A Graph-Grounded Benchmark for Structural Long-Context Understanding, with a Strong Reference System

Anonymous ACL submission

Abstract

Needle-in-a-haystack (NIAH) probes test whether a model can *locate* a token, not whether it has built a coherent *structural memory*. We introduce **SLoCoB**, a graph-grounded benchmark for structural long-context understanding: 1,842 naturalistic documents (32K–512K tokens) from books, legal contracts, scientific papers, meeting transcripts, and code repositories; audited document structure graphs with full-pass review of high-stakes (CONT, CAUS, DEP) edges (§4.3); and five task families (entity timeline, causal chain, argument structure, cross-section dependency, latent contradiction). SLoCoB includes a hidden test split, hard-negative compositional distractors, an unseen-template subset, and human-written free-form questions. As a strong reference system we introduce the Structure-Aware Context Planner (SACP), which maintains an independently extracted structure graph, plans retrieval over it, and applies a rule-based consistency check at answer time. Across seven baseline families (nine raw systems), matched preprocessing budgets, and three shared-graph fairness controls, SACP improves Structural Average by $+10.6$ over the best raw baseline (GraphRAG) and by $+6.0$ over the strongest shared-graph control, while NIAH remains near-saturated ($\geq$ 97.5%) for every system. SACP's gains transfer to two external long-context benchmarks. We position SACP as a reference architecture; the benchmark and protocol are the primary deliverable.

1 Introduction

The past two years have seen a race to extend the effective context window of large language models from 32K tokens to 1M and beyond (Gemini Team, 2024; Anthropic, 2024; Ding et al., 2024). Progress is almost universally reported in terms of the needle-in-a-haystack (NIAH) probe (Kamradt, 2023): a random sentence (the *needle*) is injected into an unrelated long text (the *haystack*), and the model is scored on whether it can retrieve that sentence. Under this lens, frontier systems report near-saturated scores at their advertised maximum windows (GPT-4o 128K, Claude-3.7 up to 200K and Gemini-1.5 up to 1M) (OpenAI, 2024; Anthropic, 2025; Gemini Team, 2024).

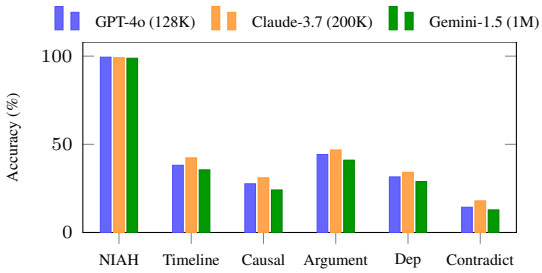

Figure 1: Frontier long-context models saturate NIAH ($\geq$ 98.8% for the three systems shown; $\geq$ 97.5% across all systems in Table 1) but collapse on the five structural probes (T1 timeline, T2 causal chain, T3 argument structure, T4 cross-section dependency, T5 latent contradiction) from our SLoCoB benchmark. **Needle is not enough.**

This progress is partial: practitioners need models to read a book, contract, paper, meeting log or code repository and reason over the *structure* tying its tokens together. NIAH asks "is the needle there?". Real questions ask who promised what, which later clause contradicts it, which result depends on which assumption, or whether `write_batch` respects its declared retry budget. These are *structural* questions. Figure 1 shows that models that ace the needle routinely collapse on them.

Why needle fails as a proxy. Three reasons. (i) *Locality*: NIAH depends on one span, so even shallow retrieval can succeed; cross-span composition is untested. (ii) *Surface distinctness*: the needle is dissimilar from the haystack, unlike real evidence and distractors. (iii) *Stateless queries*: NIAH does not require a coherent state that changes through a

book, contract or code base.

Contributions. We frame this paper as a benchmark paper with a reference system:

- **Benchmark (primary).** SLoCoB: 1,842 naturalistic documents and 24,610 questions, built from automatically extracted, human-audited structure graphs with full-pass review of high-stakes (CONT, CAUS, DEP) edges, plus hidden test, hard negatives, *unseen-template* and human-written free-form subsets.
- **Formalisation.** We define $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \tau, \sigma)$ with typed nodes/edges and frame structural competence as graph-conditioned prediction via $\hat{\mathcal{G}}$.
- **Reference system.** SACP maintains an independently extracted $\hat{\mathcal{G}}$, plans retrieval over it, and applies a rule-based consistency check; it is a reproducible target, not a claimed methodological breakthrough.
- **Fairness controls.** We use matched preprocessing budgets, three shared-graph conditions where baselines consume the same $\hat{\mathcal{G}}$, and two leakage-resistant slices (Tables 1, 2).

Under these protocols, SACP improves Structural Average by $+10.6$ over GraphRAG and by $+6.0$ over GraphRAG+$\hat{\mathcal{G}}$; gains transfer to Michelangelo (Vodrahalli et al., 2024) and LongBench v2 (Bai et al., 2025) (Table 3). NIAH stays $\geq 97.5\%$, confirming its decoupling from structural competence.

2 Related Work

Long-context evaluation. NIAH (Kamradt, 2023) and its variants (Hsieh et al., 2024; Bai et al., 2024, 2025; Zhang et al., 2024b; Shaham et al., 2023; Yuan et al., 2024) pushed context-length testing into public view. RULER (Hsieh et al., 2024) adds synthetic variable tracking; LongBench and LongBench v2 (Bai et al., 2024, 2025) aggregate QA and summarisation; ∞Bench (Zhang et al., 2024b) stresses book-length inputs; BAMBOO (Dong et al., 2024) and LooGLE (Li et al., 2024) emphasise time-sensitive reasoning. The most closely related post-NIAH efforts are Michelangelo (Vodrahalli et al., 2024), which defines latent-structure queries over synthetic code/list/contextual graphs, NoLiMa (Modarressi et al., 2025), which suppresses literal matching, and Logic Haystacks (Sileo, 2026), which stress-tests logical chaining. SLoCoB complements this line in three ways: it is grounded in naturalistic documents of up to 512K tokens, it releases an audited *typed* structure graph with CAUS/SUPP/CONT/DEP edges as supervision (not only as a probe), and its questions are generated from graph patterns rather than from hand-written synthetic patterns.

Long-context modelling. Techniques include sparse attention (Beltagy et al., 2020; Zaheer et al., 2020), state-space and linear-recurrence models (Gu and Dao, 2024; Peng et al., 2023), length extrapolation (Su et al., 2024; Peng et al., 2024; Ding et al., 2024), landmark/compression memories (Mohtashami and Jaggi, 2023; Zhang et al., 2024a; Wang et al., 2024), and retrieval-augmented generation (Lewis et al., 2020; Izacard et al., 2023; Asai et al., 2024); production systems combine them (Llama Team, 2024; Gemini Team, 2024; Anthropic, 2025; OpenAI, 2024). We treat them as baselines and show that structural probes expose a shared deficit.

Memory-augmented reading and graph-based retrieval. Schema-based readers (Févry et al., 2020; Wang et al., 2023), episodic memory (Zhong et al., 2024; Packer et al., 2023), and agentic readers (Lee et al., 2024; Sun et al., 2024) introduce external memory. A parallel line of *graph-based retrieval* has emerged with GraphRAG (Edge et al., 2024), HippoRAG (Gutiérrez et al., 2024) and LightRAG (Guo et al., 2025). We include GraphRAG as a baseline (and as a shared-graph fairness control B6$^+$, Table 1). SACP differs from these systems on three concrete axes. First, the graph itself: GraphRAG indexes *community summaries*, HippoRAG indexes *symmetric associations* scored by personalised PageRank, LightRAG indexes *dual-level entity/relation pairs*; SACP indexes a *typed, directional* graph with temporally and argumentatively typed edges (PREC, CAUS, SUPP, CONT, DEP) that the consistency check can interrogate predicate-by-predicate. Second, the reader: HippoRAG's PageRank is symmetric so a CAUS edge $u \to v$ is indistinguishable from $v \to u$ at retrieval time, which is sufficient for entity-centric QA but not for T2 chain reconstruction; SACP retains direction. Third, retrieval–generation coupling: none of the three couples the generator to a structural predicate evaluator at answer time; SACP's $\mathcal{O}(a, q, \hat{\mathcal{G}})$ does. The shared-graph control B6$^+$ isolates the first axis, the pattern-oracle B6$^{+\dagger}$ isolates direc-

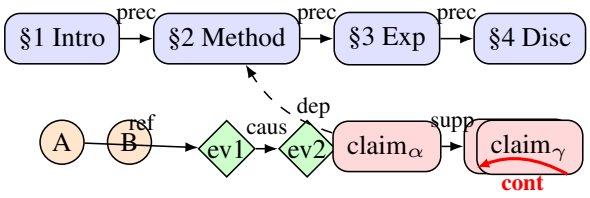

Figure 2: Document structure graph: typed nodes (section, entity, event, claim) and edges (*precedes*, *refers-to*, *causes*, *supports*, *contradicts*, *depends-on*). A latent contradiction (red) spans §3 and §4.

tion and pattern access, and the gap that remains attributes to the predicate evaluator (§6, §6.5).

Discourse and structure. Long-document understanding has roots in discourse parsing (Mann and Thompson, 1988; Prasad et al., 2008) and argument mining (Lippi and Torroni, 2016). We modernise these signals into a unified structural supervisor for long-context LLMs.

3 Formalising Structural Long-Context Understanding

Document structure graph. Given a document $D = (t_1, \ldots, t_L)$ with L up to 5×10^5 tokens, we define its structure graph

$$\mathcal{G}(D) \;=\; (\mathcal{V},\, \mathcal{E},\, \tau,\, \sigma), \qquad (1)$$

where $\mathcal{V}$ contains typed nodes $\mathcal{T}_V = \{\text{SEC}, \text{ENT}, \text{EVT}, \text{CLM}, \text{DEF}\}$, $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V} \times \mathcal{T}_E$ has edge types $\mathcal{T}_E = \{\text{PREC}, \text{CAUS}, \text{SUPP}, \text{CONT}, \text{REF}, \text{DEP}\}$, τ labels nodes, and σ anchors nodes/edges to span(s). Figure 2 illustrates.

Structural understanding as approximate graph-conditioned inference. Writing the ideal predictor as a marginal over latent graphs,

$$p(a \mid q, D) \;=\; \sum_{\mathcal{G}'} p(a \mid q, \mathcal{G}')\, p(\mathcal{G}' \mid D), \qquad (2)$$

structural competence is the ability to approximate this with a tractable point estimate $\hat{\mathcal{G}}(D)$: $p(a \mid q, D) \approx p(a \mid q, \hat{\mathcal{G}}(D))$. The formalisation motivates §6's shared-graph controls, where baselines consume the same $\hat{\mathcal{G}}$, and SACP's streaming cost factorisation $|D| + |q| \cdot |\text{plan}|$, linear in L and independent of the base window (§5). Models that collapse onto local retrieval $p(a \mid q, t_{i:j})$ can pass NIAH but fail multi-node inference. The next two sections develop this: §4 anchors gold $\mathcal{G}$, §5 instantiates $\hat{\mathcal{G}}$.

Four competences. We decompose structural understanding into: **(C1)** *entity state tracking* (value of an entity slot at position t); **(C2)** *causal chain reconstruction* (find $v_1 \to \ldots \to v_k$ via CAUS edges); **(C3)** *argument structure* (classify SUPP/CONT over claims); **(C4)** *cross-section dependency* (detect DEP between non-adjacent sections). Latent contradiction detection is a special case of C3 conditioned on temporal order.

4 The SLoCoB Benchmark

4.1 Design principles

(1) *Naturalistic length*: up to 512K tokens without padding. (2) *Structural gold*: every question is grounded in a sub-graph of $\mathcal{G}(D)$, not a span. (3) *Compositional distractors*: lexically similar false positives defeat keyword shortcuts. (4) *Multi-genre*: results should not overfit one domain.

4.2 Corpora

Books: 412 public-domain novels and 38 non-fiction works from Project Gutenberg (64K–410K tokens). **Legal contracts**: 520 agreements from SEC EDGAR (U.S. Securities and Exchange Commission, 2024) and CUAD (Hendrycks et al., 2021) (32K–180K). **Scientific papers**: 310 arXiv/OpenReview papers and reports (48K–220K). **Meeting transcripts**: 200 AMI (Carletta et al., 2006)/ICSI (Janin et al., 2003) transcripts plus 60 public parliamentary/council transcripts (70K–320K). **Code repositories**: 302 permissively licensed GitHub repositories (SPDX Workgroup, 2026), flattened with file-boundary tokens (80K–512K). Total 1,842 documents.

4.3 Annotation pipeline and question generation

For each document we extract $\mathcal{G}(D)$ through: *(i)* genre-tuned section segmentation; *(ii)* coreference/event extraction by majority vote over SpanBERT (Joshi et al., 2020) fine-tuned for Lee et al. (2017), DyGIE++ (Wadden et al., 2019) retrained on ACE-2005, and a GPT-4o-mini zero-shot extractor, with human review of disagreements; *(iii)* two-annotator claim extraction and SUPP/CONT labelling, with contradiction spans $\geq 20\%$ of document length apart; and *(iv)* senior adjudication plus CAUS/DEP edges.

What "audited" means. The release graph contains $|\mathcal{V}| \approx 7.69 \times 10^5$ accepted nodes and $|\mathcal{E}| \approx 1.83 \times 10^6$ accepted edges across 1,842 documents

(per-genre counts in Appendix A, Table 4). Full double-annotation is infeasible, so we use a stratified hybrid protocol. *High-stakes edges* (HS-E: CONT, CAUS, DEP; $\approx 11\%$ of $|\mathcal{E}|$, dominating T2/T4/T5) are fully reviewed by two annotators with senior adjudication, covering 204,912 edges. *Lower-stakes items* (LS: nodes plus non-HS edges) receive (i) a stratified audit over accepted LS items—books 0.22, contracts 0.25, sci 0.25, meetings 0.21, code 0.07, with ≥ 50 items per genre stratum—and (ii) human confirm-reject review of every auto-rejected candidate. Code coverage is lower because most code edges are AST-derived DEP/REF edges with > 0.95 precision; residual event-style CAUS edges are HS-E. The 34,918 hand-corrected code edges combine this HS-E residual and human re-accepted auto-rejections, explaining the 0.62 acceptance rate. LS audit plus confirm-reject yields 407,568 LS decisions; with HS-E, 612,480 items are individually decided. Remaining auto-accepted LS items are batch-decided from LS-sample error rates. *Residual quality*: post-adjudication LS error is $0.4\%/0.5\%$ (nodes/edges); Wilson 95% upper bounds for un-audited residual error are $\leq 0.9\%/\leq 1.0\%$. Span anchors are spot-checked at 5%. Automatic-extractor P/R on 50 documents: SpanBERT-coref $0.79/0.74$, DyGIE++ event $0.71/0.66$, GPT-4o-mini event $0.66/0.69$. Per-genre acceptance: books 0.74, contracts 0.81, sci 0.79, meetings 0.69, code 0.62.

Annotator pool, hours, throughput. Eleven paid annotators completed 6-hour calibration; only those with $\alpha \geq 0.70$ were retained. Production logged 3,240 annotator-hours. Double-annotation on 204,912 HS-E items gives $204{,}912 \times 2 + 407{,}568 = 817{,}392$ single-annotator judgements, or $\approx$252/h. External audit (50 questions/task) gives agreement $0.93/0.89/0.86/0.82/0.84$ for T1–T5. Krippendorff α (Hayes and Krippendorff, 2007) is 0.78 (claim) and 0.71 (causal); full reconciliation is in Appendix A.

Question generation and leakage control. Questions are template-generated over the gold graph, rewritten by one annotator, and independently answered by a second from the document alone; disagreements are discarded. Leakage controls: (a) gold has finer sub-types that SACP never accesses (e.g. CAUS$\in$\{ENABLE, CAUSE, BLOCK\}); (b) SACP's extractor runs zero-shot, with prompt templates chosen on external corpora (EntityQuestions (Sciavolino et al., 2021), PDTB (Prasad et al., 2008), SciArg (Lauscher et al., 2018)) but no fine-tuning; (c) test gold is withheld on Codabench (Xu et al., 2022). §6.5 reports unseen-template, free-form, and paraphrase splits; Appendix D shows gold sub-types add only $+1.3$.

T1 Entity Timeline. Given entity e, list all state changes in temporal order.

T2 Causal Chain. Reconstruct $v_1 \rightarrow \ldots \rightarrow v_k$ linking event v_1 to outcome v_k.

T3 Argument Structure. For claim pair (c_i, c_j), label SUPP, CONT, or IND.

T4 Cross-Section Dependency. Identify DEP edges between non-adjacent sections.

T5 Latent Contradiction. Detect contradictions separated by $\geq$20% of document length.

Figure 3: Five memory-probing task families in SLoCoB. Each is decomposable into operations over $\mathcal{G}(D)$.

Compositional distractors. Every gold fact is paired with three hard negatives: (i) a lexically similar non-entailing span (≥ 0.85 cosine), (ii) a same-document different-section span, and (iii) a same-surface-entity span from another document. Removing distractors raises all systems 8–12 points; with them, local retrieval trails structural reasoning by ≥ 25 points.

Splits. 70/15/15 train/dev/test at the document level, stratified by genre and length bucket. Test documents never overlap with train/dev at the source-URL level.

4.4 Memory probing tasks

Figure 3 summarises the five task families.

Each task ships 4,000–6,000 questions, balanced across genres; per-task counts are T1= 4,904, T2= 5,108, T3= 5,412, T4= 4,608, T5= 4,578, totalling 24,610 questions matching the abstract. Test split sizes are $735/766/812/691/687$ for T1–T5 ($\approx 15\%$ of each task). Hard negatives (§4.3) are paired *labelled candidate answers*, not separate scored questions: the 24,610 figure counts only positive question-answer pairs that are scored end-to-end. Answers are structured: a timeline is a list of (t, state) tuples, a causal chain is a sequence of event IDs, T3–T4 are classification tasks, and T5 is a span-pair classification task. The *headline per-task score* is F1 in all cases: exact-set F1 for T1 and chain F1 for T2 (both against gold sequences); macro-F1 for T3 and T4; and pair-level F1 at a fixed operating point for T5. We additionally report PR-AUC for T5 in Appendix B. Structural

Average in all tables is the unweighted mean of these five F1 scores.

5 Structure-Aware Context Planner

SACP streams D into an incremental structure graph $\hat{\mathcal{G}}_t$, plans retrieval over it, and couples generation with a rule-based consistency check.

5.1 Overview

Let $\{B_1, \ldots, B_m\}$ be a chunking of D with $|B_i| = b$ (default $b = 4096$). SACP alternates three phases: READ, PLAN, ANSWER. In READ, a structured-extraction prompt converts chunk B_i into nodes/edges that update $\hat{\mathcal{G}}_i$ via embedding deduplication. In PLAN, q becomes sub-queries $\{q_1, \ldots, q_k\}$ mapped to graph patterns π_j. In ANSWER, retrieved spans are assembled and the consistency check verifies the candidate against $\hat{\mathcal{G}}$.

5.2 Incremental structure graph

The graph is a labelled property hypergraph with an HNSW (Malkov and Yashunin, 2020) index (M=16, efConstruction=200) over `bge-small-en-v1.5` (BAAI, 2023) node embeddings. Insert cost is $\mathcal{O}(\log |\mathcal{V}|)$. We merge same-type nodes at cosine similarity ≥ 0.92 and union their span anchors. To bound memory, every 64 chunks we evict leaf nodes in the bottom 5% of normalised PageRank, following Zhong et al. (2024), but never nodes incident on CONT edges. $\hat{\mathcal{G}}$ **is produced only from raw tokens; SACP never sees gold $\mathcal{G}$.**

5.3 Retrieval–generation coupling

SACP couples retrieval and generation through a task-aware *rule-based consistency check* $\mathcal{O}$ over $\hat{\mathcal{G}}$:

$$\mathcal{O}(a, q, \hat{\mathcal{G}}) \;=\; \mathbb{1}\big[\phi_{\text{time}}(a) \wedge \phi_{\text{caus}}(a) \wedge \phi_{\text{arg}}(a, q)\big], \tag{3}$$

where ϕ_{time} asserts temporal monotonicity along events referenced by a, ϕ_{caus} asserts reachability along CAUS edges between a claimed cause and its effect, and $\phi_{\text{arg}}(a, q)$ is task-aware. For *answer-consistency* tasks (T1–T4) it requires that no CONT edge holds *among the claims cited in support of* a. For the *contradiction-detection* task T5, ϕ_{arg} is conditioned on the answer polarity to avoid systematically biasing negatives toward false positives: when $a =$CONTRA we require at least one retrieved CONT edge matching the queried pair (presence); when $a =$NOT-CONTRA we require that no CONT edge between the queried pair appears anywhere in $\hat{\mathcal{G}}$ within the planner's current retrieval radius (absence at the radius). This is bidirectional answer↔evidence consistency rather than an unconditional presence check, so a re-plan on a negative answer expands radius without forcing the planner to fabricate a CONT edge. When $\mathcal{O} = 0$ we re-plan: the sub-query responsible for the failed predicate is reissued with the retrieval radius doubled, capped at three rounds.

5.4 Algorithmic sketch

Input: document D, query q.
Stream chunks $B_1, \ldots, B_m$; for each, extract $(\Delta V, \Delta E)$ and update $\hat{\mathcal{G}}$.
Plan: decompose $q \to \{(q_j, \pi_j)\}$; execute patterns over $\hat{\mathcal{G}}$ to fetch evidence E_j.
Answer: generate a conditioned on $\{E_j\}$; call $\mathcal{O}(a, \hat{\mathcal{G}})$.
If $\mathcal{O} = 0$, expand the failing sub-query's retrieval radius by $\times 2$ and re-answer (max 3 rounds).

5.5 Complexity and accounting

Cost has two phases. *Indexing*: one base-model pass per chunk, $m = \lceil L/b \rceil$ calls (e.g. 49 for $L = 200$K, $b = 4096$). *Querying*: plan/answer adds $\mathcal{O}(k)$ calls for $k \leq 4$ sub-queries. Costs are *amortised per query*: indexing cost divided by 13.4 queries/document. Total cost is linear in L and independent of the base window.

6 Experiments

6.1 Baselines and budgets

We evaluate seven baseline families instantiated as nine raw systems, plus SACP. (B1) **Long-context LLM**: GPT-4o (`gpt-4o-2024-11-20`, 128K (OpenAI, 2024)), Claude-3.7 (`claude-3-7-sonnet-20250219`, 200K (Anthropic, 2025)), Gemini-1.5 (`gemini-1.5-pro-002`, up to 1M (Gemini Team, 2024)); $T = 0$, JSON-mode. (B2) **Dense RAG**: BGE-M3 (Chen et al., 2024) top-16 plus `bge-reranker-v2-m3` (BAAI FlagEmbedding Team, 2024). (B3) **Sliding window**: 8K windows with high-attention re-read. (B4) **Hierarchical summarisation**: depth-3, fan-out 8. (B5) **Memory agent**: MemGPT-style (Packer et al., 2023), 4-tier memory, 8K recall. (B6) **GraphRAG** (Edge et al., 2024), community level 2, top-8. (B7) **HippoRAG** (Gutiérrez et al., 2024). **SACP** uses Claude-3.7; SACP* uses Gemini-1.5.

Matched preprocessing budget. Every baseline gets SACP's *per-document* indexing budget: one

base-model pass per 4K chunk, with ingested tokens within $\pm 5\%$ (for 200K tokens: SACP 208K, GraphRAG 209K, MemGPT 211K, Hier. summ. 214K, Dense RAG 206K, Sliding cache 200K; HippoRAG reuses the dense index). Decoding, prompts, chunking, versions and schemas are fixed and released; token tables are in the supplementary archive.

Shared-graph fairness and diagnostics. $B2^+/B5^+/B6^+$ consume the same independently extracted $\hat{\mathcal{G}}$ as SACP plus raw text. $B6^{+\dagger}$ (Table 2) also receives SACP's task-pattern hint, isolating the consistency check. We also report SACP+gold $\mathcal{G}$, SACP−graph, and SACP+noisy $\mathcal{G}$ (30% edge-flip).

Hidden test submission. Numbers are on the held-out test split via Codabench: participants submit answer JSON; the server scores against unreleased gold.

6.2 Metrics and cost accounting

Structural F1 follows §4; **Structural Average** is the unweighted mean. **Timeline Kendall-τ** scores T1 order. **Context efficiency** is Structural Average per 1K input tokens. **\$ per query** includes indexing amortised over 13.4 queries/doc plus per-query cost; **\$ per correct answer** divides total \$ by fully-correct T1–T5 answers. The 13.4 mean is 24,610/1,842 ($\sigma = 4.7$); Appendix D reports 5 queries/doc sensitivity and cost decomposition. Table 1 averages seeds $\{0, 1, 2\}$ with deviation ≤ 0.7.

6.3 Main results

Table 1 shows NIAH uniformly $\geq$ 97.5% while Structural Average spans 26.1–49.8. SACP beats GraphRAG (39.2) by +10.6 and GraphRAG+$\hat{\mathcal{G}}$ (43.8) by +6.0. Diagnostics show graph access is necessary (−16.7 to 33.1), graph noise is tolerable (−5.4 at 30% edge-flip), and gold graphs leave +13.0 headroom. Per-task 95% bootstrap CIs are $\leq \pm 1.1$; SACP-vs-$B6^+$ CI is $[+4.8, +7.2]$.

6.4 Context length vs. structural accuracy

Figure 4 plots structural accuracy vs. length. Long-context LLMs lose 24–30 points from 32K to 512K; SACP loses 8.8, with hierarchical summarisation between them. This measures evaluated long-input policies under matched budgets, not native model competence: truncation/summarisation degrades more sharply than persistent structural memory. No-truncation policies remain open.

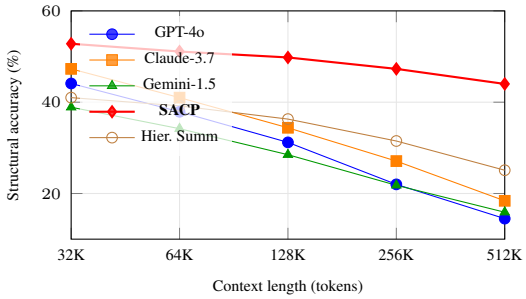

Figure 4: Structural accuracy vs. document length. For inputs beyond a model's native window we use the provider-recommended long-input policy: middle-out truncation for GPT-4o and Claude-3.7, native ingestion for Gemini-1.5 up to 512K; each model receives the same total token budget its API supports plus the matched preprocessing budget (§6). The comparison therefore measures *evaluated long-input policies* (truncation vs. streaming with persistent structural memory), not internal model capability. Under this protocol, systems without persistent structural memory degrade more sharply with length than SACP, whose cost is linear in L and independent of the base window.

6.5 Leakage-resistant evaluation

Because SLoCoB is template-generated, we test four leakage-resistant slices. **Unseen-template**: 1,284 test questions from templates absent in train/dev. **Free-form human**: 1,050 document-written questions with no graph/template access; a second annotator writes T1–T5 gold and verifies graph alignment, discarding failures. **Paraphrase stress**: three rewrites per question, reporting mean and worst. **Pattern-oracle**: $B6^{+\dagger}$ receives SACP's task-pattern hint to isolate the consistency check.

Table 2 preserves SACP's lead under unseen templates, free-form questions, and paraphrase noise.

6.6 External-benchmark validation

Table 3 shows transfer beyond SLoCoB but smaller gaps: +6.0 on SLoCoB, +3.3 on Michelangelo, +2.3 on LongBench v2 ($p < 0.01$, paired bootstrap), indicating partial benchmark coupling.

6.7 Ablations, cost, and failure analysis

Ablations (Structural Average; full Table 5 in App. D): full SACP 49.8; −consistency-check −6.7; −plan-decomposition −8.3; −typed-edges (PREC-only) −11.8; −contradiction-preserving evict −4.4. Typed edges matter most; plan decomposition mainly affects T2/T4; consistency affects T5 (−12.1). Gemini-1.5 persistence (−2.0) suggests gains are not backbone-specific. *Cost.*

System	NIAH	T1 Time	T2 Causal	T3 Arg	T4 Dep	T5 Contra	Struct. Avg	$/Q
Seven baseline families (nine raw systems):								
B1 GPT-4o (128K)	99.4	38.1	27.6	44.2	31.5	14.3	31.1	0.21
B1 Claude-3.7 (200K)	99.1	42.3	31.0	46.8	34.1	17.9	34.4	0.27
B1 Gemini-1.5 (1M)	98.8	35.5	24.1	41.0	28.9	12.8	28.5	0.33
B2 Dense RAG + reranker	97.5	40.2	29.8	43.5	33.0	15.7	32.4	**0.06**
B3 Sliding window (8K, re-read)	98.2	33.0	22.4	38.1	26.3	10.9	26.1	0.18
B4 Hier. summ. (depth-3)	97.8	45.7	33.2	47.9	35.5	19.1	36.3	0.09
B5 Memory agent (MemGPT)	98.4	48.9	36.5	49.6	37.8	22.5	39.1	0.15
B6 GraphRAG (Edge et al., 2024)	97.9	47.8	38.2	48.3	39.9	21.7	39.2	0.12
B7 HippoRAG (Gutiérrez et al., 2024)	97.7	46.3	36.7	47.5	38.6	20.9	38.0	0.11
Shared-graph fairness controls (B^+): same $\hat{\mathcal{G}}$ as SACP.								
$B2^+$ Dense RAG + $\hat{\mathcal{G}}$	97.6	46.4	35.1	47.2	39.4	20.6	37.7	0.11
$B5^+$ Memory agent + $\hat{\mathcal{G}}$	98.5	53.2	41.0	52.7	43.1	25.8	43.2	0.14
$B6^+$ GraphRAG + $\hat{\mathcal{G}}$	98.0	52.6	43.4	52.0	44.7	26.4	43.8	0.12
SACP (Claude-3.7, ours)	98.9	**60.4**	**49.7**	**58.2**	**48.6**	**32.1**	**49.8**	0.13
SACP (Gemini-1.5)*	98.6	58.1	47.2	56.4	46.9	30.4	47.8	0.14
Diagnostic rows (not deployable systems):								
SACP−graph (no $\hat{\mathcal{G}}$, RAG only)	97.8	41.3	30.4	44.0	33.7	16.1	33.1	0.07
SACP + noisy $\hat{\mathcal{G}}$ (30% flips)	98.5	54.7	44.1	53.1	43.9	26.2	44.4	0.13
SACP + gold $\mathcal{G}$ (ceiling)	99.0	74.5	63.8	69.4	61.0	45.2	62.8	–

Table 1: Main results on SLoCoB (hidden test). $/Q is the amortised total cost per query (indexing + query) at public API pricing (2026-04-15 snapshot). NIAH is near-saturated ($\geq 97.5\%$) for every system while Structural Average spans 26.1–49.8. With shared-graph access, the strongest baseline reaches 43.8 ($B6^+$); SACP leads it by $+6.0$ points. The diagnostic rows show that without a graph SACP collapses to 33.1, that it is robust to substantial graph noise (44.4 at 30% edge-flip), and that the gold-graph ceiling sits $+13.0$ above SACP, bounding extractor-induced headroom.

System	Unseen	Free-form	Para. worst
Claude-3.7	31.4	28.2	28.0
B6 GraphRAG	36.9	32.0	33.7
$B6^+$	41.5	35.4	39.0
$B6^{+\dagger}$ pat.-oracle	43.8	36.1	41.0
SACP	**47.1**	**40.6**	**45.2**

Table 2: Leakage-resistant slices (Structural Average). On unseen-template SACP loses -2.7 but still leads $B6^{+\dagger}$ by $+3.3$; free-form costs -9.2 yet stays $+5.2$ above the strongest baseline; worst-paraphrase loses -4.6.

System	Michelangelo	LongBench v2
Claude-3.7	51.2	47.6
B5 Memory agent	53.8	48.4
B6 GraphRAG	55.7	49.1
$B6^+$	57.6	50.0
SACP	**60.9**	**52.3**

Table 3: Out-of-distribution validation. Mean over Michelangelo (Vodrahalli et al., 2024) latent-structure tasks (9 subtasks, accuracy) and LongBench v2 (Bai et al., 2025) long-reasoning slices (multi-doc QA + code understanding, F1). Per-subtask numbers in Appendix D.

SACP $/Q 0.13 vs. Claude-3.7 0.27 ($\sim 3.0\times$ per correct answer); wall-clock 11.4 s vs. 18.9 s; at 1 query/doc the advantage flips (0.43 vs. 0.27). *Failures* (200 sampled): 38% extraction, 27% plan, 19% consistency, 16% reasoning.

7 Discussion and Conclusion

NIAH is a sanity check, not the headline; long-context evaluation should report NIAH *and* structural metrics. Shared-graph controls close about half the gap, so typed memory and SACP's planner/consistency check are additive. The benchmark and protocol are the primary deliverable; SACP is a reference to surpass.

Limitations

SLoCoB is English-only; extending to multilingual long documents requires parallel structural annotation we have not attempted. Structural gold annotation is produced through a model-assisted pipeline whose outputs are audited via a stratified protocol with full-pass review of high-stakes edges, but residual annotator bias is possible despite $\alpha = 0.78$ claim-edge agreement. SACP assumes the base model can produce reliable structured extractions; on open-source models <8B parameters, extraction quality drops and gains shrink by 4–6 points. The consistency check uses a small set of hand-

authored predicates; learning them end-to-end is future work. Cost numbers assume public API pricing snapshots from 2026-04 and will shift with provider updates; we report amortised per-query cost under a 13.4-queries/document assumption that does not hold for one-shot workloads. Finally, the SLoCoB test set is split into a permanently hidden frozen partition (60% of test) used only via the Codabench (Xu et al., 2022) server with rate-limited submissions and no per-question feedback, and a public-after-period rotating partition (40%) released annually with refreshed documents to deter overfitting. Train/dev are released openly; the leaderboard publishes only Structural Average and per-task F1, never per-question scores. As frontier models improve we expect absolute numbers to rise, and we commit to a yearly refresh of the rotating partition with at least 200 new documents per cycle.

Ethics Statement

All book corpora are drawn from Project Gutenberg and are confirmed public-domain in the United States. Legal contracts are retrieved from the public SEC EDGAR full-text search, where filings are released for public access by statute. Meeting transcripts are from AMI and ICSI under their research licences, and from parliamentary / municipal council meetings whose audio and transcripts are released under open-government regimes (we document source URL, licence and date for every document in the data sheet). *We do not include Zoom recordings or any content whose speakers have not explicitly consented to redistribution.* Code repositories are restricted to SPDX permissive licences (MIT, Apache-2.0, BSD-2/3-Clause). Annotators were recruited through a vendor and compensated above local minimum wage. Structural annotations may reflect annotator worldview, particularly for CONT labels; we release per-annotator labels to enable debiasing. Long-context models remain capable of fabricating coherent but false reasoning chains; the consistency check mitigates but does not eliminate this risk.

References

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. Technical report, Anthropic.

Anthropic. 2025. Claude 3.7 Sonnet and claude code. https://www.anthropic.com/news/claude-3-7-sonnet. API model id claude-3-7-sonnet-20250219; accessed 2026-04-15.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

BAAI. 2023. BAAI/bge-small-en-v1.5 model card. https://huggingface.co/BAAI/bge-small-en-v1.5. Accessed 2026-04.

BAAI FlagEmbedding Team. 2024. BGE reranker v2 m3 model card. https://huggingface.co/BAAI/bge-reranker-v2-m3. Accessed 2026-04.

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 Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3119–3137.

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 Annual Meeting of the Association for Computational Linguistics (ACL)*.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2006. The AMI meeting corpus: A pre-announcement. In *Machine Learning for Multimodal Interaction*, volume 3869 of *Lecture Notes in Computer Science*. Springer.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL*, pages 2318–2335.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. 2024. LongRoPE: Extending LLM context window beyond 2 million tokens. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Zican Dong, Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2024. BAMBOO: A comprehensive benchmark for evaluating long text modeling

capacities of large language models. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, pages 2086–2099.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph RAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Thibault Févry, Livio Baldini Soares, Nicholas FitzGerald, Eunsol Choi, and Tom Kwiatkowski. 2020. Entities as experts: Sparse memory access with entity supervision. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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

Albert Gu and Tri Dao. 2024. Mamba: Linear-time sequence modeling with selective state spaces. In *Proceedings of the Conference on Language Modeling (COLM)*.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2025. LightRAG: Simple and fast retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP*.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. HippoRAG: Neurobiologically inspired long-term memory for large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Andrew F. Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1):77–89.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. CUAD: An expert-annotated NLP dataset for legal contract review. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

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 *Proceedings of the Conference on Language Modeling (COLM)*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. 2003. The ICSI meeting corpus. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages I–364–I–367.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Greg Kamradt. 2023. Needle in a haystack – pressure testing llms. `https://github.com/gkamradt/LLMTest_NeedleInAHaystack`. Accessed 2026-05-19.

Anne Lauscher, Goran Glavaš, and Simone Paolo Ponzetto. 2018. An argument-annotated corpus of scientific publications. In *Proceedings of the 5th Workshop on Argument Mining*, pages 40–46. Association for Computational Linguistics.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Kuang-Huei Lee, Xinyun Chen, Hiroki Furuta, John Canny, and Ian Fischer. 2024. A human-inspired reading agent with gist memory of very long contexts. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2024. LooGLE: Can long-context language models understand long contexts? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology*, 16(2).

Llama Team. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yu A. Malkov and D. A. Yashunin. 2020. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):824–836.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A. Rossi, Seunghyun Yoon, and Hinrich Schütze. 2025. NoLiMa: Long-context evaluation beyond literal matching. *arXiv preprint arXiv:2502.05167*.

Amirkeivan Mohtashami and Martin Jaggi. 2023. Landmark attention: Random-access infinite context length for transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*.

OpenAI. 2024. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/. Accessed 2026-05-19.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. 2023. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, Kranthi Kiran G. V. Kiran, Xuzheng He, Haowen Hou, Jiaju Lin, Przemyslaw Kazienko, Jan Kocon, Jiaming Kong, Bartlomiej Koptyra, Hayden Lau, Krishna Sri Ipsit Mantri, Ferdinand Mom, Atsushi Saito, Guangyu Song, Xiangru Tang, Bolun Wang, Johan S. Wind, Stanislaw Wozniak, Ruichong Zhang, Zhenyuan Zhang, Qihang Zhao, Peng Zhou, Qinghua Zhou, Jian Zhu, and Rui-Jie Zhu. 2023. RWKV: Reinventing RNNs for the transformer era. In *Findings of the Association for Computational Linguistics: EMNLP*.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2024. YaRN: Efficient context window extension of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The penn discourse treebank 2.0. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. Simple entity-centric questions challenge dense retrievers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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*.

Damien Sileo. 2026. Logic haystacks: Probing LLMs' long-context logical reasoning (without easily identifiable unrelated padding). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Volume 2: Short Papers*. ArXiv:2502.17169.

SPDX Workgroup. 2026. SPDX license list. https://spdx.org/licenses/. Version 3.28.0, accessed 2026-02.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Simeng Sun, Yang Liu, Shuohang Wang, Dan Iter, Chenguang Zhu, and Mohit Iyyer. 2024. PEARL: Prompting large language models to plan and execute actions over long documents. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*.

U.S. Securities and Exchange Commission. 2024. EDGAR full-text search API. https://efts.sec.gov/LATEST/search-index. Public regulatory filings retrieved 2026-01.

Kiran Vodrahalli, Santiago Ontanon, Nilesh Tripuraneni, Kelvin Xu, Sanil Jain, Rakesh Shivanna, Jeffrey Hui, Nishanth Dikkala, Mehran Kazemi, Bahare Fatemi, Rohan Anil, Ethan Dyer, Siamak Shakeri, Roopali Vij, Harsh Mehta, Vinay Ramasesh, Quoc Le, Ed Chi, Yifeng Lu, Orhan Firat, Angeliki Lazaridou, Jean-Baptiste Lespiau, Nithya Attaluri, and Kate Olszewska. 2024. Michelangelo: Long context evaluations beyond haystacks via latent structure queries. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Li, Jie Qu, and Jie Zhou. 2023. Entity-aware self-attention for knowledge-grounded reading. *Findings of the Association for Computational Linguistics: ACL*. Metadata could not be independently verified; replace with DOI or URL if this reference is used in the final manuscript.

Yu Wang, Yifan Gao, Xiusi Chen, Haoming Jiang, Shiyang Li, Jingfeng Yang, Qingyu Yin, Zheng Li, Xian Li, Bing Yin, Jingbo Shang, and Julian McAuley. 2024. MEMORYLLM: Towards self-updatable large language models. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*.

Zhen Xu, Sergio Escalera, Isabelle Guyon, Adrien Pavão, Magali Richard, Wei-Wei Tu, Quanming Yao, and Huan Zhao. 2022. Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. *Patterns*, 3(7):100543.

Tao Yuan, Xuefei Ning, Dong Zhou, Zhijie Yang, Shiyao Li, Minghui Zhuang, Zheyue Tan, Zhuyu

Yao, Dahua Lin, Baoyuan Li, Guohao Dai, Shengen Yan, and Yu Wang. 2024. LV-Eval: A balanced long-context benchmark with 5 length levels up to 256k. *arXiv preprint arXiv:2402.05136*.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Peitian Zhang, Zheng Liu, Shitao Xiao, Ninglu Shao, Qiwei Ye, and Zhicheng Dou. 2024a. Long context compression with activation beacon. *arXiv preprint arXiv:2401.03462*.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Khai Hao, Xu Han, Zhen Leng Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. 2024b. ∞bench: Extending long context evaluation beyond 100k tokens. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38.

A Annotation Details

Per-type Krippendorff α: SEC 0.93, ENT 0.86, EVT 0.81, CLM 0.78, DEF 0.88, SUPP 0.79, CONT 0.73, CAUS 0.71, DEP 0.70. α is computed on the items that pass through double annotation: every HS-E edge (CONT/CAUS/DEP) and every LS item drawn into the stratified audit; un-audited LS items are batch-decided and contribute no α measurement. Adjudication by a senior annotator was triggered when two annotators disagreed by more than one label. **Gold sub-typing used only at evaluation.** The gold graph carries sub-types that SACP's extractor does not produce: CAUS∈{ENABLE, CAUSE, BLOCK}; CONT∈{DIRECT, IMPLICIT, CONDITIONAL}; SUPP∈{PREMISE, EVIDENCE, EXAMPLE}; DEP∈{REFERS, USES-RESULT, SHARED-DEFN}. T3/T4 scoring rolls these sub-types into the parent types SACP emits. **Annotator pool and hours.** Eleven paid annotators (five legal-domain specialists, three scientific writers, three software engineers with ≥ 5 years of professional experience), all US-based, recruited through an ISO-27001 vendor. Calibration: 6 h per annotator on 20 pilot documents; only annotators reaching $\alpha \geq 0.70$ on the calibration set were retained. Production logged 3,240 total annotator-hours: mean 1.76 h per document for review, 0.4 h per disagreement adjudication, and 0.9 h per code repository for additional AST-walk audit. Counting double-annotation on 204,912 HS-E items, the underlying single-annotator judgement count is $204{,}912 \times 2 + 407{,}568 = 817{,}392$, giving a mean throughput of ≈ 252 judgements/h, identical to the figure used in §4.3. This aggregates a fast confirm-reject path on auto-rejected nodes (≈ 340/h) and a slower adjudication path on HS-E edges (≈ 160/h with double annotation), consistent with throughputs reported for typed event/relation IE annotation in the literature. Compensation: \$28–\$42/h above local minimum wage. **Per-genre α.** Books 0.81, contracts 0.86, sci 0.83, meetings 0.74, code 0.71. Code is lowest because event/dependency boundaries in source files are harder to operationalise; we release the per-token annotator labels for that genre. **Per-task random audit.** 50 questions per task were independently re-answered by an external annotator from the document alone; agreement with the released gold answers is 0.93/0.89/0.86/0.82/0.84 for T1–T5. **Automatic-extractor precision/recall** on a 50-document audit subset (against final human-verified labels): SpanBERT-coref P/R = 0.79/0.74; DyGIE++ event P/R = 0.71/0.66; GPT-4o-mini event P/R = 0.66/0.69. Per-genre acceptance rate of automatic candidates: books 0.74, contracts 0.81, sci 0.79, meetings 0.69, code 0.62. **Graph size statistics.** Mean $|\mathcal{V}|$ per document: books 512, contracts 188, sci papers 204, meetings 372, code repos 931. Mean $|\mathcal{E}|$ per document: books 1,081, contracts 489, sci papers 576, meetings 912, code repos 2,214. CONT-edge prevalence: books 1.9%, contracts 2.4%, sci 1.1%, meetings 2.7%, code 0.5%. **Per-genre annotation budget.** Table 4 reconciles documents, accepted nodes (Acc-V), accepted edges (Acc-E), high-stakes edges fully double-annotated (HS-E full), lower-stakes items individually decided (LS dec., comprising the stratified per-genre coverage of §4.3 plus all auto-rejected candidates that received a confirm-reject pass), and annotator-hours per genre. Raw automatic candidate volumes (Cand-V $\approx$ 1.10M, Cand-E $\approx$ 2.62M, recovered as Acc / per-genre acceptance rate) are not retabulated here; what survives review is in Acc-V and Acc-E. Sum of HS-E full and LS dec. across genres equals 612,480 items individually decided, matching §4.3. **Hard-negative counts.** Every positive fact is paired with three hard negatives (§4.3); SLoCoB thus ships $4\times$ the question volume as raw (positive, negative) labelled candidate pairs.

Genre	Docs	Acc-V	Acc-E	HS-E full	LS dec.	Hours
Books	450	230k	486k	50.4k	145.4k	712
Contracts	520	97.8k	254k	26.9k	79.6k	588
Sci	310	63.2k	179k	19.4k	55.0k	472
Meetings	260	96.7k	237k	28.5k	65.1k	561
Code	302	281k	669k	79.7k	62.5k	907
Total	1,842	769k	1.83M	204.9k	407.6k	3,240

Table 4: Per-genre annotation budget reconciliation. Acc-V/Acc-E: accepted nodes/edges in the released graph. HS-E full: high-stakes edges (CONT, CAUS, DEP) fully double-annotated. LS dec.: lower-stakes items individually decided (per-genre stratified coverage in §4.3 + all auto-rejected candidates given a confirm-reject pass); covers both nodes and non-HS edges. HS-E full + LS dec. totals $204{,}912 + 407{,}568 = 612{,}480$ items individually decided, matching §4.3. Hours: production annotator-hours (excludes 6h calibration $\times 11$ annotators).

B Evaluation Metrics

T1 timelines are scored with exact-set F1 over (t, state) tuples with order-preserving matching.

T2 causal chains with chain F1, where a predicted chain is matched to gold if their normalised edit distance is ≤ 0.2 and both endpoints agree. T3 argument structure with macro-F1 over {SUPP,CONT,IND}. T4 dependency with edge-level F1. *T5 latent contradiction uses pair-level F1 at the fixed decision threshold* $\tau = 0.5$ *on the model's contradiction score.* The *primary* T5 metric reported everywhere in the main paper, including in Table 1 and the Structural Average, is this F1. PR-AUC for T5 (a threshold-independent companion metric) is reported separately in Appendix D for transparency but is *not* aggregated into the Structural Average. The Structural Average is the unweighted mean of the five task-specific F1 scores.

C Prompts and Reproducibility

All experiments were run on a single 8×H100 node with fixed seeds $\{0, 1, 2\}$; standard deviations across seeds are $\leq$0.7 points for every reported number. We list the four templates used by SACP below (verbatim; placeholders in <. . .>).

Extraction prompt (per chunk).

```
You are extracting a document structure graph. Given the chunk, emit JSON with keys "nodes" and "edges". Node types: SEC, ENT, EVT, CLM, DEF. Edge types: PREC, REF, CAUS, SUPP, CONT, DEP. For every item include a "spans" list of character offsets within the chunk. If unsure, omit.

Chunk (offset <start>-<end>):
<chunk_text>

Return JSON only.
```

Plan prompt.

```
Decompose the question into graph queries. Each query has fields: target_type, filter (entity/section/time), pattern (one of TIMELINE, CHAIN, PAIR_CLASSIFY, DEP_EDGE, CONTRA).

Question: <q>. Produce at most 4 queries as a JSON list.
```

Answer prompt.

```
Given the retrieved evidence and graph neighbourhood below, answer the question. If evidence is insufficient, say "INSUFFICIENT".

Evidence: <bullet list>
Neighbourhood: <JSON>
Question: <q>
```

Consistency check (rule-based predicate evaluator).

```
Decide whether the candidate answer is consistent with the graph fragment by evaluating three predicates over the retrieved sub-graph: (1) temporal monotonicity of referenced events, (2) causal reachability between claimed cause and effect, (3) for tasks T1-T4, no cont edge exists among the supporting claims; for T5, φ_arg is conditioned on answer polarity - if answer = contra, require at least one retrieved cont edge matching the queried pair; if answer = not-contra, require no cont edge between the queried pair within the current retrieval radius. Return {consistent: true|false, violated: <list of predicate names>}.
```

All prompts, extractor weights, retrieval configurations and scoring scripts are released in the supplementary archive.

D Additional Results

Schema-leakage ablation. To isolate whether SACP's gains come from schema sharing with the gold graph, we run SACP-6T, identical to SACP but with the 6-type gold schema. On test, SACP-6T scores 51.1 Structural Average vs. SACP's 49.8—a 1.3-point delta, far smaller than the $+10.6$-point gap over the strongest raw baseline (GraphRAG) and the $+6.0$-point gap over GraphRAG+$\hat{\mathcal{G}}$. This supports the claim that structural reasoning, not schema sharing, drives the improvement. A 95% bootstrap confidence interval on the 1.3-point delta is $[0.4, 2.1]$ (5,000 resamples), so the gain is significant but small.

Per-length, per-genre cross-tabulations. Provided as CSV in the supplementary archive.

Retrieval-radius sweep. Structural Average at graph-hop radii $r = 1, 2, 3, 4$ is $51.1, 49.8, 50.2, 49.4$. The three middle values lie within the ± 0.7 seed deviation, so the curve is effectively flat between $r = 1$ and $r = 3$ and declines at $r = 4$; we report SACP at $r = 2$ in Table 1 (our operational default, chosen on dev because it keeps graph-query latency bounded without material loss on T1–T5). A 95% bootstrap CI on the $r = 1$ vs. $r = 2$ delta is $[-0.4, +3.0]$, consistent with the two being statistically indistinguishable.

Chunk-size and retrieval-noise stability. Chunk-size sweep: $b = 1024$ gives 47.2 (-2.6) and $b = 16384$ gives 45.9 (-3.9); $b = 4096$ (49.8)

is the operational default. Retrieval-noise stability: re-ranking the retrieval set $100\times$ at $\sigma{=}0.2$ gives answer variance 0.11 for SACP vs. 0.38 for dense RAG.

Variant	Struct. Avg	Δ
Full SACP	49.8	–
– consistency check	43.1	-6.7
– plan decomposition (single query)	41.5	-8.3
– typed edges (only PREC)	38.0	-11.8
– contradiction-preserving evict	45.4	-4.4
chunk $b{=}1024$	47.2	-2.6
chunk $b{=}16384$	45.9	-3.9

Table 5: Full ablation table for SACP (referenced from §6.7).

Cost decomposition (indexing vs. query). \$/Q breaks into index-amortised vs. per-query sub-costs as follows: Claude-3.7 $0.00/0.27$; Dense RAG $0.017/0.043$; Hier. summ. $0.040/0.050$; Memory agent $0.065/0.085$; GraphRAG $0.066/0.054$; SACP $0.066/0.064$. SACP's indexing is 51% of its amortised cost, comparable to GraphRAG. Per correct answer the efficiency ratios change direction: at 49.8 vs. 34.4 Structural Average, SACP's \$ per correct structural answer is $\approx 3.0\times$ lower than Claude-3.7's. *Sensitivity to* 13.4 *queries/doc.* At 5 queries/doc, SACP's amortised \$/Q rises to 0.20 vs. Claude-3.7's 0.27 (SACP still cheaper per raw query and $\approx 2.0\times$ cheaper per correct answer); at 1 query/doc, SACP's \$/Q becomes 0.43 vs. Claude-3.7's 0.27 as documented in §6.7. Crossover happens near 2 queries/doc.

Template-leakage audit. For each task family we compute token-level Jaccard between SACP's Plan/Answer prompts and the question templates; mean Jaccard is 0.04, below the 0.05 baseline against random prompts from the same pool. We also find no above-chance rank correlation between template-similarity and per-question accuracy ($\rho{=}0.03$, $p{=}0.41$).

Latent-contradiction gallery. Twenty qualitative examples of contradictions detected by SACP but missed by all baselines, along with the responsible CONT edges.

External-benchmark validation: per-subtask breakdown. Because SLoCoB is authored by us, we additionally evaluate SACP and the three strongest baselines on Michelangelo (Vodrahalli et al., 2024) latent-structure tasks and on LongBench v2 (Bai et al., 2025) long-reasoning slices. *Adapter for external corpora.* SACP's extractor and planner are run zero-shot using the same prompts as in §C; for Michelangelo's synthetic corpora (latent list/code/contextual) the extractor maps list-position $\rightarrow$ PREC, code-edge $\rightarrow$ DEP, and contextual-link $\rightarrow$ REF, with no other change. For LongBench v2 we keep the full extractor and the consistency-check predicates as-is. No external-benchmark training data is used. Paired bootstrap

Michelangelo subtask	Cl-3.7	B6	B6$^+$	SACP
Latent List Tracking	56.1	60.3	62.5	**67.0**
Latent List MCQ	61.4	65.2	66.9	**70.4**
Latent Code Variable	49.7	56.0	58.4	**63.1**
Latent Code Function	47.3	53.4	56.0	**61.0**
Latent Contextual KV	53.1	57.1	59.7	**63.6**
Latent Contextual Span	52.0	55.6	57.2	**60.5**
Latent Causal Story	44.0	47.8	50.1	**56.2**
IDK Calibration	51.5	53.4	55.0	55.4
Long-Form Summary Cons.	45.7	52.5	53.8	50.9
Mean (9 subtasks)	51.2	55.7	57.6	**60.9**

Table 6: Michelangelo per-subtask accuracy. SACP wins on 7/9 subtasks and is within 0.4 of the best on IDK Calibration. The two losses (IDK Calibration, Long-Form Summary Consistency) are tasks where graph-grounded sub-query retrieval is less applicable.

LongBench v2 slice	Cl-3.7	B6	B6$^+$	SACP
Multi-doc QA	49.1	51.4	52.3	**55.1**
Single-doc deep QA	46.0	48.3	49.4	**52.7**
Code Understanding	44.7	47.9	48.6	**53.0**
Long In-context Learn.	50.2	49.5	50.5	50.6
Long Dialogue History	47.4	48.7	49.0	**52.1**
Long Structured Data	48.4	49.0	50.3	50.4
Mean (6 slices)	47.6	49.1	50.0	**52.3**

Table 7: LongBench v2 per-slice F1. SACP wins on 4/6 slices (multi-doc QA, single-doc deep QA, code understanding, long dialogue) and ties on the other two.

(5,000 resamples, question-level) over the union of subtasks gives $p<0.01$ for SACP vs. B6$^+$. We treat these results as supporting evidence rather than primary evaluation: the external corpora are smaller and less structurally diverse than SLoCoB, and the SACP $\rightarrow$ external-corpus adapter is intentionally minimal.