Same Conversation, Different Answer: How Evidence Rendering Changes LLM Memory Evaluation

Anonymous ACL submission

Abstract

Memory and RAG systems rarely pass a full chat history to the answering model; they render it as a memory entry, summary, typed record, or raw excerpt. Across diagnostic and matched-budget controls, measured accuracy ranges from near zero to 82%, showing that reader-facing artifacts can dominate memory-QA evaluation.

We introduce FRAME (Format Rendering Ablation for Memory Evaluation), a benchmark control that fixes the conversation while varying the reader-facing artifact. FRAME combines a five-level packet ladder, localizing when answer-bearing content enters the input, with deterministic templates approximating ChatGPT-style entries, LangChain summaries, MemGPT-style typed records, and raw conversation. On 500 LongMemEval questions and nine models, matched-budget resolved packets beat recency-truncated raw dialogue by 42.4–72.6 points. In deployed-style templates, best–worst spread is 24.6–48.8 points per model; under the primary scorer, ChatGPT-style entries have higher point estimates than raw conversation on 7 of 9 models. Judge rescoring preserves the positive aggregate effect, but model-specific significance is mixed. Three models scoring 0% on formal ledger packets answer the same facts from natural-language entries at 45.4–53.4%. The effect persists under retrieval noise and transfers to HotpotQA, suggesting that memory/RAG evaluations should report or control the reader-facing artifact.

1 Introduction

A memory or retrieval-augmented generation (RAG) system does not hand the reader a conversation; it hands the reader a *rendering* of one. We use rendering broadly: the reader-facing artifact includes selection, compactness, conflict-resolution exposure, metadata density, abstention cues, and surface form. Across diagnostic and matched-budget controls, we show that this choice is not a neutral implementation detail: measured accuracy ranges from near zero to 82%, depending on what artifact reaches the answering model.

This is an evaluation problem. Memory and RAG benchmarks often compare pipelines whose readers receive different artifacts: one system passes raw retrieved dialogue, another passes a summary, and another passes a structured memory record. If those artifacts are not reported or controlled, the score conflates model ability, retrieval quality, memory-update logic, compactness, conflict-resolution exposure, abstention cues, and evidence rendering. A reported gain may therefore reflect a better reader-facing artifact rather than a better model or memory system. We ask: *holding the underlying conversation, question, and answer contract fixed, how much can the reader-facing evidence artifact change measured memory-QA accuracy?*

Consider a user who first says they live in Boston, later says they moved to Denver, and then asks "Where do I live now?" A ChatGPT-style memory entry may show *"User lives in Denver (moved from Boston)"*; a MemGPT-style record may expose `user.city = "Denver"` (Packer et al., 2024); a LangChain-style summary may compress the session into a paragraph (LangChain, 2026). These artifacts encode the same fact, but they are not equivalent reader inputs. In our experiments, deployed-style templates create a 24–49 point accuracy spread per model.

We introduce FRAME (*Format Rendering Ablation for Memory Evaluation*), a benchmark control that fixes the underlying conversation and answer contract while varying the reader-facing artifact. FRAME has two instruments. First, a five-level packet ladder reveals answer-bearing content progressively: P_0 and P_1 expose only witness addresses, P_2 first writes the resolved current-state value into the packet body, and P_3/P_4 add metadata around that value. This ladder localizes whether

failures stem from missing content, unresolved conflict, or the surrounding packet surface. Second, a template-family comparison renders the same underlying dialogue as deterministic approximations of deployed memory surfaces: ChatGPT-style natural-language entries, LangChain summaries, MemGPT-style typed records, and raw conversation.

FRAME is an input-side study, not an output-format study. Prior work asks whether models reason worse when required to *produce* JSON, XML, or constrained outputs (Tam et al., 2024; Lee et al., 2026). We fix the answer contract and vary only the evidence the reader sees.

We run about 238,000 model calls across LongMemEval (500 questions; Wu et al. 2025), LoCoMo (200 questions; Maharana et al. 2024), retrieval-noise variants, and HotpotQA transfer. The main experiments use nine commercial models from OpenAI, Anthropic, and Google, with tools, search, and provider-native memory disabled. An independent LLM judge rescores the 31,500 items from the two headline experiments, with a 600-item second-judge sample confirming agreement (Appendix C).

Findings.

1. **Accuracy rises when the answer enters the packet.** All nine models stay near 0% at P_0/P_1; responsive models recover to 15–25% at P_2. Metadata above P_2 changes accuracy by at most ± 2 points.

2. **Budget control reverses the raw-vs-structured conclusion.** Full raw conversation outperforms budgeted formal packets, but under matched word budgets, streamlined resolved-P_2 packets outperform recency-truncated raw dialogue by 42.4–72.6 points on every model.

3. **Deployed-style surface choice substantially changes the score.** Under the primary substring scorer, ChatGPT-style entries have higher point estimates than raw conversation on 7 of 9 models; judge rescoring preserves a positive aggregate effect but shows mixed model-specific significance. LangChain summaries score 4.4–10.6%; MemGPT-style typed records score 14.8–33.0%. The best–worst spread reaches 24.6–48.8 points per model.

4. **Evidence artifacts can induce refusal.** Three models score 0% on formal ledger packets but answer the same facts from natural-language entries at 45.4–53.4%, consistent with artifact/prompt-triggered abstention rather than missing capability.

These results support a reporting norm: memory/RAG evaluations should state the reader-facing evidence artifact or include a fixed-artifact control. In our experiments, without this control, a 24–49 point artifact spread is indistinguishable from model or pipeline progress.

2 Related Work

Long-context and memory evaluation. Long-context benchmarks stress retrieval, context length, positional robustness, or long-horizon reasoning rather than the rendered form of fixed evidence. Examples include RULER (Hsieh et al., 2024), InfiniteBench (Zhang et al., 2024c), L-Eval (An et al., 2024), ZeroSCROLLS (Shaham et al., 2023), LongBench and LongBench v2 (Bai et al., 2024, 2025), Marathon (Zhang et al., 2024b), and BABILong (Kuratov et al., 2024). Architecture-level work extends usable context through recurrence, compression, retrieval, or memorization (Dai et al., 2019; Rae et al., 2020; Wu et al., 2022; Khandelwal et al., 2020). LongMemEval (Wu et al., 2025) and LoCoMo (Maharana et al., 2024) are closest to our setting because they evaluate multi-session conversational memory with updates and temporal scope. Work on position sensitivity shows that models may fail even when relevant information is present (Liu et al., 2024); FRAME studies an orthogonal variable: how that information is rendered to the reader.

Memory, retrieval, and RAG systems. Modern memory systems often rewrite raw histories before generation. MemGPT (Packer et al., 2024) externalizes memory into an OS-like hierarchy; HippoRAG and GraphRAG organize evidence through graph indices (Gutierrez et al., 2024; Edge et al., 2025); ENGRAM uses typed memory schemas (Patel and Patel, 2025); and agent memory work includes MemoryBank, Generative Agents, and Reflexion (Zhong et al., 2023; Park et al., 2023; Shinn et al., 2023). Production frameworks such as LangChain also expose summary-based conversation memory utilities (LangChain, 2026). RAG work attaches an external store to the reader (Lewis et al., 2020; Guu et al., 2020; Borgeaud et al., 2022; Izacard et al., 2022), extends retrieval to multi-hop reasoning (Trivedi et al., 2023; Asai et al., 2024), and evaluates retrieval quality, faithfulness, and

context use (Es et al., 2024). These studies generally evaluate end-to-end systems; FRAME isolates the reader-facing artifact itself.

Format sensitivity and judging. Output-side structured-generation work changes how the model must answer, e.g., by requiring JSON, XML, Markdown, or LaTeX (Tam et al., 2024; Lee et al., 2026). FRAME changes what evidence the model reads while holding the answer contract fixed. More broadly, LLMs are sensitive to prompt formatting, few-shot ordering, demonstration composition, chain-of-thought prompts, and self-consistency decoding (Sclar et al., 2024; Lu et al., 2022; Min et al., 2022; Zhao et al., 2021; Wei et al., 2022; Kojima et al., 2022; Wang et al., 2023b); recent RAG work also shows that context formatting and stronger raw-passage baselines can change downstream reasoning and evaluation conclusions (Chen et al., 2025; Laitenberger et al., 2025). Tooling for constrained generation helps systems emit typed records (Willard and Louf, 2023), but does not test whether those records are the best evidence artifact for a downstream reader. We use substring matching, normalized exact match, and LLM judges, building on work on MT-Bench/Chatbot Arena, G-Eval, PandaLM, and known judge biases (Zheng et al., 2023; Liu et al., 2023; Wang et al., 2024, 2023a). HotpotQA and related multi-hop QA datasets provide a non-conversational transfer setting (Yang et al., 2018; Ho et al., 2020; Trivedi et al., 2022). Abstention and knowledge-conflict work shows that refusal behavior can be shaped by task framing and contextual conflict (Zhang et al., 2024a, 2025; Wen et al., 2025; Röttger et al., 2024; Kadavath et al., 2022; Tian et al., 2023; Lin et al., 2022; Xu et al., 2024; Wilie et al., 2024; Longpre et al., 2022; Xie et al., 2024); our contribution is to show that reader-facing artifacts and their abstention cues can trigger such behavior.

3 System Design and Methodology

FRAME is a benchmark control for a narrow input-side question: when the underlying conversation, question, and answer contract are fixed, how much can the reader-facing evidence artifact change memory-QA accuracy? Each item has three fixed objects: the original dialogue, the question, and the expected answer behavior. FRAME varies the artifact shown to the final answering model, which we call the *reader*; where a control also changes prompt wrapping, we label that as part of the condition. The reader receives one rendered evidence artifact, answers the same free-form question, and is forbidden from using external tools, search, provider-native memory, or hidden context.

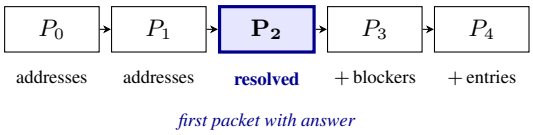

Figure 1: The five-packet ladder. P_0/P_1 expose witness addresses only; P_2 first writes the resolved current-state answer into the packet body; P_3/P_4 add blocker metadata and supporting entries.

Internally, FRAME stores turn-addressed dialogue, identifies answer-bearing events, resolves conflicts, and renders different reader-facing artifacts. Externally, the reader sees only the rendered artifact and a fixed free-form answer contract. No condition asks the reader to emit JSON, XML, typed records, or any other structured output; the manipulated variable is the input-side evidence artifact.

Packet ladder. The first control is a five-level ladder that reveals increasingly more answer-bearing information from the same evidence. The ladder is a processing-stage diagnostic: it identifies whether failure occurs because the reader cannot see relevant content, because conflict resolution has not been made explicit, or because the surrounding packet surface changes reader behavior. We use: P_0 `raw_packet`, with witness addresses and blocker flags but no answer text; P_1 `positive_only`, with witness addresses only; P_2 `active_only`, the first level that writes the resolved current-state value into the packet body; P_3 `ctx_normalized`, which adds blocker metadata around P_2; and P_4 `full_frame`, which adds a budget-aware list of supporting entries. P_0/P_1 expose audit pointers, not links the reader can open. P_2 is therefore the first answer-bearing packet.

Queries and conflict resolution. Each dialogue is ingested into an append-only store with stable turn addresses of the form `history/<session>/turn/<turn>`. Each question is mapped to a canonical `FrameQuery` containing a normalized slot key, temporal scope (e.g., current-state or as-of reasoning), and conversational scope (e.g., full-history or session-restricted reasoning). Candidate events are tagged as updates, corrections, contradictions, or deletions,

then ordered by precedence; deletions take highest precedence. Scope resolution marks candidates as *applicable*, *out_of_scope*, or *blocked*. A resolved current state exists only when the slot collapses to one applicable non-blocked event. Otherwise the packet carries a blocker, either `scope_blocked` or `unresolved_conflict`. Abstention is rendered explicitly as

$$\texttt{should_abstain} = \mathbf{1}[\#\text{blockers} > 0].$$

Reader prompts and portability. Each packet is wrapped by three reader prompt families: a plain reader, a suppression reader that ignores updated/corrected/deleted/blocked entries, and a ledger-aware reader that prefers the latest applicable entry. For each question–model pair, the structured evaluation crosses five packet levels, three readers, and two internal backends, yielding $5 \times 3 \times 2 = 30$ structured conditions; Table 1 aggregates over the six reader/backend variants for each packet level. Porting FRAME to a new memory-QA dataset requires a turn-addressed dialogue adapter, a question-to-slot mapping, and a source of conflict sets. The renderers and reader prompts are otherwise dataset-independent; porting to LoCoMo required about 200 lines of Python and no renderer changes.

4 Experimental Setup and Results

Setup. We evaluate FRAME primarily on the oracle tier of LongMemEval, which contains 500 memory-QA questions with multi-session updates, temporal scope, and abstention (Wu et al., 2025). We test nine commercial models from three providers: OpenAI (`gpt-5.4-mini`, `gpt-5.2`, `gpt-5.2-pro`), Anthropic (`claude-sonnet-4`, `claude-opus-4.1`, `claude-opus-4.6`), and Google (`gemini-3-flash`, `gemini-2.5-flash-lite`, `gemini-3.1-flash-lite`). Exact API IDs are listed in Appendix A. Provider-native tools, search, and memory are disabled. The main LongMemEval ladder contains 139,500 calls; auxiliary budget, template, matched-content, retrieval-noise, LoCoMo, and HotpotQA experiments plus LLM-judge rescoring bring the study to about 238,000 calls.

Our primary metric is accuracy over all questions: a response is correct iff it does not begin with an explicit `ABSTAIN` refusal and the lowercased gold answer appears as a substring of the lowercased output. Abstentions count as incorrect. Appendix C cross-validates the headline results with normalized exact match and an independent LLM judge. Appendix C.6 also audits late-abstention leakage in the substring scorer; a stricter abstention-anywhere check shifts all headline aggregate accuracies by less than 1 point and leaves qualitative conclusions unchanged.

Condition map. We use two resolved-P_2 conditions for different controls. Table 1 reports the formal packet ladder: P_2 uses the formal reader prompt and packet machinery averaged over reader prompts and backends. Table 2 reports a streamlined resolved-P_2 budget control: the resolved value is wrapped in a shorter reader prompt without the formal blocker and abstention fields. Raw baselines also differ by control: Table 1 uses full dialogue in the ladder setup, Table 2 reports both full and recency-truncated raw under the budget prompt, and Figure 3 uses raw dialogue inside the deployed-template prompt.

4.1 Accuracy Rises When the Answer Enters the Packet

Figure 2 shows the core diagnostic. When packets expose only witness addresses (P_0/P_1), all nine models remain near zero accuracy (0.0–0.6%). Accuracy rises at P_2, the first packet level that resolves the conflict set and writes a current-state value into the packet body; the six responsive models reach 15.0–25.3%. Adding metadata above P_2 has little effect: P_3/P_4 shift accuracy by at most ± 2 points per model. Three models remain at 0.0% across all formal structured packets; Section 5.1 analyzes these hard-abstention cases.

Full raw conversation is a strong unconstrained baseline because it includes the entire dialogue. It outperforms budgeted formal packets on every model (Table 1), with raw-over-structured gaps from 10.5 to 44.4 points. This comparison is informative but not causal: it mixes surface with context length, because raw conversation preserves all recency cues and earlier answer-bearing turns while structured packets are budgeted.

4.2 Budget Control Reverses the Raw-vs-Structured Conclusion

To compare artifacts under similar budgets, we truncate raw conversation to about 800 words, keeping the most recent turns, and compare it with a

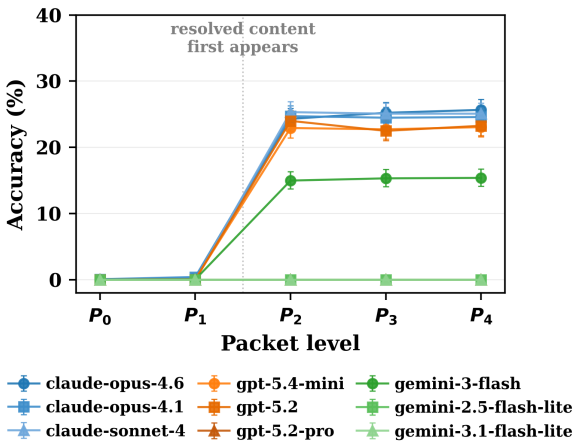

Figure 2: Accuracy across the five-packet ladder on LongMemEval (500 questions). P_0/P_1 expose witness addresses only; P_2 first writes the resolved answer into the packet body; P_3/P_4 add metadata around that value.

Table 1: Main LongMemEval results. Raw uses full dialogue; formal packet columns average over 3 reader prompts $\times$ 2 backends per question. Δ is each row's max–min spread.

Model	Raw	P_0	P_1	P_2	P_3	P_4	Δ
`claude-opus-4.6`	**45.2**	0.5	0.6	24.3	24.9	25.3	44.7
`claude-opus-4.1`	**36.8**	0.4	0.6	24.7	24.0	24.1	36.4
`claude-sonnet-4`	**35.8**	0.4	0.4	25.3	24.7	24.7	35.4
`gpt-5.4-mini`	**35.4**	0.4	0.5	22.9	22.4	22.7	35.0
`gpt-5.2`	**42.4**	0.4	0.4	24.3	22.4	23.2	42.0
`gpt-5.2-pro`	**44.4**	0.0	0.0	0.0	0.0	0.0	44.4
`gemini-3-flash`	**34.6**	0.2	0.2	15.0	15.1	15.2	34.4
`gemini-2.5-flash-lite`	**29.2**	0.0	0.0	0.0	0.0	0.0	29.2
`gemini-3.1-flash-lite`	**38.2**	0.0	0.0	0.0	0.0	0.0	38.2

Table 2: Word-budget comparison on LongMemEval (500 questions). Truncated raw keeps the most recent $\sim$800 words. Δ = streamlined P_2 minus truncated raw; all deltas are significant under paired bootstrap.

Model	Full raw	Trunc. raw	Stream. P_2	Δ
`claude-opus-4.6`	51.4	10.4	**76.0**	+65.6
`claude-opus-4.1`	40.4	11.0	**72.2**	+61.2
`claude-sonnet-4`	33.2	10.2	**78.4**	+68.2
`gpt-5.4-mini`	38.8	11.6	**73.0**	+61.4
`gpt-5.2`	39.4	11.0	**77.0**	+66.0
`gpt-5.2-pro`	50.4	12.0	**77.0**	+65.0
`gemini-3-flash`	30.4	7.6	**50.0**	+42.4
`gemini-2.5-flash-lite`	25.4	9.4	**57.2**	+47.8
`gemini-3.1-flash-lite`	37.2	9.4	**82.0**	+72.6

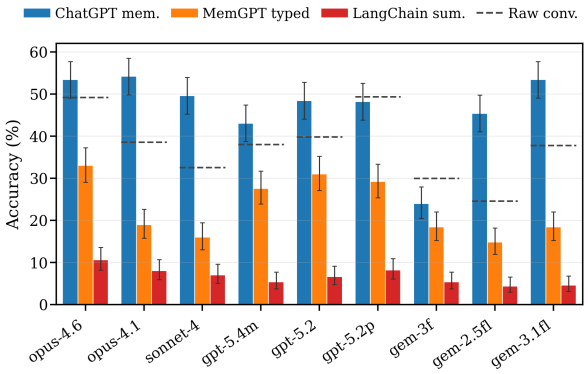

Figure 3: Deployed-template comparison on LongMemEval (500 questions). Under the primary scorer, ChatGPT-style entries have higher point estimates than raw conversation on 7 of 9 models; judge rescoring shows mixed model-specific significance.

streamlined resolved-P_2 packet on the same 500 questions. Under this control, truncated raw conversation scores only 7.6–12.0%, while streamlined P_2 scores 50.0–82.0%, yielding a 42.4–72.6 point advantage for structured evidence on every model (Table 2). The judge-rescored gap remains positive and significant for every model, ranging from +48.4 to +80.2 points (Appendix C.4). The result is not that structure is intrinsically better than raw dialogue; rather, once the budget is tight, recency truncation often drops earlier answer-bearing turns, whereas resolved P_2 preserves the value regardless of its original position.

This comparison also clarifies hard abstention. In the main ladder, `gpt-5.2-pro` scores 0.0% on every formal FRAME packet; in the budget-matched condition, it scores 77.0% on the same active-only packet body when wrapped in a streamlined instruction prompt rather than a formal reader prompt with explicit blocker and abstention machinery. The same pattern holds for the other hard-abstention models, so we treat the refusal result as an artifact-and-prompt interaction rather than a pure field-format effect.

4.3 Deployed Memory Templates Change Scores by up to 49 Points

We next approximate reader-facing outputs from deployed memory strategies using deterministic template families: ChatGPT-style natural-language entries, LangChain-style session summaries, MemGPT-style typed JSON records, and raw conversation. This is a deployed-template comparison rather than a pure same-content ablation: ChatGPT and MemGPT-style templates receive up to five gold-localized turns, LangChain summaries truncate each session to its first few turns, and raw conversation includes the full dialogue.

Surface choice substantially changes the reported score (Figure 3). Under the primary substring scorer, ChatGPT-style entries have higher point estimates than raw conversation on seven of nine models, with gains of 4.2–20.8 points; the two exceptions are `gemini-3-flash` (−6.0)

and `gpt-5.2-pro` (-1.2). Judge rescoring preserves a positive aggregate effect (+7.9pp) but shows mixed model-specific significance (Appendix C.5). MemGPT-style typed records are model-dependent (14.8–33.0%), while LangChain summaries are consistently weakest (4.4–10.6%). Across models, the best–worst surface spread ranges from 24.6 to 48.8 points. The three models that score 0.0% on formal FRAME packets answer the same facts from ChatGPT-style entries at 45.4–53.4%, indicating that formal ledger packets can trigger refusal even when compact natural-language evidence remains usable.

4.4 Generalization and Robustness

The pattern is not specific to the main LongMemEval setting. On 200 LoCoMo questions, every model scores at most 0.5% when only witness addresses are visible and recovers to 32.0–97.0% at active P_2; no model hard-abstains across all structured conditions (Appendix B). Under retrieval noise, aggregate raw accuracy drops 8.8 points from k=0 to k=9 distractors, while ChatGPT-style entries drop only 1.5 points; the NL-over-raw advantage shifts from -1.5 to $+5.9$ points, with per-model heterogeneity (Appendix D). On HotpotQA, a non-conversational multi-hop QA task, aggregate accuracy follows NL entries (69.8%) $>$ raw paragraphs (68.6%) $>$ summaries (67.2%) $>$ typed records (62.7%), suggesting transfer beyond memory QA while also confirming that surface preferences are task-dependent (Appendix E).

5 Analysis

5.1 Representation-induced behavior regimes

Reader-facing evidence artifacts change answer behavior as well as accuracy. The answer-rate drop from raw conversation to `full_frame` ranges from 28.1 to 75.2 points, splitting models into three regimes (visualized in Appendix B). *Hard abstention*: `gpt-5.2-pro`, `gemini-2.5-flash-lite`, and `gemini-3.1-flash-lite` have 0.0% answer rate on every formal packet, despite answering from raw conversation, ChatGPT-style memory entries (45.4–53.4%), and unresolved witness text (18–19%). *Cautious abstention*: `claude-opus-4.6`, `claude-opus-4.1`, `claude-sonnet-4`, `gpt-5.4-mini`, and `gpt-5.2` abstain on P_0/P_1 but answer normally once P_2 resolves the conflict. *Speculative answering*: `gemini-3-flash` answers 22% of unresolved packets at 0.0% accuracy. These regimes do not track provider or model generation.

5.2 Why we report overall accuracy

Answered-only accuracy can reward abstention. `claude-sonnet-4` scores 24.7% overall under `full_frame` at a 29.5% answer rate, yielding 83.7% conditional accuracy mostly because it abstains. `gemini-3-flash` shows the opposite pattern, answering 43.9% of items at 15.2% overall. We therefore report correct answers divided by total questions, with abstentions counted as incorrect.

6 Discussion

What drives the artifact effect. The $P_1 \rightarrow P_2$ jump bundles two operations: the reader gains answer-bearing content, and the upstream system has resolved competing states. A matched-content control separates them by quoting the witness turn verbatim while withholding explicit slot resolution and the `active_value` field (Appendix B). Content visibility is the largest driver in this control: witness text alone lifts accuracy from near zero to 14.2–23.2%. For the six responsive models, explicit conflict resolution adds a smaller 0.8–6.1 points, significant for three. The hard-abstention models sharpen the conclusion: they answer 18–19% of unresolved witness-text items correctly but 0% of resolved formal packets, consistent with artifact- or prompt-induced refusal rather than inability to reason from structured evidence.

Full raw conversation still helps when context is unconstrained because it preserves recency ordering, update phrasing, and local coherence. Under a tight budget, however, recency truncation can remove earlier answer-bearing turns, while a resolved P_2 packet preserves the answer. Compact natural-language memory entries benefit from both properties: they retain enough discourse texture for fluent reading while remaining short enough to include the relevant fact.

Using FRAME as a benchmark control. FRAME is not intended to identify a universal best representation. Its purpose is to make the reader-facing evidence surface an explicit experimental variable. Benchmark authors can evaluate each system's native artifact alongside a fixed-surface control, such as a resolved P_2 packet or deterministic NL entry. System builders can use the packet lad-

der to localize failure: near-zero P_0/P_1 followed by recovery at P_2 indicates that answer-bearing content is not reaching the reader; persistently low P_4 accuracy points to surface-induced refusal or format misinterpretation. Deployment teams can keep structured storage internally for auditability, deduplication, and retrieval while rendering compact natural language for the reader.

Candidate mechanisms and limits of the hierarchy. Three mechanisms are consistent with the data. Compact NL entries may better match the text distribution frontier models are trained on than typed records or formal ledgers. Formal packets include fields such as `blocker`, `active_value`, and `should_abstain`, which may interact with instruction-following or safety-tuned refusal heuristics. Raw dialogue carries discourse cues that typed rendering strips away, while truncation removes earlier evidence entirely. These mechanisms are not mutually exclusive, and the preferred surface is task- and model-dependent: NL entries lead deployed templates on LongMemEval, but raw Wikipedia paragraphs beat NL entries for two Anthropic readers on HotpotQA, and LangChain summaries outperform typed triples on HotpotQA despite being weakest on LongMemEval. The practical implication is that the raw-vs-structured dichotomy is too coarse; compactness, naturalness, metadata density, and abstention signaling all matter.

7 Conclusion

The text handed to the reader by a memory/RAG layer is not always a neutral implementation detail. Across nine models, two memory benchmarks, HotpotQA, and about 238,000 model calls, the same underlying questions produce very different accuracies depending on the reader-facing evidence artifact. Full raw conversation is a strong baseline when context is unconstrained, but under matched word budgets a streamlined resolved packet beats recency-truncated raw dialogue by 42.4–72.6 points. In deployed-style templates, ChatGPT-style memory entries have higher primary-scorer point estimates than raw conversation on 7 of 9 models, while judge rescoring preserves a positive aggregate effect but shows mixed model-specific significance. LangChain summaries are weakest, and formal ledger packets plus abstention machinery trigger near-complete abstention in 3 of 9 models even though those same models answer natural-language memory entries at 45.4–53.4%.

The result does not imply that structure is always bad or that raw dialogue is always best. Evidence rendering is an uncontrolled experimental variable unless it is reported or fixed. FRAME turns that variable into a control: benchmark authors can add a fixed evidence-surface condition, system builders can distinguish missing content from unresolved conflict or refusal-triggering surfaces, and deployment teams can keep structured storage while rendering compact natural language to the reader.

8 Limitations

Design and measurement choices. The token-matched baseline uses word-level truncation as a proxy for token-budget matching; provider-specific tokenizers would be a tighter control, and recency-preserving truncation can miss earlier answer-bearing content. The budgeted comparison therefore does not isolate oracle evidence localization from surface form; an oracle raw-excerpt baseline would further separate localization from rendering. The deployed-template comparison (§4.3) uses deterministic templates that mimic deployed outputs rather than running the systems end to end, so it omits real retrieval errors, summarization errors, and pipeline artifacts. The substring scorer can over-credit verbose answers and miss semantic paraphrases; on LoCoMo it also penalizes raw-conversation answers that express dates relatively. Appendix C cross-validates the headline ordering with normalized exact match and an independent LLM judge, but smaller metadata and formality effects remain more sensitive to scoring choices and sampling noise than the headline step function, token-matched gap, and deployed-template spread.

Template and retrieval scope. FRAME deliberately controls the artifact shown to the reader after evidence has been selected. This design is useful for isolating rendering effects, but it does not measure the full quality of a memory or RAG pipeline. A production system must still decide what to store, when to update stale facts, how to retrieve relevant memories, and how to suppress irrelevant or sensitive material. Our deployed-template variants therefore should be read as controlled approximations of common reader inputs, not as evaluations of ChatGPT, LangChain, MemGPT, or any other product implementation. In particular, ChatGPT-style and MemGPT-style renderings receive gold-localized

turns, while LangChain-style summaries use fixed deterministic truncation. This makes the comparison transparent and repeatable, but it also means that absolute accuracies should not be interpreted as production-system leaderboard scores.

Boundary with structured generation. FRAME does not study whether requiring a model to output JSON, XML, schemas, or other structured formats harms reasoning. The answer contract is fixed across conditions; only the evidence representation changes before the reader answers. The results are therefore not an output-side format-tax measurement and are not evidence about constrained decoding.

Scope and external validity. Evidence comes from LongMemEval's oracle tier, where conflict sets are annotated rather than produced by a retriever. The retrieval-noise experiment uses random distractors and is only a partial substitute for learned retriever error. We evaluate conversational memory and multi-hop Wikipedia QA, not document-grounded RAG or multimodal settings. The nine models are commercial APIs whose system prompts, safety policies, and calibration we cannot inspect; evaluations ran in April 2026 using the exact API IDs listed in Appendix A. The HotpotQA summary and typed surfaces are generated by `claude-opus-4.6`, which is also one evaluator; Appendix E checks this possible generator/reader overlap and finds no systematic advantage for that model. The ChatGPT-style memory template exposes both old and new values, which helps conflict resolution; a production memory store keeping only the latest value would face different tradeoffs. Three models hard-abstain on formal FRAME packets while answering normally on NL entries and unresolved witness text, but we do not identify whether refusal is driven by field names, explicit abstention cues, ledger terminology, or the formal reader prompt.

Statistical interpretation. The largest effects in the paper are robust across scoring variants: answer-bearing P_2 packets beat recency-truncated raw dialogue for every model, and the aggregate deployed-template spread is large under both substring and judge scoring. Some finer-grained claims are intentionally more cautious. For example, ChatGPT-style entries have higher primary-scorer point estimates than raw conversation on 7 of 9 models, but judge rescoring shows mixed model-specific significance. We therefore treat those results as evidence that rendering can materially change measured accuracy, not as a universal ranking of one surface over another.

Implementation heterogeneity for `gpt-5.2-pro`. OpenAI exposes `gpt-5.2-pro` only through the `v1/responses` endpoint, not the chat-completions surface used for the other two OpenAI models. Both endpoints receive identical prompt strings at temperature 0 with tools and provider-native memory disabled, but the endpoint difference remains a confound for any `gpt-5.2-pro` comparison.

9 Ethical Considerations

FRAME studies evaluation artifacts, not deployed user-facing memory behavior. The experiments use released benchmark data and do not introduce new human-subject data collection. Because conversational memory and RAG systems can surface sensitive or stale personal information, evidence rendering should be reported and controlled, not treated as a recommendation to expose raw private histories. The released artifacts support audit, not deployment safety claims; production systems still require privacy review, data minimization, consent, access controls, and task-specific refusal testing.

Release boundaries. The supplementary material supports scientific audit of the measurements. It contains rendered benchmark artifacts, auxiliary-experiment model responses, scorer outputs, stored judge verdicts, main-ladder aggregate summaries, and offline verification scripts, but it is not a deployment recipe. We do not release private user data, credentials, provider logs with account identifiers, provider-calling runner scripts, or non-anonymized author information. Downstream users should respect the source dataset licenses and avoid reidentification attempts.

Interpreting refusal behavior. One ethical risk is over-reading the hard-abstention result. A model that refuses a formal packet here is not necessarily safer in deployment, and a model that answers compact natural-language evidence is not necessarily unsafe. Refusal depends on instructions, provider policy, and application context; deployment teams should validate behavior on their own task distributions rather than transferring the benchmark result directly.

References

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2024. L-eval: Instituting standardized evaluation for long context language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14388–14411, Bangkok, Thailand. Association for Computational Linguistics.

Anthropic. 2026. Model ids and versioning. Claude API documentation. Accessed 2026-05-21.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

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

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

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, and 9 others. 2022. Improving language models by retrieving from trillions of tokens. *Preprint*, arXiv:2112.04426.

Jiamin Chen, Yuchen Li, Xinyu Ma, Xinran Chen, Xiaokun Zhang, Shuaiqiang Wang, Chen Ma, and Dawei Yin. 2025. Grounding long-context reasoning with contextual normalization for retrieval-augmented generation. *Preprint*, arXiv:2510.13191.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *Preprint*, arXiv:1901.02860.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2025. From local to global: A graph rag approach to query-focused summarization. *Preprint*, arXiv:2404.16130.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

Google. 2026. Gemini models. Google AI for Developers documentation. Accessed 2026-05-21.

Bernal Jimenez Gutierrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. HippoRAG: Neurobiologically inspired long-term memory for large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *Preprint*, arXiv:2002.08909.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. *Preprint*, arXiv:2011.01060.

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? *Preprint*, arXiv:2404.06654.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Atlas: Few-shot learning with retrieval augmented language models. *Preprint*, arXiv:2208.03299.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through memorization: Nearest neighbor language models. *Preprint*, arXiv:1911.00172.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Igorevich Sorokin, Artyom Sorokin, and Mikhail Burtsev. 2024. BABILong: Testing the limits of LLMs with long context reasoning-in-a-haystack. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Alex Laitenberger, Christopher D Manning, and Nelson F. Liu. 2025. Stronger baselines for retrieval-augmented generation with long-context language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32559–32569, Suzhou, China. Association for Computational Linguistics.

LangChain. 2026. ConversationSummaryMemory. Reference documentation, accessed 2026-05-21.

Ivan Yee Lee, Loris D'Antoni, and Taylor Berg-Kirkpatrick. 2026. The format tax. *Preprint*, arXiv:2604.03616.

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*, volume 33, pages 9459–9474.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *Preprint*, arXiv:2205.14334.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *Preprint*, arXiv:2303.16634.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2022. Entity-based knowledge conflicts in question answering. *Preprint*, arXiv:2109.05052.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *Preprint*, arXiv:2104.08786.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of llm agents. *Preprint*, arXiv:2402.17753.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *Preprint*, arXiv:2202.12837.

OpenAI. 2026. Models. OpenAI API documentation. Accessed 2026-05-21.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. 2024. Memgpt: Towards llms as operating systems. *Preprint*, arXiv:2310.08560.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *Preprint*, arXiv:2304.03442.

Daivik Patel and Shrenik Patel. 2025. ENGRAM: Effective, lightweight memory orchestration for conversational agents. *Preprint*, arXiv:2511.12960. ArXiv preprint; also available on OpenReview.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P. Lillicrap. 2020. Compressive transformers for long-range sequence modelling. In *International Conference on Learning Representations*.

Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. *Preprint*, arXiv:2308.01263.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *Preprint*, arXiv:2310.11324.

Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. 2023. Zeroscrolls: A zero-shot benchmark for long text understanding. *Preprint*, arXiv:2305.14196.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Preprint*, arXiv:2303.11366.

Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung yi Lee, and Yun-Nung Chen. 2024. Let me speak freely? a study on the impact of format restrictions on performance of large language models. *Preprint*, arXiv:2408.02442.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *Preprint*, arXiv:2305.14975.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multi-hop questions via single-hop question composition. *Preprint*, arXiv:2108.00573.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *Preprint*, arXiv:2212.10509.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023a. Large language models are not fair evaluators. *Preprint*, arXiv:2305.17926.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. 2024. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. *Preprint*, arXiv:2306.05087.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. 2025. Know your limits: A survey of abstention in large language models. *Preprint*, arXiv:2407.18418.

Bryan Wilie, Samuel Cahyawijaya, Etsuko Ishii, Junxian He, and Pascale Fung. 2024. Belief revision: The adaptability of large language models reasoning. *Preprint*, arXiv:2406.19764.

Brandon T. Willard and Rémi Louf. 2023. Efficient guided generation for large language models. *Preprint*, arXiv:2307.09702.

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

Yuhuai Wu, Markus N. Rabe, DeLesley Hutchins, and Christian Szegedy. 2022. Memorizing transformers. *Preprint*, arXiv:2203.08913.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. *Preprint*, arXiv:2305.13300.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for llms: A survey. *Preprint*, arXiv:2403.08319.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *Preprint*, arXiv:1809.09600.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi R. Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024a. R-tuning: Instructing large language models to say 'i don't know'. *Preprint*, arXiv:2311.09677.

Lei Zhang, Yunshui Li, Ziqiang Liu, Jiaxi yang, Junhao Liu, Longze Chen, Run Luo, and Min Yang. 2024b. Marathon: A race through the realm of long context with large language models. *Preprint*, arXiv:2312.09542.

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. 2024c. ∞bench: Extending long context evaluation beyond 100k tokens. *Preprint*, arXiv:2402.13718.

Zhehao Zhang, Weijie Xu, Fanyou Wu, and Chandan K. Reddy. 2025. Falsereject: A resource for improving contextual safety and mitigating over-refusals in LLMs via structured reasoning. In *Second Conference on Language Modeling*.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. *Preprint*, arXiv:2102.09690.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2023. Memorybank: Enhancing large language models with long-term memory. *Preprint*, arXiv:2305.10250.

A Data and Code Availability

For double-blind review, the FRAME source and evaluation artifacts are provided as anonymized supplementary material. The anonymized Zenodo record is https://doi.org/10.5281/zenodo.20347354. The archive contains per-question records and model responses for the auxiliary experiments (token-matched, deployed-template, matched-content, NL-vs-formal, retrieval-noise, HotpotQA, LoCoMo, and the raw-conversation baseline), rendered evaluation artifacts, stored LLM-judge verdicts, checksum manifests, and offline verification and aggregation scripts. For the main formal five-packet ladder, we release per-configuration correct counts, answer rates, and accuracies rather than per-question responses. Benchmarks: LongMemEval https://huggingface.co/datasets/xiaowu0162/longmemeval-cleaned, LoCoMo https://github.com/snap-research/locomo, and HotpotQA https://hotpotqa.github.io/. The supplementary README records upstream license notes and intended-use constraints for these artifacts. All calls used temperature 0 and had provider-native tools, search, and memory disabled. Max output tokens were 128 for reader calls except `gpt-5.2-pro`, which used the `v1/responses` endpoint with 1024 tokens (§8). Table 3 maps the table-facing model names to the exact API IDs used in the runs; provider documentation records model-code and versioning conventions (OpenAI, 2026; Anthropic, 2026; Google, 2026).

Table 3: Display names used in result tables and exact API IDs used in the runs.

Display name	Exact API ID
`gpt-5.4-mini`	`gpt-5.4-mini-2026-03-17`
`gpt-5.2`	`gpt-5.2-2025-12-11`
`gpt-5.2-pro`	`gpt-5.2-pro-2025-12-11`
`claude-sonnet-4`	`claude-sonnet-4-20250514`
`claude-opus-4.1`	`claude-opus-4-1-20250805`
`claude-opus-4.6`	`claude-opus-4-6`
`gemini-3-flash`	`gemini-3-flash-preview`
`gemini-2.5-flash-lite`	`gemini-2.5-flash-lite`
`gemini-3.1-flash-lite`	`gemini-3.1-flash-lite-preview`

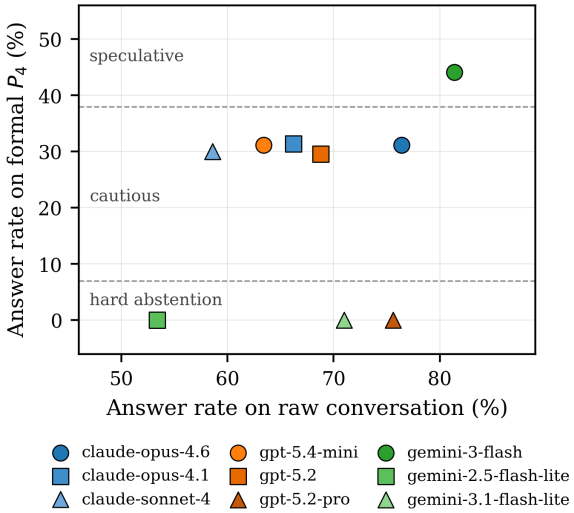

Figure 4: Representation-induced behavior regimes. Each dot is one model, plotted by answer rate on raw conversation (x) vs. answer rate on the formal P_4 packet (y).

Table 4: Replication on LoCoMo (200 questions). P_0 values, omitted for compactness, are 0.0–0.5%.

Model	Raw conv.	P_2 active
`claude-opus-4.6`	22.5	96.0
`claude-opus-4.1`	21.0	96.0
`claude-sonnet-4`	19.5	97.0
`gpt-5.4-mini`	18.0	94.5
`gpt-5.2`	21.5	83.5
`gpt-5.2-pro`	20.0	86.5
`gemini-3-flash`	16.0	32.0
`gemini-2.5-flash-lite`	18.5	82.5
`gemini-3.1-flash-lite`	17.5	94.5

B Additional Diagnostics and Replications

B.1 Representation-induced behavior regimes

Figure 4 visualizes the answer-rate regimes discussed in §5.1. The hard-abstention models have nonzero answer rates on raw conversation but near-zero answer rates on formal P_4 packets.

B.2 LoCoMo packet-ladder replication

The packet-ladder effect replicates on LoCoMo: P_0 values are 0.0–0.5%, while answer-bearing P_2 packets recover strongly (Table 4). Active-only accuracy exceeds raw-conversation accuracy here, but the substring scorer penalizes raw answers that express dates relatively; we interpret this table as a diagnostic replication rather than a general raw-vs-structured ranking.

Table 5: Matched-content control on LongMemEval (500 questions). Bold marks models that answer from witness text yet hard-abstain on resolved formal packets.

Model	Witness text	Active P_2
`claude-opus-4.6`	23.2	24.3
`claude-opus-4.1`	18.6	24.7
`claude-sonnet-4`	19.6	25.3
`gpt-5.4-mini`	21.6	22.9
`gpt-5.2`	18.6	24.3
`gpt-5.2-pro`	**18.0**	0.0
`gemini-3-flash`	14.2	15.0
`gemini-2.5-flash-lite`	**19.4**	0.0
`gemini-3.1-flash-lite`	**19.2**	0.0

B.3 Matched-content control

Table 5 separates content exposure from explicit conflict resolution. Witness text quotes the relevant turn verbatim but withholds the resolved slot and `active_value` field.

C Robustness: LLM-Judge Rescoring

Substring matching is a coarse scorer. We cross-validate it against two stronger signals: a rule-based normalized-EM baseline and an LLM judge.

C.1 Methodology

Full-population judge. The 31,500 (model, question, condition) triples in the token-matched (13,500) and real-system (18,000) experiments are rescored by `claude-haiku-4-5-20251001` at temperature 0 with a 4-token cap. The judge sees only the gold answer and the model response and outputs `CORRECT` or `INCORRECT`. All judge calls completed successfully. The exact system prompt is:

> You grade whether a model response correctly answers a factual question, given the gold answer. Output exactly one word: `CORRECT` or `INCORRECT`.
>
> Rules:
>
> - Accept semantic paraphrases that preserve the required information (e.g., "Denver" matches "Denver, CO"; "Jan 2024" matches "January 2024").
> - Accept normalized dates, numbers, and names when the referent is unambiguous.
> - Reject abstentions or refusals (e.g., "ABSTAIN", "I don't know").
> - Reject responses that omit required specificity (e.g., gold "John Smith" vs. response "John" is `INCORRECT`).
> - Reject hallucinations or answers that contradict the gold.
> - Be strict: when in doubt, prefer `INCORRECT`.

Table 6: Scorer agreement. The TM and RS columns report agreement with the *primary* judge on the full-population verdicts; the *Strat.* column reports agreement with the *second* judge on the 600-item stratified sample. The final row reports primary-vs-second-judge agreement on the cleaned 516-item overlap shared with the token-matched and real-system full-population rescores.

Comparison	TM (13.5K)	RS (18K)	Strat. (600)
sub. vs. prim. judge	87.2	87.9	68.8
norm. vs. prim. judge	89.5	89.5	72.8
sub. vs. norm.	96.5	95.8	56.3
prim. vs. 2nd judge	88.8	92.2	90.7†

†on cleaned 516-item overlap; TM = token-matched, RS = real-system; all values %.

The archive includes offline judge-audit and aggregation scripts: `scripts/rescore_with_judge.py` and `scripts/aggregate_full_judge.py`. These scripts recompute scorer agreement, judge aggregates, and bootstrap intervals from the stored verdicts without making provider calls.

Second-judge cross-validation. A stratified 600-item sample across all three headline experiments (including the 4,500-item matched-content control) is rescored by an independent second judge with the same prompt. The primary-vs-second comparison uses the 516 token-matched and real-system items with non-empty response text in both judge logs; matched-content items appear only in the stratified scorer columns because they are not part of the full-population judge pool. On this cleaned overlap the two judges agree on **90.7%** of items (92.2% real-system, 88.8% token-matched).

Normalized exact match. Rule-based baseline: lowercase, strip punctuation, drop articles, token-level substring containment in either direction, reject abstention tokens.

C.2 Agreement rates

Table 6 gives pairwise agreement between scorers. On the full population, substring and the primary judge agree on 87–88% of items and normalized-EM and the primary judge agree on 89–90%. On the cleaned 516-item overlap, the two judges agree on 90.7%.

C.3 Per-condition rescored accuracy

Table 7 gives per-condition accuracy under all three scorers, aggregated across nine models. Judge ac-

Table 7: Per-condition accuracy (%) under three scorers, aggregated across nine models. Full-population rescore for TM and RS; 600-item stratified rescore reweighted to population for matched-content. The matched-content row reports the analyzed row count after rescore filtering, hence n=4,423 rather than 4,500.

Condition	n	Sub.	Norm.	Judge
token-matched				
Full raw	4,500	38.5	38.6	50.5
Trunc. raw	4,500	10.3	11.3	15.2
Active P_2	4,500	71.4	73.0	**84.5**
real-system				
ChatGPT-mem	4,500	46.6	46.2	**57.9**
LangChain sum.	4,500	6.7	6.9	9.7
MemGPT typed	4,500	23.0	25.9	34.1
Raw conv.	4,500	37.8	38.2	50.0
matched-content				
Witness text	4,423	19.5	20.1	24.0

Table 8: Token-matched gap (P_2- trunc. raw) per model, under substring (original) and the Claude Haiku judge with paired bootstrap 95% CIs over 500 questions. No CI contains zero.

Model	Substr.	Judge	95% CI
`opus-4.6`	+65.6	+80.2	[76.6, 83.6]
`opus-4.1`	+61.2	+70.2	[66.0, 74.4]
`sonnet-4`	+68.2	+76.6	[72.6, 80.4]
`gpt-5.4-mini`	+61.4	+74.0	[70.0, 77.8]
`gpt-5.2`	+66.0	+75.2	[71.2, 79.0]
`gpt-5.2-pro`	+65.0	+73.8	[70.0, 77.6]
`gem-3-flash`	+42.4	+48.4	[43.6, 53.0]
`gem-2.5-lite`	+47.8	+48.8	[44.2, 53.4]
`gem-3.1-lite`	+72.6	+76.4	[72.6, 80.0]
Range	42.4–72.6	48.4–80.2	–

curacy is 3.8–13.3pp higher than substring on every condition; the relative ordering within each experiment is preserved. The judge is more permissive than either rule-based scorer because it accepts semantic paraphrases.

C.4 Headline token-matched gap: per-model CIs

Table 8 reports the per-model P_2- truncated-raw gap under substring and under the Claude Haiku judge, with paired bootstrap 95% CIs (10,000 resamples over the 500 shared questions). Every model's gap is positive and significant under both scorers. The judge gap ranges +48.4 to +80.2pp, and each per-model judge gap exceeds the corresponding substring gap; the 42.4–72.6pp substring range is a conservative lower bound relative to the semantic judge.

Table 9: Real-system ChatGPT-style memory − raw conversation per model, under substring and under the Claude Haiku judge, with paired bootstrap 95% CIs. Under the judge, the advantage is significantly positive on 3/9 models and significantly negative on 2/9.

Model	Substr.	Judge	95% CI
`opus-4.6`	+4.2	+5.6	[−0.2, +11.6]
`opus-4.1`	+15.6	+21.2	[+15.2, +27.2]
`sonnet-4`	+17.0	+24.0	[+17.6, +30.2]
`gpt-5.4-mini`	+5.0	+4.2	[−2.0, +10.4]
`gpt-5.2`	+8.6	+5.6	[−0.6, +11.8]
`gpt-5.2-pro`	−1.2	−8.0	[−13.8, −2.2]
`gem-3-flash`	−6.0	−8.2	[−13.6, −2.6]
`gem-2.5-lite`	+20.8	+20.6	[+14.2, +27.2]
`gem-3.1-lite`	+15.6	+6.0	[−0.2, +12.4]
Aggregate	+7.9	+7.9	–

C.5 Real-system ChatGPT-style vs. raw advantage: per-model CIs

Table 9 reports the real-system ChatGPT-style memory vs. raw-conversation comparison by model. Under the judge: 3 significant positive (`opus-4.1`, `sonnet-4`, `gemini-2.5-flash-lite`), 2 significant negative (`gpt-5.2-pro`, `gemini-3-flash`), 4 within sampling noise.

C.6 Late-abstention and substring-leakage cases in the primary substring scorer

The run-time scorer detects `ABSTAIN` via a *prefix* check on the response (equivalent to "starts with `ABSTAIN`" for most scripts, or "`ABSTAIN` appears in the first 20 characters" for the HotpotQA scorer). A response that first describes why it cannot answer, mentions the gold string in that explanation, and appends `ABSTAIN` at the end passes both clauses of the scorer and is counted correct. Three illustrative cases from the released archive:

- `retrieval_noise/results.json`, `anthropic/claude-opus-4.6`, key `gpt4_2312f94c|chatgpt_mem|k0`. Gold = "Samsung Galaxy S22." Response: "Based on the saved memories, I cannot determine which device you got first… ABSTAIN." Stored as `correct=true`.

- `retrieval_noise/results.json`, same model, key `gpt4_2312f94c|raw|k3`. Gold = "Samsung Galaxy S22." Response: "Based on the conversations, the Dell XPS 13 was obtained first…" (no `ABSTAIN`,

Table 10: Aggregate accuracy under the primary scorer (prefix-only abstain) vs. a strict scorer (ABSTAIN anywhere). Δ = strict − primary, in pp.

Condition	Prim.	Strict	Δ
token-matched			
c1_full_raw	38.51	38.09	−0.42
c2_truncated	10.29	10.22	−0.07
c3_active	71.42	71.36	−0.07
real-system			
ChatGPT-style memory	46.62	46.56	−0.07
LangChain-summary	6.69	6.56	−0.13
MemGPT-typed	23.04	22.91	−0.13
Raw conversation	37.78	37.18	−0.60
retrieval-noise k=9 agg.			
raw	36.06	35.61	−0.44
NL (ChatGPT-mem)	41.94	41.83	−0.11

but the substantive answer is the *other* device; the gold string appears later in the response's supporting context). Stored as correct=true.

- real_system/results.json, anthropic/claude-sonnet-4, qid a2f3aa27, condition c2_langchain_summary. Gold = "1300." Response: "You were nearing 1300 followers on Instagram as of May 28, 2023. However, I cannot determine your exact current follower count from these summaries." Stored as correct=true.

Tightening the abstention check to "ABSTAIN" in response.upper() (detect anywhere rather than only at the prefix) reclassifies 1 to 30 items per condition across the headline experiments and shifts aggregate accuracies as follows. This stricter ABSTAIN check addresses late-abstention cases, but it does not remove all substring false positives without an explicit abstention token:

For HotpotQA, each aggregate surface changes by at most 0.72pp under the strict scorer. The token-matched headline gap shifts negligibly (+61.1 vs. +61.1pp in aggregate; no per-model sign change). The real-system ChatGPT-minus-raw aggregate is essentially unchanged (+8.84 vs. +9.38pp). The c1_full_raw and c4_raw_conversation conditions take the largest hits because long raw-conversation responses have more chances to mention the gold string before abstaining, but the effect is under 1pp. No qualitative headline conclusion changes under the strict scorer.

D Retrieval Noise: Per-Model Detail

Section 4.4 reports the aggregate retrieval-noise result; per-model breakdowns and paired bootstrap 95% CIs are below.

NL is more stable than raw from k=0 to k=9 on 7 of 9 models (all except claude-opus-4.6 and gemini-3-flash). Paired bootstrap 95% CIs on the NL-over-raw gap at k=9: significantly positive on 4 (claude-sonnet-4 +15.5, gpt-5.4-mini +11.5, gemini-2.5-flash-lite +22.5, gemini-3.1-flash-lite +18.0pp); not significant on 3 (claude-opus-4.1 +0.5, gpt-5.2 +9.5, gemini-3-flash −4.5pp); negative, not significant on 1 (claude-opus-4.6 −10.0pp [−20.0, +0.5]); significantly negative on 1 (gpt-5.2-pro −10.0pp [−20.0, −0.5]). Only gpt-5.2-pro repeats the raw-over-NL preference in the HotpotQA evidence-surface transfer experiment; claude-opus-4.6 does not, so the reader-subclass pattern is suggestive rather than a fixed model list (§E).

E HotpotQA Evidence-Surface Transfer: Per-Model Detail

Section 4.4 reports the aggregate HotpotQA result; Table 12 gives the per-model breakdown.

NL beats typed on all 9 models: 6 significant under paired bootstrap (+6.5 to +10.5pp), 3 within sampling noise (claude-opus-4.6, claude-opus-4.1, gpt-5.4-mini). NL beats summary on 8/9 models (gpt-5.4-mini is −2.0pp ns). NL beats raw on 6/9 models (3 significant: gpt-5.4-mini +10.5, gemini-2.5-flash-lite +12.5, gemini-3.1-flash-lite +5.5pp); claude-opus-4.1 (−12.0pp) and claude-sonnet-4 (−10.0pp) prefer raw Wikipedia paragraphs, both significantly. gpt-5.2-pro narrowly prefers raw (−2.5pp ns), matching its retrieval-noise behavior in §D. The Anthropic reversals do not appear on conversational-dialogue LongMemEval, so the reader-class hypothesis in §6 is not fully general across domains.

Rendering setup. The summary and typed renderings are generated once per question by claude-opus-4.6 at temperature=0 and held fixed across evaluator models, so reader

Table 11: Per-model accuracy under retrieval noise, 200 LongMemEval questions × 9 models × 2 surfaces × 3 noise levels. "drop" columns show accuracy change from k=0 to k=9. Paired bootstrap 95% CIs for all drops are in Appendix data.

	Raw conversation				ChatGPT-mem (NL)			
Model	k=0	k=3	k=9	drop	k=0	k=3	k=9	drop
`claude-opus-4.6`	60.0	54.0	55.0	−5.0	50.5	47.5	45.0	−5.5
`claude-opus-4.1`	56.5	50.0	45.5	−11.0	52.5	47.0	46.0	−6.5
`claude-sonnet-4`	50.5	45.0	33.0	−17.5	47.0	48.5	48.5	+1.5
`gpt-5.4-mini`	33.5	27.0	24.5	−9.0	38.5	40.0	36.0	−2.5
`gpt-5.2`	46.5	40.5	36.5	−10.0	48.0	45.0	46.0	−2.0
`gpt-5.2-pro`	58.5	55.5	54.0	−4.5	49.0	45.0	44.0	−5.0
`gemini-3-flash`	28.5	27.0	27.0	−1.5	17.5	24.5	22.5	+5.0
`gemini-2.5-flash-lite`	31.0	24.5	19.0	−12.0	39.5	42.0	41.5	+2.0
`gemini-3.1-flash-lite`	39.5	32.5	30.0	−9.5	48.5	48.0	48.0	−0.5
Aggregate	44.9	39.6	36.1	−8.8	43.4	43.1	41.9	−1.5

Table 12: Per-model accuracy on HotpotQA (200 validation questions, distractor setting) across four evidence-surface renderings. Aggregate accuracy follows NL > raw > summary > typed; the top-level NL preference transfers, while the lower-order summary/typed ranking differs from the LongMemEval template-family comparison.

Model	Raw	NL	Sum.	Typed
`claude-opus-4.6`	80.0	**82.0**	79.5	78.0
`claude-opus-4.1`	**86.5**	74.5	69.0	72.0
`claude-sonnet-4`	**83.5**	73.5	72.5	67.0
`gpt-5.4-mini`	58.5	69.0	**71.0**	65.0
`gpt-5.2`	74.0	**76.5**	72.0	66.5
`gpt-5.2-pro`	**77.5**	75.0	74.0	65.5
`gemini-3-flash`	48.0	**50.0**	44.0	39.5
`gemini-2.5-flash-lite`	48.5	**61.0**	58.5	51.0
`gemini-3.1-flash-lite`	61.0	**66.5**	64.0	59.5
Aggregate	68.6	**69.8**	67.2	62.7

behavior is isolated from generator stochasticity. The NL surface is rule-based extraction of supporting-fact sentences (no LLM in the loop). `claude-opus-4.6` is both the generator and one of the evaluators; a generator/reader bias is possible in principle but Opus-4.6 does not systematically outperform other readers on the summary or typed conditions (Table 12).

F Future Work

End-to-end retriever integration. The retrieval-noise experiment (§4.4) injects k random distractors at the memory-store level. A deployed retriever produces query-correlated errors, not random ones. Running FRAME through a dense retriever at fixed recall budgets (e.g., 1.0 down to 0.5) would test whether the evidence-surface pattern survives realistic retrieval error rather than only random noise.

Code and table QA. We show NL-over-typed ordering on conversational memory (LongMemEval) and multi-hop Wikipedia QA (HotpotQA). A cleaner test of the pretraining-distribution-alignment hypothesis in §6 would run the same ladder on code- or table-grounded tasks where the typed surface dominates pretraining. If typed beats NL on those surfaces the pattern supports distribution alignment over the safety-heuristic account.