FROM RELEVANT FRAGMENTS TO SUFFICIENT ANSWERS: VERIFIABLE EVIDENCE CHAINS IN LONG DIALOGUES

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-context language models can process large dialogue histories, yet supplying more context does not reliably improve memory-based question answering. We study this through a fixed-panel reader experiment at increasing input lengths and through 50K-token, 100K-token, full-history, and oracle-evidence conditions on three long-dialogue benchmarks. Performance is non-monotonic and dataset-dependent: 50K packages are close to Full on RHELM and LoCoMo Long but substantially worse on EverMemBench, and coverage of Gold Evidence does not guarantee answer correctness. These gaps motivate a hypothesis of evidence sufficiency: when a complete answer is present in a finite textual context, at least one linear chain of source facts, events, and objects joined by verifiable lexical keys supports it. A manual audit of 24 blindly selected questions illustrates chain reconstruction and exposes limitations of official annotations, including dependencies on external metadata.

1 INTRODUCTION

A long dialogue history poses two related problems for a memory system: finding the information needed for a particular question in the primary source, and passing it to a model in a form that enables a justified answer. When the entire history is supplied, relevant utterances may be lost among distracting events, an effect described for long contexts by Liu et al. (2024). When only a short set of retrieved chunks is supplied, an intermediate fact may be omitted or a reference may be resolved incorrectly. Thus, the relevance of an individual fragment to the question does not establish the sufficiency of the evidence set as a whole.

We treat the original dialogue as material for retrieval and recomposition, rather than replacing it with a pre-extracted entity graph or a summary of events. We first measure how a reader uses increasingly long inputs on a fixed question panel. We then compare reduced packages of original text with complete histories and diagnostic Gold Evidence packages on three corpora. This separates context volume, coverage of annotated evidence, and answer correctness: none of these quantities alone describes the entire process.

The observed discrepancies lead to the main question of this work: how can we verify that retrieved fragments do not merely concern the question's topic, but form a sufficient textual basis for the answer? We formulate a hypothesis about chains of source facts, events, and object descriptions connected by verifiable, lexically expressed keys. Its scope is limited to finite textual contexts; metadata and other external information are considered separately. A manual question audit illustrates how proposed chains are constructed and checked, and identifies cases in which the question or Gold Evidence itself lacks a sufficient basis.

2 READER UNDER LONG CONTEXT

Even when a dialogue fits within a model's context window, a reduced package must preserve the evidence needed for an answer and present it at a usable volume.

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

In structured memory, extracted entities, events, and relations can facilitate retrieval (e.g., Chhikara et al., 2025). Yet that transformation adds an interpretive layer: qualifications, referents, or relations in the original text may be lost or recorded incorrectly. This does not imply that graph-based or other structured memory is generally unsuitable for dialogue systems. Here we test a narrower approach: select fragments of the original conversation for answering while minimizing the transmitted context and without replacing the primary source with an automatically composed summary.

Our starting point was an experiment on the effective context window of contemporary language models. We assembled a fixed panel of 20 LongMemEval questions (Wu et al., 2025) and selected GPT-5.6 Luna, whose advertised context window is 1,050,000 tokens (OpenAI, n.d.-a). By effective window, we mean not the technical ability to accept a request, but the ability to extract the information needed for a correct answer.

The panel contained 7 temporal-relation questions, 8 questions combining information across sessions, and 5 questions updating a previously stated fact. At every length level, the same 20 questions, prompt template, and complete source sessions containing the evidence were retained; whole distracting sessions were added to increase volume. Thus, both input length and the position of relevant evidence changed. No fragment retrieval was performed at this stage: the model answered from the prepared dialogue package. All conditions used GPT-5.6 Luna, reasoning_effort=low, and a 1,200-token output limit. Repeated runs at a given level received identical inputs and differed only in independent answer generation. Correctness was determined by a separate blind GPT-5.6 Sol judgment against the reference answer; negative judgments in upper-level runs were manually reviewed. The table reports mean *actual API input tokens*, including the question and prompt template, rather than the nominal package target. The panel and repeat-run logs are retained in the experiment records.

Table 1: Results at selected input lengths.

Mean actual input tokens	**Independent-run results**	**Mean correct-answer rate**
47,011	20/20; 19/20; 19/20	96.7%
178,194	15/20; 20/20	87.5%
264,041	14/20; 16/20	75.0%
435,976	15/20; 16/20	77.5%
606,324	17/20; 12/20	72.5%
891,466	11/20; 12/20	57.5%

The short control near 47K tokens yielded 96.7% correct answers. At 178K-606K tokens, mean results ranged from 72.5% to 87.5%, with repeated runs diverging at 606K (17/20 versus 12/20). At 891K, the mean fell to 57.5%, 39.2 percentage points below the short control. Longer inputs also added distracting sessions, changing both volume and the position of necessary evidence.

This observation raised a further question: can a long dialogue context be packed into a volume that the model handles effectively without losing the source facts, conclusions, and events needed for the answer?

3 THREE-DATASET RETRIEVAL SETTING

To address this question, we developed an experiment on selecting evidence from long dialogues. It uses three corpora with different actual history lengths: RHELM (Zhang et al., 2026), EverMemBench (Hu et al., 2026), and LoCoMo Long (Maharana et al., 2024). A Main panel of 100 questions was fixed for each corpus, and each question was tested under every comparison condition. The corpora were prepared differently: RHELM contains five long histories; EverMemBench uses histories from three topics constrained by a source-text budget; and ten original LoCoMo dialogues were supplemented with independent distracting sessions. Full-history lengths were therefore not equalized and should not be treated as a property of three identically constructed datasets.

In the retrieval branch, each history was split into 1,024-token windows with a 256-token overlap. We computed 3,072-dimensional OpenAI text-embedding-3-large (OpenAI, n.d.-b) vectors for the windows. Queries were evaluated in parallel by dense search and BM25 (Robertson and Zaragoza, 2009); their ranks were merged by reciprocal rank fusion (Cormack et al., 2009), after which a

limited candidate pool was reranked with Voyage rerank-2.5 (Voyage AI, n.d.). Overlapping portions of the source text were not duplicated when assembling the reader input. Candidate-pool size and packing rules differed among panels; their configurations are retained in the experiment records.

We compared four reader-input conditions: 50K and 100K, upper budgets for selected original context; Full, the entire available history; and Oracle, a diagnostic package of annotated evidence.

Oracle's contents depended on the annotations of each corpus and were not produced by automated retrieval. GPT-5.6 Luna generated answers in every condition with low reasoning effort. Final answer accuracy was calculated from blind binary judgments by two independent judges, GPT-5.6 Sol and Terra; coverage of officially annotated evidence was measured separately from answer correctness. The fixed LongMemEval panel of Section 2 was not included in this three-corpus Main evaluation.

Table 2: Mean actual reader input, thousands of tokens; 100 questions per corpus.

Corpus	**50K**	**100K**	**Full**	**Oracle**
RHELM	49.8	99.8	639.5	1.33
EverMemBench	36.1	44.6	574.3	3.50
LoCoMo Long	29.6	29.9	349.9	0.11

The reduced conditions passed substantially less text to the reader than Full. Increasing the budget from 50K to 100K barely changed the mean package size on LoCoMo and increased it only moderately on EverMemBench: on those corpora, the two budgets produced identical packages for 96/100 and 68/100 questions, respectively. A larger budget does not by itself add new retrieved fragments.

Table 3: Questions for which a reduced package fully covered official Gold Evidence.

Corpus	**50K**	**100K**
RHELM	61/100	63/100
EverMemBench	5/100	5/100
LoCoMo Long	79/88	79/88

Gold Evidence coverage changed little when the budget increased to 100K and was particularly low on EverMemBench. The LoCoMo denominator includes only the 88 questions with applicable Gold annotations, and annotated units differ among corpora. Complete coverage of official annotations means that marked fragments were present, not that they suffice for an answer.

Table 4: Answer accuracy on the Main panels: mean of two judges' binary assessments, 100 questions per corpus.

Corpus	**50K**	**100K**	**Full**	**Oracle**
RHELM	66.5%	63.0%	66.0%	70.5%
EverMemBench	55.5%	57.5%	72.5%	89.5%
LoCoMo Long	62.5%	62.5%	55.5%	64.5%

Context reduction did not have a uniform effect across corpora. For 50K, the difference from Full was +0.5 percentage points on RHELM, -17 on EverMemBench, and +7 on LoCoMo. Question-level 95% bootstrap intervals for RHELM [-7.5, +8.5] and LoCoMo [-1.5, +15.5] include zero; that for EverMemBench [-26, -8] does not. Thus, the observed failure of the reduced condition on EverMemBench should not be generalized to all three corpora, and the numerical differences on RHELM and LoCoMo should not be called established improvements.

4 WHERE ANSWER EVIDENCE IS LOST

Tables 2-4 describe different stages of one process: how much original text reached the reader, which annotated evidence was included, and whether the model produced a correct answer. They do not assign one cause to all errors. Misses already occur at retrieval: on RHELM, the full set of annotated target utterances was present in the candidate pool for 63/100 questions, but after

reranking and 50K packing, for 61/100. Almost all incomplete-coverage cases there therefore arose before final packing, not because a 50K budget truncated already retrieved evidence.

Gold coverage is not answer sufficiency. On RHELM under 50K, nineteen answers were judged correct despite missing at least one annotated utterance, whereas fifteen were judged incorrect despite complete annotation coverage. On LoCoMo, complete Gold coverage was attained for 79 of 88 applicable questions, while 50K accuracy was 62.5% and Oracle accuracy was 64.5%. Official annotations and answer judgments measure different things; without reading the utterances, one cannot automatically distinguish insufficient Gold Evidence from a reader error.

On EverMemBench, all annotated messages were included in the reduced package for only 5/100 questions. Under 50K, multi-hop questions yielded 8/33 correct answers, compared with 30/34 for single-hop questions. This indicates particular difficulty with dispersed evidence. An additional audit of 25 deliberately selected difficult cases found that every answer depended on technical message metadata: 23 required timestamps to order two events, one required message dates to establish working-day boundaries, and one required speaker and project-group information for attribution. Thus, 24/25 cases depended on service time. These questions test not only relations in dialogue text, but also correct attachment of events to message structure; quality losses cannot be reduced to retrieval of topically relevant fragments.

The main question raised by this experiment goes beyond retrieval quality: how are the dispersed facts, events, and object descriptions needed for a justified answer connected within the original dialogue? The next section proposes a testable hypothesis: these connections can be traced as a linear chain of textual references from question to answer. We must first separate dialogue text from additional information, lest a connection reconstructed from metadata be incorrectly attributed to the text.

5 SEMANTIC-CHAIN HYPOTHESIS

Textual cohesion through reference and lexical reiteration (Halliday and Hasan, 1976), discourse coherence through relations and inferences (Hobbs, 1979), and lexical chains (Morris and Hirst, 1991) provide related accounts of textual connection. In multi-hop QA, HotpotQA (Yang et al., 2018) annotates supporting sentences and MuSiQue (Trivedi et al., 2022) requires connected steps; WorldTree V2 (Xie et al., 2020) and EntailmentBank (Dalvi et al., 2021) construct explanatory graphs and trees. We instead examine linear reference chains in the original dialogue, without a separate explanation structure, distinguishing dialogue text from other information available to participants.

Scope of the hypothesis.

The contract T is the complete communicative frame, including the original dialogue text and additional information:

$$T = (C, R)$$

where:

- C is the original textual context of the dialogue;
- R is information available to participants outside that dialogue: preceding context, roles, message metadata, headers, and high-level descriptions.

The rules of language, reference resolution, and textual coherence are treated as fixed interpretive conditions rather than additional facts in R. Related ideas of common ground in communication were discussed by Clark and Brennan (1991), and of linguistic coordination by Lewis (1975).

In this analysis we use only the dialogue text: additional information is excluded ($R = \varnothing$).

5.1 INITIAL HYPOTHESIS

If an answer is complete and sufficient for the question, it cannot be entirely lexically disconnected from that question within the original context.

$$\mathrm{Full}(C, Q, A) \Rightarrow \exists K_0 : \; Q \overset{K_0|C}{\longleftrightarrow} A$$

Notation:

- C is the finite textual dialogue context within the stated scope;
- Q is the question;
- A is its answer;
- $\mathrm{Full}(C, Q, A)$ means that A is complete, sufficient for Q, and supported by C; this criterion is defined independently of the chain;
- K_0 is a lexically expressed key linking Q and A through C; it is drawn from Q but is not fixed to the chain's first transition;
- $Q \overset{K_0|C}{\longleftrightarrow} A$ means that the connection between Q and A through K_0 is traceable in C; wording may change across references, but each connection must be checked.

Drawing on Clark and Brennan's (1991) concept of common ground, we treat the original dialogue text in the present case as a record of a contract between participants. This is a shared field of facts, events, and assertions that participants regard as agreed in advance or establish during the conversation. Every element of this contract is available to both participants for interpreting utterances; a connection between distinct elements must still be shown in the text itself.

Let $F_i(P)$ denote participant P's interpretation of agreed fact i. Let P_1 ask a question and P_2 answer. In the present model we assume:

$$F_i(P_1) = F_i(P_2)$$

This equality records a shared interpretation of the same agreed fact: when referring to it, both participants rely on the same content within contract T.

Next, we trace how an agreed fact refers to an object in the text. In Clark and Wilkes-Gibbs (1986), interlocutors jointly establish which object an expression denotes and how it may be referred to. A reference may be expressed by a name, pronoun, abbreviation, or another lexical construction that directly points to the original object.

Checking references first requires delimiting the contract and establishing the presence of the original object within it. If the contract is infinite, a chain of references to an object may also be infinite, and the object may remain unidentified. We therefore take the original dialogue text to be finite in the present model. Tracing references to earlier expressions either yields an unambiguous designation of the object or reveals a broken chain. In the latter case the reference remains unresolved, and the assertion containing it cannot provide an agreed basis for a complete answer. Thus, in an admissible complete answer, the object is identified through a finite chain of references in the original text.

The same rule applies to facts, events, and assertions that refer to an object or to one another. If such a connection is expressed by a pronoun, abbreviation, alternative name, or another textual reference, it can be traced backward to its original fragment. Each accepted connection of this type in the contract can therefore be reconstructed as a finite reference chain; a broken chain leaves the connection unconfirmed.

The shared interpretation of agreed facts and the presence of the original object in the contract give us a connection between question and answer. If the question concerns an object, a respondent giving a complete and sufficient answer must identify the same object in the dialogue text. The participants may name it differently, for example, "Michael" and "boy", but both expressions point to the same original object in the contract. A lexically expressed key from the question can therefore be traced through the dialogue to the answer, as asserted by the initial formula.

While working with the initial hypothesis of lexical correspondence between question and answer, we formulated an extended version:

If the original context contains a complete and sufficient answer to a question, it contains at least one linear chain of facts joined by lexically expressed keys that connects the question to the answer and justifies it.

$$\mathrm{Full}(C,Q,A) \Rightarrow \exists K_0\, \exists G\, \big(G \in \mathcal{G}(C,Q,A) \wedge \mathrm{Trace}_C(G,K_0,A)\big)$$
$$G = Q \overset{K_1}{\longleftrightarrow} Z_1 \overset{K_2}{\longleftrightarrow} \cdots \overset{K_n}{\longleftrightarrow} Z_n \overset{K_{n+1}}{\longleftrightarrow} A$$
$$Z_i = (k_i, V_i)$$

Notation:

- $\mathcal{G}(C,Q,A)$ is the set of admissible chains from C whose facts jointly suffice to answer every part of Q with A;
- G is one chain in $\mathcal{G}(C,Q,A)$; its link order need not follow event chronology or fragment order in C, but every adjacent transition must be verifiable;
- Z_i = (k_i, V_i) is a link consisting of local key k_i and its neighborhood V_i, containing a C-supported statement about a fact, event, or object;
- k_i is the key used to extract Z_i; it may coincide with a transition key K;
- V_i is the neighborhood of k_i that preserves an assertion minimally sufficient for Q or part of Q. Its length is not fixed: a compound fragment is permitted when needed to preserve meaning and the basis of a connection; the mechanically shortest fragment is not required;
- K_0 is the key linking Q to A through C;
- K_1, ..., K_{n+1} are lexically expressed transition keys in G: K_1 links Q to Z_1, intermediate keys connect neighboring links, and K_{n+1} links Z_n to A. A key may be a repeated word, phrase, or inflected form of the same lemma, or a pair of textual expressions that unambiguously designate the same object in C. For a pronoun, alternative name, or description, the reference basis is checked against C; class-name similarity is insufficient;
- $\mathrm{Trace}_C(G,K_0,A)$ is a condition on a completed chain G: the K_0 connection between Q and A appears in at least one link and remains traceable through C without switching the designated object or the question's target.

The initial formula $\mathrm{Full}(C,Q,A) \Rightarrow \exists\ K_0 : Q \overset{K_0|C}{\longleftrightarrow} A$ entails that K_0 exists for a complete and sufficient answer. Thus, a chain G supporting such an answer must preserve this connection: at least one of its links contains a lexically expressed reference through which K_0 can be traced from the question to the answer. This fixes neither the number of links nor the position of that link in G.

To connect links, we extend the same principle from an individual object to a fact. A fact is a larger unit of context: it may include events, objects, and nested assertions. Its boundaries depend on the content needed for the answer, not on a fixed fragment length. A reference to a fact may be direct or may use an object it contains, a pronoun, an abbreviation, or another lexical construction pointing to that specific fact.

Consider any two adjacent links. If one refers to a fact or object in the other, that reference can be traced through finite C to its basis in the same way as for the initial hypothesis. An established reference yields a transition key between the links. If it breaks or fails to identify a unique basis, the transition cannot support the answer. The rule is the same for every adjacent pair and does not depend on chain length; transitions from Q to the first link and from the last link to A are checked in the same manner.

Thus, K_0 links question and answer, while verified transition keys connect the links of a proposed linear chain. The sufficiency of the whole chain for A is established from its assertions taken together.

This provides a rule for checking a proposed chain G, but does not establish that such a chain exists for every $\mathrm{Full}(C,Q,A)$ instance. That existence claim remains under investigation and is not yet supported by a complete mathematical proof.

5.2 Sequential chain check

$$B_0 = Q; \quad B_{i+1} = B_i \overset{K_{i+1}}{\longleftrightarrow} Z_{i+1}.$$
$$\Phi_i = \Phi(C,Q,B_i); \quad \Delta\Phi_{i+1} = \Phi_{i+1} - \Phi_i.$$

$$G = B_n \overset{K_{n+1}}{\longleftrightarrow} A.$$

Notation:

- i is the construction-step number;
- B_0 = Q is the initial chain state;
- B_i for $i \geq 1$ is the chain after i links, ending at Z_i; B_{i+1} is the chain after appending the next link;
- Z_{i+1} is the next link from C, joined by verified key K_{i+1} to the preceding Z_i, or to Q on the first step;
- Φ(C, Q, B_i) assesses how far the links in B_i, already assembled and confirmed against C, cover the parts of Q; it neither tests for K_0 nor replaces Trace_C. Facts in C that have not entered B_i are not part of the answer;
- Φ_i is a formal numerical assessment of answer completeness supported by B_i for Q;
- $\Delta\Phi_{i+1}$ is the change after adding Z_{i+1}: current Φ_{i+1} minus previous Φ_i.

Thus, Φ assesses the completeness of the answer supported by current chain B_i, and $\Delta\Phi_{i+1}$ shows the change after adding Z_{i+1}. After closing B_n on A, $\mathrm{Trace}_C(G, K_0, A)$ is checked for the resulting candidate chain. The difference formula does not specify how the next link is selected. In the manual analysis, a human and an LLM checked completeness; numerical Φ_i values were not measured. Accepting a chain as G $\in \mathcal{G}(C, Q, A)$ additionally requires checking that its facts suffice for all parts of Q and that the final transition to A holds.

6 Source-locked audit and Dense comparison

The manual experiment used RHELM, EverMemBench, and LoCoMo, together with the chunks and embeddings prepared for the main experiment (Section 3). Eight questions were selected blindly from each corpus, without prior inspection of their contents by the author; 24 questions were examined in total. Each case was analyzed separately. The author and an LLM jointly assembled chains using BM25 search over the original context, but the author checked and ultimately approved every proposed fact and transition.

First, the LLM extracted initial lexical keys from the unchanged question and used them as BM25 queries. For each question we created a two-column table: the search key on the left, and each exact occurrence in an original utterance, with up to eight surrounding words on each side, on the right. Occurrences were ordered by the BM25 rank of their containing chunk, retaining lower-ranked matches. When a short neighborhood did not suffice to understand an assertion, a larger passage of the original utterance was opened. We then analyzed the question itself: which objects and events it concerned, and which relations and answer components had to be established. Gold Evidence was not used at these stages.

The occurrence table yielded an initial relevant assertion. We then searched for further facts through its expressed keys and tested alternative ways of linking them. Each link had to preserve a verifiable source fact, and each transition had to connect adjacent links through a shared lexically expressed key referring to the same object, event, or assertion. Where alternatives existed, the author reviewed the primary source, rejected false matches, and selected a chain addressing all essential parts of the question. Merely collecting topically similar phrases was insufficient: answer completeness, unambiguous references, and preservation of source meaning were checked separately. A transition or answer component that could not be justified was marked as a break, not filled by conjecture.

Two techniques were used to delimit links. *Escalation* traced a reference, such as "he", through neighboring sentences to a person or object named unambiguously earlier. This established who a needed fact concerned; the recovered name was not presented as a literal word of the source utterance. *Surgery* removed parts of an utterance irrelevant to the question, including a second object's description, while retaining a minimally sufficient grammatical assertion. The surviving words kept their original order. After shortening, we rechecked grammar and verified that negation, condition, time, participant, or another meaning-changing qualification had not been removed. Sometimes both techniques were applied to one utterance.

For example, an original utterance says "I finally managed to meet with Jeroen at the electronics shop about my laptop situation", followed by "He set me up with a temporary external monitor and keyboard". Escalating from "He" to the unambiguously named person yields "Jeroen set me up with a temporary external monitor and keyboard"; this is a composite extraction, not a continuous quotation. Surgery shortens "the monitor kept flickering and wouldn't sync properly" to "the monitor wouldn't sync properly": the flickering detail, unnecessary for the answer, is removed while the synchronization-failure fact remains.

After author approval, the links were joined into one textual Dense query without chain delimiters. Repeated fragments at link boundaries were removed when doing so preserved each assertion's grammar and meaning. The query ranked the same source chunks in the prebuilt Dense index; chunk boundaries and embeddings were unchanged for the comparison. Only then did we open official Gold Evidence, identify the ranks of its chunks, and separately mark source chunks used by our chain or containing additional facts needed for the answer. A second Dense query concatenated the bodies of all Gold Evidence utterances and ranked the same index. The final cross-table recorded, for *both queries*, the ranks of *both chunk sets*: officially annotated chunks and those selected during the manual audit. The table and the decision about answer sufficiency were recorded for each question. Corrections made after Gold inspection were saved as separate versions rather than substituted for the initial result.

7 Results and evidence-quality audit

Table 5: Mean Dense rank (lower is better) on the same union of audited and Gold chunks for both queries; shared chunks count once, and n is the union size.

Corpus	Question	n	Our chain	Gold Evidence
LoCoMo	conv-41-q077	3	15.33	26.00
LoCoMo	conv-41-q083	–	–*	–*
LoCoMo	conv-41-q099	2	14.00	23.50
LoCoMo	conv-42-q002	8	9.00	25.00
LoCoMo	conv-42-q019	1	1.00	1.00
LoCoMo	conv-42-q020	3	2.00	2.00
LoCoMo	conv-42-q025	3	29.00*	26.67*
LoCoMo	conv-42-q039	1	1.00	1.00
RHELM	fact_b86427	–	–*	–*
RHELM	fact_90d49f	4	157.25*	165.00*
RHELM	temporal_0cdf60	4	43.25*	15.00*
RHELM	fact_2a5182	–	–*	–*
RHELM	fact_5a7990	3	41.00*	60.33*
RHELM	fact_82e780	4	252.00*	45.00*
RHELM	aggregation_9e30df	6	100.50*	82.83*
RHELM	fact_b530ec	6	41.83*	26.33*
EverMemBench	F_MH_Top03_007	22	220.82	145.86
EverMemBench	F_MH_Top03_011	32	143.94*	193.16*
EverMemBench	F_MH_Top03_019	22	64.77*	111.55*
EverMemBench	F_MH_Top04_003	31	146.52	185.32
EverMemBench	F_MH_Top04_039	32	176.06	193.59
EverMemBench	F_MH_Top05_030	31	136.97	198.42
EverMemBench	F_SH_Top03_041	13	31.31*	92.85*
EverMemBench	F_MH_Top05_047	–	–	–

An asterisk marks a question discussed in Table 6. A dash means that no numerical rank pair was accepted for that entry, not a zero rank.

Table 6: Issues found in question wording and Gold Evidence annotations.

Question	Observed issue
conv-41-q083	The source does not confirm the awarded certificate presupposed by the question; Gold shows neither the certificate nor study completion.
conv-42-q025	The official answer drops "two weekends later"; Gold omits the preceding utterance establishing the time reference.
fact_b86427	Gold does not support the question's "at home" qualification for the piano activity.
fact_90d49f	Neither Gold nor the breakfast statement establishes that the sweet potatoes came from the March 11 purchase.
temporal_0cdf60	Gold omits the named therapy session used as the question's event-order reference.
fact_2a5182	The directly relevant source episode contradicts the question's premise.
fact_5a7990	Gold omits the direct statement that Jasper added his drawing to the blog post, though the answer remains contextually supported.
fact_82e780	The episode has no retelling of the bicycle discovery to the group; Gold contains only route utterances.
aggregation_9e30df	Gold points to a snack discussion, not the most frequent dinner dish in the cooking journal for the stated period.
fact_b530ec	Test results were pending at the cafe visit; Gold also omits an earlier meal with Lucas needed for the comparison.
F_MH_Top03_011	Gold includes far more messages than needed to establish the two dated events used in the calculation.
F_MH_Top03_019	Two event messages suffice for the interval; the other 35 of 37 Gold messages are unnecessary.
F_SH_Top03_041	Gold omits the project's source designation needed to identify the task and includes unnecessary messages.

8 Discussion, limitations, and conclusion

We began by testing the effective, rather than merely advertised, context window of GPT-5.6 Luna. At tested input lengths from 47K to 891K tokens, fact-extraction quality was non-uniform: the highest effectiveness was observed at up to 50K tokens, with variable regression at longer inputs. This motivated reducing the context passed to the model while retaining the original facts and their relations. We next compared, on three datasets, a setup intended as an alternative to graph-based and causal memory systems: dialogue was not transformed into a separate backend structure; its original fragments were used to retrieve facts, events, and relations. This approach allows blocks of context to be mechanically recomposed without distorting or paraphrasing nested assertions and causal connections. Even at this technically simple stage, results broadly comparable to Full were obtained on long histories exceeding 50K-100K tokens. Dialogue compaction remains an open question, but in our view, working from the primary source is among the most promising directions for these systems, as distinct from graph structures and separate stores of extracted entities.

The new hypothesis developed during this experiment is still being examined with respect to a complete mathematical proof. Its principles have, however, been tested on questions with complex causal relations: in development runs, we examined many cases from the datasets in detail. In the examined cases, we found no grounds for a counterexample to the hypothesis. Difficulties that arose concerned information outside the primary context, namely metadata and additional annotations needed to answer the questions. When this external frame was taken into account, the examined questions did not exhibit identification problems requiring further metadata-based clarification.

Twenty-four questions were selected blindly; more than half had Gold Evidence problems, including missing facts, incorrect annotation, answer noise, and other inconsistencies. Public Gold Evidence should therefore not be treated as exhaustive justification. In these dialogue corpora, annotations identify answers and individual evidence pieces, but not complete chains of facts and connections. Annotating those chains could sharpen retrieval benchmarks and distinguish retrieval errors, insufficient evidence, and answer-construction errors.

REFERENCES

Chhikara, P., Khant, D., Aryan, S., Singh, T., & Yadav, D. (2025). Mem0: Building Production-Ready AI Agents with Scalable Long-Term Memory. https://arxiv.org/abs/2504.19413. arXiv:2504.19413.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. https://web.stanford.edu/~clark/1990s/Clark%2C%20H.H.%20_%20Brennan%2C%20S.E.%20_Grounding%20in%20communication_%201991.pdf.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. https://web.stanford.edu/~clark/1980s/Clark,%20H.H.%20_%20Wilkes-Gibbs,%20D.%20_Referring%20as%20a%20collaborative%20process_%201986.pdf. *Cognition*, 22(1), 1-39.

Cormack, G. V., Clarke, C. L. A., & Büttcher, S. (2009). Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. https://doi.org/10.1145/1571941.1572114. *SIGIR '09*, 758-759.

Dalvi, B., Jansen, P., Tafjord, O., Xie, Z., Smith, H., Pipatanangkura, L., & Clark, P. (2021). Explaining Answers with Entailment Trees. https://aclanthology.org/2021.emnlp-main.585/. *EMNLP*, 7358-7370.

Halliday, M. A. K., & Hasan, R. (1976). Cohesion in English. https://www.routledge.com/Cohesion-in-English/Halliday-Hasan/p/book/9780582550414. Longman.

Hobbs, J. R. (1979). Coherence and coreference. https://onlinelibrary.wiley.com/doi/10.1207/s15516709cog0301_4. *Cognitive Science*, 3(1), 67-90.

Hu, C., Li, T., Gao, X., et al. (2026). Evaluating Long-Horizon Memory for Multi-Party Collaborative Dialogues. https://arxiv.org/abs/2602.01313. arXiv:2602.01313.

Lewis, D. (1975). Languages and language. https://sites.ualberta.ca/~francisp/Phil426/LewisLanguagesLanguage75.pdf. In K. Gunderson (Ed.), *Language, Mind, and Knowledge*, 3-35. University of Minnesota Press.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the Middle: How Language Models Use Long Contexts. https://aclanthology.org/2024.tacl-1.9/. *Transactions of the Association for Computational Linguistics*, 12, 157-173.

Maharana, A., Lee, D.-H., Tulyakov, S., Bansal, M., Barbieri, F., & Fang, Y. (2024). Evaluating Very Long-Term Conversational Memory of LLM Agents. https://aclanthology.org/2024.acl-long.747/. *ACL*, 13851-13870.

Morris, J., & Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. https://aclanthology.org/J91-1002/. *Computational Linguistics*, 17(1), 21-48.

OpenAI (n.d.-a). GPT-5.6 Luna model documentation. https://developers.openai.com/api/docs/models/gpt-5.6-luna.

OpenAI (n.d.-b). Embeddings guide: text-embedding-3-large. https://developers.openai.com/api/docs/guides/embeddings.

Robertson, S., & Zaragoza, H. (2009). The Probabilistic Relevance Framework: BM25 and Beyond. https://www.nowpublishers.com/article/DownloadEBook/INR-019. *Foundations and Trends in Information Retrieval*, 3(4), 333-389.

Trivedi, H., Balasubramanian, N., Khot, T., & Sabharwal, A. (2022). MuSiQue: Multihop questions via single-hop question composition. https://aclanthology.org/2022.tacl-1.31/. *Transactions of the Association for Computational Linguistics*, 10, 539-554.

Voyage AI (n.d.). rerank-2.5 documentation. https://docs.voyageai.com/docs/reranker.

Wu, D., Wang, H., Yu, W., Zhang, Y., Chang, K.-W., & Yu, D. (2025). LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory. https://proceedings.iclr.cc/paper_files/paper/2025/hash/d813d324dbf0598bbdc9c8e79740ed01-Abstract-Conference.html. *ICLR*.

Xie, Z., Thiem, S., Martin, J., Wainwright, E., Marmorstein, S., & Jansen, P. (2020). WorldTree V2: A Corpus of Science-Domain Structured Explanations and Inference Patterns supporting Multi-Hop Inference. https://aclanthology.org/2020.lrec-1.671/. *LREC*, 5456-5473.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W., Salakhutdinov, R., & Manning, C. D. (2018). HotpotQA: A dataset for diverse, explainable multi-hop question answering. https://aclanthology.org/D18-1259/. *Proceedings of EMNLP*, 2369-2380.

Zhang, H., Tang, Z., Yu, X., et al. (2026). Beyond Static Dialogues: Benchmarking Realistic, Heterogeneous, and Evolving Long-Term Memory. https://arxiv.org/abs/2605.31086. arXiv:2605.31086.

AI USE STATEMENT

Generative models were used in the experiments to generate and evaluate answers and, in the manual audit, to identify search keys and suggest chain links. During preparation of this work, AI was also used to discuss and refine the hypothesis, mathematical formulations, and proof sketches; develop methodology; write and revise experimental code; analyze literature and individual cases; interpret results; translate the manuscript into English; and draft and edit text. The author checked proposed facts and transitions in the manual audit against the original dialogues, made the final decisions, and takes responsibility for the paper's contents.