcome despite another unknown. The denominator is one assigned answer, not a vote or criterion. Objective checks are source-backed, predeclared routes, not extra independent votes; locating a cited span does not establish entailment. No applicable constraint is not applicable, whereas an applicable unresolved constraint is unknown. Neither is silently zero.

Qualification. The v1 JudgeMeta design uses eight mechanisms with 30 probes each, split by ancestry into 120 development and 120 acceptance probes. Qualification probes are not additional benchmark tasks. Objective expectations require demonstrable sources; metamorphic probes support consistency only. V1 engineering targets and V4N/V5 gates are distinct versions. Acceptance items used for tuning cease to be hidden. Per-role/per-mechanism coverage, unknowns, invalidity, error detection, benign preservation, and audit disagreement must be reported. The production panel is not qualified; no human-agreement claim follows from synthetic tests or panel agreement.

5.2 Paired estimands, uncertainty, and claim gates

For policy p and reference p_0, the proposed request-weighted effect is $\Delta_Q = \sum_{i,k} w_{ik}(Y_{ipk} - Y_{ip_0k})$, with nonnegative weights summing to one. Equal-replicate designs use equal unit weights; session-balanced analysis is a different sensitivity estimand. For known success/failure/unknown, use $[L, U] = [1, 1]/[0, 0]/[0, 1]$. Retaining all assigned units gives

$$[\Delta_L, \Delta_U] = \left[\sum_{i,k} w_{ik}(L_{ipk} - U_{ip_0k}), \sum_{i,k} w_{ik}(U_{ipk} - L_{ip_0k}) \right]. \tag{6}$$

This identification interval is not a sampling confidence interval. Numerical bounds cannot validate an unqualified quality release. Observed-only means are secondary coverage diagnostics. Constraint comparisons use the same pre-answer applicable-unit mask in both conditions.

With complete usage and positive reference total, token saving is

$$S_T = 1 - \frac{\sum_{i,k} w_{ik} T^{\text{native}}_{ipk}}{\sum_{i,k} w_{ik} T^{\text{native}}_{ip_0 k}}, \tag{7}$$

not the mean of per-answer savings percentages. Paired resampling keeps all conditions and repeats in a sampled session/verified-lineage component and recomputes this ratio. The inspected comparator clusters by session only; cross-session lineage and partially identified outcomes need explicit extensions. Fewer than two clusters yields no bootstrap CI; a small cluster count does not assure adequate coverage or power.

Non-inferiority requires a suitable lower confidence bound on Δ_Q above a predeclared $-\delta_Q$, conservatively using Δ_L under unknown outcomes. Savings separately require a positive lower bound on S_T, and constraint safety an upper difference bound within a predeclared tolerance. A joint claim requires all three and valid releases. The existing comparator uses Bonferroni allocation across declared pairs and two-sided percentile intervals; margins, alpha, family, tail conventions, and bootstrap settings remain E19. A CLI example margin is not adopted. Nonsignificance is not non-inferiority.

The statistical freeze must also identify the population to which a claim applies. Development and exposed Historical observations remain exploratory or diagnostic. Confirmation on unseen conversations additionally needs a verified Fresh freeze and result-independent sample-size decision. Cluster resampling preserves observed dependence; it does not transform an exposed convenience sample into a representative or unseen population. Reporting a session-balanced sensitivity estimate cannot replace the request-weighted primary estimand without acknowledging the change. Endpoint uncertainty, constraint applicability, and failed release gates remain visible even when an observed subset yields a favorable mean.

6 Available Evidence and Pending Results

The source layers answer different questions. B0 supports input profiles and candidate counts. A later development report describes 40 rubrics and re-evaluation of 400 previously generated answers: 40 targets, five conditions (None, Full, Last1, Recent25, BM25-25), two generators, and 34 sessions. It is neither a new formal run nor evidence for all eight interfaces. The report marks formal-input readiness, panel validation, and formal-publication eligibility false.

A read-only audit found 192 invalid votes among 1,049: A 26/400, B 133/400, and C 33/249. All 62/400 unknown task outcomes involved critical protocol errors, not simply irreducible semantic ambiguity. Of 172 null constraint outcomes, 110 were inapplicable and 62 applicable but unresolved. These distinctions preclude treating null as zero or dropping unknown rows.

Table 3 summarizes the later qualification. The 120-item V4N qualification reports task-level expected agreement of 120/120, 89/120, and 117/120 for A/B/C; their role gates are pass/fail/fail. Aggregated panels nevertheless achieve 120/120 task decisions. That agreement does not override the failed evidence-binding gate: the overall status remains `LEN_GATE_FAIL`, with quality release ineligible. A also has ten field-level judgment mismatches despite perfect task agreement. The 120 qualification items are not the 120 Development candidates.

Appendix E retains the complete supplied qualification counts, invalidity breakdown, old identification intervals, and source-reported answer-token comparisons, including unfavorable evidence. Those accounting comparisons exclude a complete policy ledger and do not establish deployment savings. V5 adds offline implementation and source-reported tests, not a new authorized qualification or a rerun of the old answer matrix.

No accepted paired formal analysis, budget curve, controlled routing ablation, slice result, deployment overhead trace, or authorized real case is supplied. Empty figure interfaces are therefore not plotted as observations or zeros. Those formal results remain unavailable. Quality preservation, method rankings, and unseen generalization remain unestablished.

Table 3: V4N role-only diagnostics. Objective errors, positive items, and pairs have distinct denominators; task agreement does not clear the failed overall evidence-binding gate.

Measure	A	B	C
Task expected exact	120/120	89/120	117/120
Format valid	120/120	89/120	117/120
Objective errors detected	45/45	37/45	43/45
Positive items preserved	75/75	52/75	74/75
Both items in pair exact	60/60	36/60	57/60
Role gate	Pass	Fail	Fail

7 Limitations

The present evidence supports a task contract, framework interfaces, historical profiles, and adverse development diagnostics, not the intended empirical contribution. All imported inputs are exposed; Fresh and validated Challenge are absent. Static replay excludes method-dependent dialogue evolution, while within-session and lineage dependence limit effective sample size. Structural verification and reported permission flags are not independent provenance or release authorization. Release scope remains unconfirmed.

Rubric availability in a later development report does not validate B0 labels or resolve failed judge qualification. Full-history access, structured output, and agreement do not establish semantic correctness. Independent human calibration is unavailable (E13). Reference-token reduction is not native-usage reduction, and native tokens across tokenizers do not imply equal compute or cost. Missing calls, usage, run identities, and live traces block full reproduction and deployment claims (E11). The anonymous companion reproduces supplied aggregates and typesetting, not private histories, model generations, or qualified labels.

8 Conclusion

SCM-Bench frames context efficiency as a target-turn quality–resource question under explicit history, routing, and accounting boundaries. The available historical profile and implementation specify what a controlled comparison must preserve; development diagnostics demonstrate why qualification cannot be replaced by agreement. No qualified empirical quality–efficiency conclusion is yet supported. Until the evidence and release requirements are resolved, the manuscript is an internal scientific draft, not a completed quality–efficiency evaluation.

AI Use Statement

AI assistance was used in this manuscript workflow for research-material synthesis, draft integration and wording, organization of methods and evaluation descriptions, figure/table preparation, LaTeX construction, and generation of offline checking and reproduction scripts. This includes AI-assisted figure design, not only proofreading. This integration used local aggregate checks, source inspection, compilation, and rendered-page review; it did not run model inference or new experiments. These checks are not independent human scientific validation. Authors remain responsible for the complete AI-use history and all claims, citations, and artifacts.

Ethics Statement

Historical conversations may contain private or sensitive information. Structural eligibility and an internal permission field do not establish permission to publish transcripts or linked records. The review companion therefore omits raw dialogues, answers, original identifiers, relinkable record hashes, service endpoints, and private repository locations, retaining only aggregate statistics, interface descriptions, prompt templates, and offline reproduction utilities. Underlying conversations and models are not claimed publicly redistributable; source categories are provenance, not demo-

graphic labels. Data rights, review requirements, release authorization, and relevant risks require author confirmation.

Reproducibility Statement

Supplementary Appendices A–G and the anonymous companion supply aggregate data, available prompt bytes, descriptive contracts, editable tables, and offline figure/checking code. They reproduce aggregate arithmetic and typesetting, not end-to-end experiments. Private inputs, executed request/usage ledgers, immutable model bindings, and a qualified quality release remain unavailable; Appendix F details the required handoff and author confirmations.

References

Zexue He, Yu Wang, Churan Zhi, Yuanzhe Hu, Tzu-Ping Chen, Lang Yin, Ze Chen, Tong Arthur Wu, Siru Ouyang, Zihan Wang, Jiaxin Pei, Julian McAuley, Yejin Choi, and Alex Pentland. MemoryArena: Benchmarking agent memory in interdependent multi-session agentic tasks, 2026. URL https://arxiv.org/abs/2602.16313v2.

Qitian Jason Hu, Jacob Bieker, Xiuyu Li, Nan Jiang, Benjamin Keigwin, Gaurav Ranganath, Kurt Keutzer, and Shriyash Kaustubh Upadhyay. RouterBench: A benchmark for multi-LLM routing system, 2024. URL https://arxiv.org/abs/2403.12031v2.

Yucheng Li, Huiqiang Jiang, Qianhui Wu, Xufang Luo, Surin Ahn, Chengruidong Zhang, Amir H. Abdi, Dongsheng Li, Jianfeng Gao, Yuqing Yang, and Lili Qiu. SCBench: A KV cache-centric analysis of long-context methods. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2412.10319v2.

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)*, pp. 13851–13870, 2024. doi: 10.18653/v1/2024.acl-long.747. URL https://aclanthology.org/2024.acl-long.747/.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs with preference data. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2406.18665v4.

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Menglin Xia, Xufang Luo, Jue Zhang, Qingwei Lin, Victor Rühle, Yuqing Yang, Chin-Yew Lin, H. Vicky Zhao, Lili Qiu, and Dongmei Zhang. LLMLingua-2: Data distillation for efficient and faithful task-agnostic prompt compression. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 963–981, 2024. doi: 10.18653/v1/2024.findings-acl.57. URL https://aclanthology.org/2024.findings-acl.57/.

Ved Sirdeshmukh, Kaustubh Deshpande, Johannes Mols, Lifeng Jin, Ed-Yeremai Cardona, Dean Lee, Jeremy Kritz, Willow Primack, Summer Yue, and Chen Xing. MultiChallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18632–18702, 2025. doi: 10.18653/v1/2025.findings-acl.958. URL https://aclanthology.org/2025.findings-acl.958/.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. JudgeBench: A benchmark for evaluating LLM-based judges. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.12784v2.

Bo Wang, Yuqian Yao, Enxi Wang, Luozhijie Jin, Yang Liu, Yiran Suo, Yuxuan Cai, Enyu Zhou, Yufei Gao, Honglin Guo, Tianyu Huai, Li Ji, Zhikai Lei, Bufan Li, Lizhi Lin, Jinxiu Liu, Jie Yang, Jiazheng Zhou, Maosen Zhou, Pengfang Qian, Shichun Liu, Guanshan Liu, Hao Zheng,

Yunhao Yu, Hang Yan, Jihua Kang, Xinchi Chen, and Xipeng Qiu. ContextWeave: A real-world workflow benchmark, 2026. URL https://arxiv.org/abs/2608.04830v1.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.10813v2.