CORAL: CONTEXT-ORIENTED RETENTION WITH ANCHOR-BASED LONG-TERM MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) increasingly rely on long-term conversational memory to maintain consistency across multi-turn and multi-session interactions. However, continuously accumulating uncompressed conversational histories introduces substantial redundancy, making task-relevant information increasingly difficult to retrieve and utilize. To address these challenges, we introduce CORAL (Context-Oriented Retention with Anchor-Based Long-Term Memory), a Multi-Agent System for constructing compact and retrieval-efficient conversational memory. CORAL first extracts salient information as structured anchors and then uses these anchors to guide LLM-based compression under a constrained memory budget. The resulting memory is indexed using Retrieval-Augmented Generation (RAG), enabling relevant evidence to be retrieved for downstream reasoning. By separating information preservation, compression, and query-time reasoning, CORAL aims to retain information with high future utility while removing repetitive or low-value conversational content. We evaluate CORAL on LoCoMo using question answering and event summarization, and systematically analyze factors such as compression fraction, compressor size, compression method, and model family. Overall, our results show that effective conversational memory compression requires preserving the right information, rather than simply retaining more content or using larger models. The code is anonymously available online.[1]

1 INTRODUCTION

Traditional multi-turn dialogue systems suffer from several challenges. First, they have difficulty maintaining contextual coherence in scenarios involving long-term dependencies. Second, they typically require extensive domain knowledge acquired through curated task-specific training data, as well as explicit dialogue state tracking. Third, user inputs can be diverse and unexpected and may not conform to patterns seen during training. Finally, such systems often struggle to capture implicit meaning, leading to semantically inaccurate or inappropriate responses (Yi et al., 2025). In contrast to traditional task-oriented dialogue pipelines, numerous approaches have been proposed to advance end-to-end task-oriented dialogue (EToD) research (Qin et al., 2023). These approaches can be broadly categorized into two types: modularly end-to-end systems (Mehri et al., 2019; Le et al., 2020), which consist of four interconnected modules—natural language understanding (NLU), dialogue state tracking (DST), dialogue policy learning (DPL), and natural language generation (NLG)—and fully end-to-end systems (Eric & Manning, 2017; Wu et al., 2019), typically formulated as unified sequence-to-sequence models. More recently, with the emergence of Large Language Models (LLMs), the performance of dialogue systems has significantly improved. Large-scale pretraining has endowed LLMs with broad domain knowledge, while Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) further enables access to external knowledge, thereby enhancing their generality. Accordingly, challenges related to maintaining context across multiple turns, acquiring domain knowledge, handling diverse scenarios, and understanding implicit context have been alleviated to a considerable extent.

However, long context is not a complete solution to long-term conversational memory. As dialogue accumulates, retaining the full history increases computational cost and introduces substantial re-

[1] https://anonymous.4open.science/r/CORAL-3C63/

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.

dundancy. RAG reduces the amount of context passed to the LLM, but still relies on the quality of the underlying memory: uncompressed histories may retrieve irrelevant content, while conventional summaries may discard details needed later. This motivates a central question: what should be retained when future information needs are unknown? We view long-term conversational memory as a problem of *selective memory construction*, where redundancy is removed while potentially useful information—including preferences, goals, events, relations, and temporal cues—is preserved. To address this problem, we propose CORAL, a multi-agent memory compression system consisting of a *Preservation Agent*, a *Compression Agent*, and a *Conversation Agent*. Together, they identify salient information, construct compact memory under a token budget, and retrieve it for downstream reasoning.

We evaluate CORAL on LoCoMo across compression fractions, model scales, compression methods, and model families using question answering and event summarization tasks. Our contributions are: (1) formulating long-term memory as selective memory construction; (2) introducing CORAL for structured preservation and budget-constrained compression; (3) proposing anchor-guided compression; (4) systematically evaluating memory efficiency and downstream utility and (5) releasing the code of CORAL and experiments for reference.

2 Related Work

2.1 Benchmark

Early work on long-term dialogue extended conversational context across multiple sessions. Multi-Session Chat (MSC) Xu et al. (2022) introduced repeated human-human open-domain interactions, while LoCoMo Maharana et al. (2024) constructed much longer conversations grounded in personas and temporal event graphs and evaluated memory through question answering, summarization, and dialogue generation. LongMemEval Wu et al. (2025) further targets long-term memory in chat assistants, covering information extraction, multi-session reasoning, temporal reasoning, knowledge updates, and abstention. Recent benchmarks broaden memory evaluation beyond factual recall. EvolMem Shen et al. (2026) studies declarative and non-declarative memory, while EvoMemBench Wang et al. (2026) distinguishes in-episode from cross-episode memory and knowledge from execution experience. LoCoMo-Plus Li et al. (2026) focuses on beyond-factual memory, testing whether agents can preserve latent constraints such as user states, goals, and values under semantically different future cues.

2.2 Context Compression Methods

These benchmarks naturally raise a fundamental question: given a limited memory budget, which parts of a long interaction should be retained explicitly, and which can be abstracted? Existing LLM-based memory systems have explored several approaches to address this challenge. Global summarization compresses an entire interaction segment into a shorter abstract representation, but may omit fine-grained details. Local turn-level compression instead summarizes smaller units independently, preserving local information at the cost of potentially weakening cross-turn relationships. Token-level compression retains selected portions of the original context; LLMLingua (Jiang et al., 2023) introduces budget-controlled coarse-to-fine prompt compression, while LLMLingua-2 (Pan et al., 2024) formulates prompt compression as a token-level classification task, enabling efficient and faithful information compression. Retrieval-based methods reduce effective context by identifying relevant chunks or spans at query time. SeCom (Pan et al., 2025) combines conversation segmentation, compression-based denoising, and memory retrieval, demonstrating that memory construction granularity can significantly affect downstream retrieval quality. Structured-memory approaches (Xu et al., 2026) represent conversations as atomic facts, preferences, events, or relations. Such representations can improve factual accessibility but may discard contextual dependencies between extracted items. Hierarchical compression methods progressively aggregate local representations into higher-level summaries, enabling long histories to be represented at multiple levels of granularity (Sun et al., 2026). Finally, learned latent-memory approaches, such as R^3Mem (Wang et al., 2025), encode long histories into compact virtual or continuous representations, achieving high compression at the cost of reduced interpretability and potentially stronger model dependence. Unlike query-conditioned selection, latent memory, or purely fact-based compression, CORAL, instead,

performs query-agnostic textual compression while retaining human-readable contextual representations.

3 METHODOLOGY

3.1 OVERVIEW

CORAL treats long-term conversational memory as a selective memory construction problem. Rather than retaining the complete dialogue history, the framework constructs a compact memory representation that preserves information likely to remain useful for future retrieval and reasoning.

As illustrated in Figure 1, CORAL consists of three functional components: a *Preservation Agent*, a *Compression Agent*, and a *Conversation Agent*. The Preservation Agent identifies salient information that should be protected from lossy compression, the Compression Agent constructs a compact memory representation under a predefined token budget, and the Conversation Agent retrieves relevant compressed memories to support downstream reasoning.

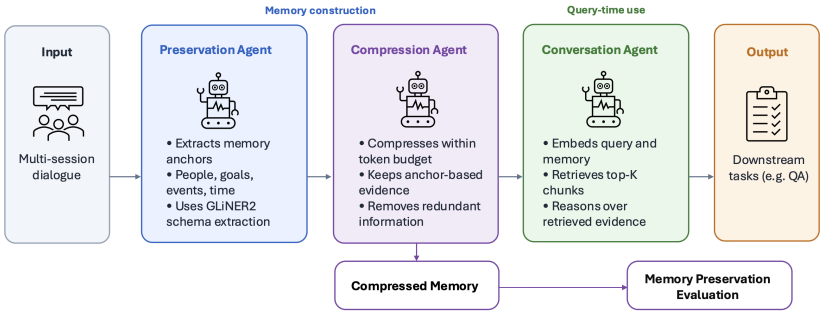


Figure 1: Overview of CORAL. Three agents support memory preservation, compression, and retrieval-augmented reasoning.

3.2 MEMORY CONSTRUCTION

Given a multi-session conversation, CORAL first partitions the conversation into approximately 300-token chunks while preserving the timestamp associated with each session. Timestamps provide essential temporal context for information involving explicit dates, relative temporal expressions, event ordering, and temporal relations.

For each conversational chunk x_i , the Preservation Agent identifies information that may be important for future interactions, while the Compression Agent constructs a compact representation conditioned on these preservation signals. The resulting memory units are subsequently indexed for retrieval by the Conversation Agent.

Preservation Agent. Lossy compression can discard information whose future relevance is unknown at the time of compression. CORAL therefore introduces a Preservation Agent to identify potentially important information before compression. We implement the Preservation Agent using GLiNER2 (Zaratiana et al., 2025), a schema-driven information extraction model that supports extraction according to user-defined semantic categories. Unlike conventional named entity recognition systems with predefined entity types, GLiNER2 allows the extraction schema to be adapted to the information needs of conversational memory. For CORAL, the extraction schema covers people and identities, preferences, goals, plans, constraints, locations, events, relationships, dates and temporal expressions, temporal relations, and updates to previously stated information. Given a conversational chunk x_i , the Preservation Agent produces a set of memory anchors:

$$A_i = P_\theta(x_i), \tag{1}$$

where A_i denotes the extracted memory anchors and P_θ represents the Preservation Agent. These anchors serve as preservation signals rather than the final memory representation.

Compression Agent. The Compression Agent receives the original conversational chunk x_i , its session timestamp t_i, the extracted memory anchors A_i, and a target token budget B. It generates a compact memory representation:

$$m_i = C_\phi(x_i, A_i, t_i; B), \tag{2}$$

where C_ϕ denotes the Compression Agent and m_i is the resulting memory unit. The compressor is instructed to prioritize information associated with the memory anchors while preserving temporal information, including session dates, explicit dates and durations, relative temporal expressions, event ordering, and temporal relations. Redundant statements, greetings, filler, and other low-value conversational content are removed to satisfy the target token budget. The compressor is additionally constrained not to introduce information that is absent from the original conversation. This anchor-guided design separates the identification of potentially important information from the compression process itself. Rather than requiring a single generation step to simultaneously determine what should be retained and how it should be compressed, CORAL first identifies preservation targets and then performs lossy compression conditioned on these targets.

Conversation Agent. The Conversation Agent is responsible for accessing the constructed memory and using relevant information for downstream tasks. CORAL adopts a retrieval-augmented generation (RAG) architecture, in which relevant memory units are retrieved from an external memory store before response generation. After compression, each memory unit m_i is encoded using a sentence embedding model and stored in a vector index. Given a query q, the same encoder produces a query representation $e(q)$, and cosine similarity is used to rank the stored memory units:

$$s(q, m_i) = \cos\left(e(q), e(m_i)\right). \tag{3}$$

The top-K memory units are then retrieved and provided to the Conversation Agent together with their associated session timestamps:

$$R(q) = \{m_{(1)}, \ldots, m_{(K)}\}, \quad where\ \ s(q, m_{(1)}) \geq \cdots \geq s(q, m_{(|M|)}). \tag{4}$$

where M denotes the set of stored memory units. The Conversation Agent generates a response based on the retrieved memory rather than the complete conversation history. Formally, given a query q and the retrieved memory set $R(q)$, the Conversation Agent produces a response as

$$y = G_\psi\left(q, R(q)\right), \tag{5}$$

where G_ψ denotes the Conversation Agent and y denotes the generated response.

Overall, CORAL separates long-term memory processing into three complementary stages: *identifying* information that should be protected, *constructing* compact representations under a memory budget, and *retrieving* relevant representations when they are needed. This separation allows CORAL to explicitly control the trade-off between information preservation and memory efficiency.

4 Experiments

4.1 Dataset

We use LoCoMo (Maharana et al., 2024), a benchmark for very long-term conversational memory, to evaluate CORAL. Each conversation of LoCoMo contains approximately 600 turns and 16K tokens on average, spanning up to 32 sessions.

LoCoMo constructs long-term conversations through a multi-stage generative pipeline. Two virtual agents, $\mathcal{L}_1$ and $\mathcal{L}_2$, are first assigned a seed persona statement p_c from the MSC dataset (Xu et al., 2022). The underlying language model, $\mathcal{M}$ (`gpt-3.5-turbo`), then expands the seed statement into a full persona profile p_i, typically covering objectives, past experiences, daily habits, interpersonal relationships, and basic demographic attributes such as name, age, etc. (Gao et al., 2023).

For each agent $\mathcal{L}_i$, a temporal event graph $\mathcal{G}_i = (V_i, E_i)$ is then generated accordingly by $\mathcal{M}$ (`text-davinci-003`) from its persona statement, where each node $e_j \in V_i$ denotes a life event associated with a date of occurrence t_j, and each edge $(e_j, e_h) \in E_i$ represents a causal relation between events. Event generation proceeds iteratively: an initial batch of $k = 3$ events is generated

and then used as context to produce the next batch of k events, with this process repeated until the event sequence is complete.

Each agent $\mathcal{L}_i$ includes modules from a generative-agent architecture (Park et al., 2023). The agent has two functions: (1) reflect & respond, (2) image sharing & image reaction. Based on p_i and $\mathcal{G}_i$, the agents conduct multi-session conversations using the architecture with short-term and long-term memory. At session $k+1$, the response is generated based on the latest summary w_k of session k, reflections based on the retrieved relevant observations o, the ongoing session's conversation history h_{k+1} and persona statement p_i. The architecture also supports image sharing and reaction. Finally, human annotators edit the generated conversations to remove inconsistencies, align dialogues with event graphs, and filter irrelevant images.

In this paper, we primarily evaluate CORAL on question answering and event summarization tasks, which allows us to examine how well task-relevant information remains retrievable after memory compression. Our objective is to examine the balance between *memory efficiency* and *downstream performance*. Accordingly, we evaluate CORAL not only in terms of task performance, but also with respect to the number of tokens required to achieve that performance.

We organize the evaluation into three experiments. Experiment 1 identifies effective settings for the compression fraction and compressor model size. Experiment 2 compares different compression methods on the same downstream tasks. Experiment 3 examines performance across model families.

4.2 Experimental Setup

All experiments were conducted using a single NVIDIA A100 GPU with 80 GB of GPU memory. Model inference, memory compression, retrieval, and evaluation were performed within the same environment. Due to computational constraints, we evaluate the first 10 questions from the first LoCoMo conversation. While this represents a limited evaluation setting, the results still revealed clear comparative patterns across compression fractions, model sizes, compression methods, and model families.

Our evaluation separates memory construction from downstream reasoning. The evaluated conversation is first converted into compressed memory, organized into fixed-length retrieval units with session timestamps retained, and indexed using the Sentence Transformer model `all-MiniLM-L6-v2`. For each question, the top-K ($K = 10$) units were retrieved and provided to a fixed QA model. For event-summary evaluation, compressed memories were directly compared with LoCoMo's reference session summaries using ROUGE recall, without retrieval. Thus, across compression methods, only memory construction changed, while retrieval, question set, and downstream tasks were held constant. Following Idreos et al. (2010), we measure compression efficiency using the actual compression fraction

$$\mathrm{CF} = \frac{n_{\mathrm{comp}}}{n_{\mathrm{total}}}, \tag{6}$$

where n_{total} and n_{comp} are the token counts before and after compression; lower values indicate stronger compression.

For QA, human-evaluated accuracy is used as the primary metric because semantically correct generated answers may differ lexically from the reference answer. We also report token-level F1 as an auxiliary lexical-overlap metric, which may underestimate semantic correctness when equivalent answers use different wording. Human judgments therefore provide a more direct assessment of answer correctness, including cases that require semantic or temporal reasoning. Statistical significance is assessed using paired non-parametric tests, with full results reported in Appendix A.

To measure general memory preservation independently of QA, we compare the compressed memory with LoCoMo's reference session event summaries using ROUGE-1, ROUGE-2, and ROUGE-L recall. These scores measure how much of the reference conversational information remains represented after compression.

4.3 Experiment 1: Compression Settings

We examine how CORAL's internal configuration affects memory compression and task performance by varying two factors: the compression fraction and the compressor model size, while keeping the downstream pipeline fixed.

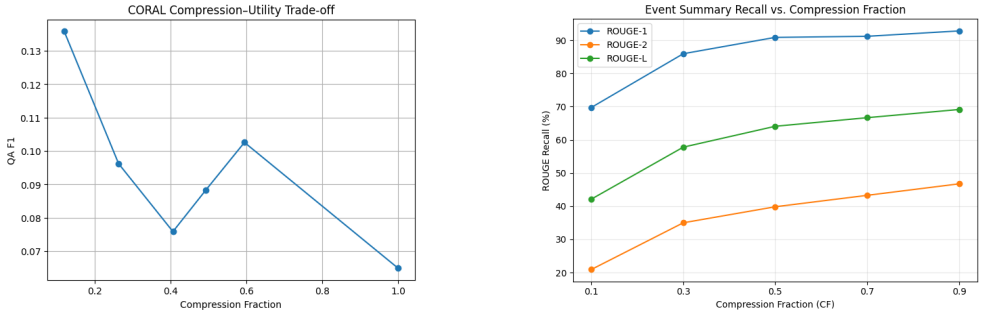
Compression Fraction. We first examine how the compression fraction affects downstream QA performance and general memory preservation. Using Qwen3-8B as the compressor, we evaluate target compression fractions of $\{0.1, 0.3, 0.5, 0.7, 0.9\}$, with uncompressed memory as a reference.

Table 1: Experiment 1A results across compression fractions.

Target CF	Actual CF	QA F1	Human Acc.	ROUGE-1 rec.	ROUGE-2 rec.	ROUGE-L rec.
0.1	0.119	**0.136**	**60%**	69.75	20.92	42.12
0.3	0.263	0.096	30%	85.95	35.00	57.79
0.5	0.406	0.076	30%	90.87	39.83	64.08
0.7	0.493	0.088	40%	91.19	43.27	66.67
0.9	0.595	0.103	**60%**	**92.82**	**46.74**	**69.14**
Raw	1.000	0.065	50%	70.07	28.24	54.50

As shown in Table 1 and Figure 2a, QA performance is non-monotonic with respect to compression fraction. Human accuracy ranges from 30% to 60% across compressed settings, compared with 50% for uncompressed memory, while F1 is likewise non-monotonic and reaches its highest value at $\mathrm{CF} = 0.1$.

In contrast, Figure 2b shows that ROUGE recall generally increases as more memory is retained. From $\mathrm{CF} = 0.1$ to $\mathrm{CF} = 0.9$, ROUGE-1 increases from 69.75% to 92.82%, ROUGE-2 from 20.92% to 46.74%, and ROUGE-L from 42.12% to 69.14%. This suggests that retaining more conversational content improves general memory preservation, but does not necessarily improve downstream QA.



(a) QA performance across compression fractions.

(b) Event summary ROUGE recall.

Figure 2: Effect of compression fraction on downstream QA tasks and general memory preservation.

Overall, $\mathrm{CF} = 0.1$ provides the strongest compression while achieving 60% human QA accuracy, compared with 50% for uncompressed memory, and still retains substantial general conversational information. We therefore use $\mathrm{CF} = 0.1$ in subsequent experiments.

Compressor Model Size. We next examine the effect of compressor model size while fixing the target compression fraction at $\mathrm{CF} = 0.1$. CORAL is instantiated with Qwen3-4B, Qwen3-8B, Qwen3-14B, and Qwen3-32B, while the retrieval pipeline and downstream QA model are held fixed.

As shown in Table 2, increasing compressor size does not produce a monotonic improvement in downstream QA. Human accuracy rises from 40% with Qwen3-4B to 60% with Qwen3-8B and remains unchanged for the 14B and 32B models, while Qwen3-8B achieves the highest F1 score of 0.119.

Table 2: Effect of compressor model size at target CF $= 0.1$.

Compressor	Actual CF	QA F1	Human Acc.	ROUGE-1 rec.	ROUGE-2 rec.	ROUGE-L rec.	Comp. Time (s)
Qwen3-4B	0.117	0.103	40%	62.55	20.40	43.60	169.9
Qwen3-8B	0.119	**0.119**	**60%**	69.75	20.92	42.12	173.0
Qwen3-14B	0.120	0.088	**60%**	**75.25**	**27.44**	**48.07**	191.6
Qwen3-32B	0.120	0.105	**60%**	68.90	23.99	46.07	292.8

Event-summary preservation shows a different pattern. Qwen3-14B achieves the highest ROUGE-1, ROUGE-2, and ROUGE-L recall of 75.25%, 27.44%, and 48.07%, respectively, whereas Qwen3-8B achieves lower ROUGE recall despite stronger downstream QA performance. This suggests that a larger compressor may preserve more general conversational information without necessarily producing a memory representation that is more useful for downstream retrieval and reasoning.

The actual compression fractions remain nearly identical across models, ranging from 0.117 to 0.120, making the comparison well controlled with respect to memory budget. Compression time, however, increases with model size, reaching 292.8 seconds for Qwen3-32B compared with 173.0 seconds for Qwen3-8B. Overall, Qwen3-8B provides the best balance between downstream QA performance and computational cost, and is therefore used as the compressor in subsequent experiments.

4.4 Experiment 2: Comparison with Compression Baselines

We next compare CORAL against two compression baselines, Chunk Summary and LLMLingua-2, with uncompressed memory serving as the uncompressed reference. All compressed methods use the target compression fraction selected in Experiment 1, while the retrieval procedure, question set, and downstream QA model are held fixed.

Table 3: Comparison of compression methods on QA and event-summary preservation.

Method	QA F1	Human Acc.	ROUGE-1 rec.	ROUGE-2 rec.	ROUGE-L rec.
CORAL	**0.105**	**50%**	69.75	20.92	42.12
Chunk Summary	0.089	10%	**76.02**	**35.77**	52.17
LLMLingua-2	0.079	30%	31.75	4.24	21.64
Raw	0.078	**50%**	70.07	28.24	**54.50**

As shown in Table 3, CORAL achieves the highest QA F1 of 0.105 and 50% human accuracy, matching the uncompressed condition in human-evaluated correctness while outperforming it on token-level F1. Chunk Summary achieves higher ROUGE recall than CORAL, with ROUGE-1, ROUGE-2, and ROUGE-L scores of 76.02%, 35.77%, and 52.17%, respectively, yet its human QA accuracy drops to only 10%. LLMLingua-2 reaches 30% human accuracy but retains substantially less event information, with ROUGE-1, ROUGE-2, and ROUGE-L recall of 31.75%, 4.24%, and 21.64%.

These results highlight a distinction between general information preservation and downstream memory utility. Higher event-summary ROUGE does not necessarily translate into better question answering. Although Chunk Summary preserves greater lexical overlap with the reference summaries, CORAL produces a compressed memory representation that is more effective for downstream retrieval and reasoning. Overall, CORAL achieves the strongest downstream QA performance among the compressed methods while preserving substantial event information under the same target memory budget.

4.5 Experiment 3: Generalization Across Model Families

Finally, we examine how CORAL behaves across different language-model families. Using the same target compression fraction of CF $= 0.1$ and the same downstream pipeline, we instantiate the compressor with Qwen3-8B, DeepSeek-R1-Distill-8B, Llama-3.1-8B-Instruct, and GPT-5.6 Sol. The downstream QA model remains fixed across conditions so that differences primarily reflect the compressor model.

Table 4: CORAL performance across compressor model families at target CF $= 0.1$.

Compressor	Actual CF	QA F1	Human Acc.	ROUGE-1 rec.	ROUGE-2 rec.	ROUGE-L rec.	Comp. Time (s)
Qwen3-8B	0.119	**0.125**	**50%**	69.75	20.92	42.12	166.9
DeepSeek-R1-Distill-8B	0.124	0.061	0%	31.10	1.95	27.08	129.8
Llama-3.1-8B	0.128	0.086	30%	**76.16**	**27.34**	**49.70**	**128.3**
GPT-5.6 Sol	0.028	0.056	30%	38.38	4.64	22.04	264.2

As shown in Table 4, CORAL's downstream performance varies substantially across compressor families. Qwen3-8B achieves the highest QA F1 of 0.125 and the highest human accuracy of 50%. Llama-3.1-8B obtains lower QA performance, with an F1 of 0.086 and 30% human accuracy, but achieves the strongest event-summary preservation, with ROUGE-1, ROUGE-2, and ROUGE-L recall of 76.16%, 27.34%, and 49.70%, respectively. DeepSeek-R1-Distill-8B performs substantially worse on both downstream QA and event preservation, reaching 0% human accuracy and markedly lower ROUGE recall.

The OpenAI condition behaves differently from the local 8B models. Although all compressors are given the same target fraction of 0.1, GPT-5.6 Sol reaches an actual compression fraction of only 0.028, compared with 0.119–0.128 for the other models. Its stronger-than-target compression is accompanied by lower QA F1 and event-summary recall, making direct comparison less controlled. Overall, these results show that CORAL can be instantiated across multiple model families, but the quality of the resulting memory representation remains strongly model-dependent. Higher lexical preservation or stronger compression does not necessarily translate into better downstream QA performance.

5 Discussion

Our results suggest that effective long-term conversational memory compression is not determined by compression ratio, model size, or lexical preservation alone. Across all experiments, the most useful compressed memories are those that preserve information in a form that remains accessible to downstream retrieval and reasoning.

Aggressive Compression Can Preserve Downstream Performance. Experiment 1A shows that QA performance is not monotonic in the amount of memory retained. Although ROUGE recall generally improves as the compression fraction increases, the strongest QA result is obtained at the most aggressive tested setting, with an actual compression fraction of 0.119. At this setting, CORAL reaches 60% human accuracy compared with 50% for uncompressed memory. This indicates that retaining more text is not always beneficial for downstream reasoning: removing redundant or distracting content can yield a smaller memory that remains highly useful for question answering.

More Capacity Does Not Guarantee Better Memory. Experiment 1B reveals a similar disconnect between compressor scale and downstream utility. Qwen3-8B achieves the highest QA F1, while larger Qwen3-14B and Qwen3-32B models do not improve QA despite higher computational cost. Qwen3-14B does achieve the strongest event-summary ROUGE recall, suggesting that larger models may preserve more general conversational content without producing a representation that is better aligned with the downstream task. Compressor selection should therefore consider both memory quality and computational efficiency rather than model size alone.

Information Preservation and Task Performance Are Distinct Objectives. Experiment 2 further demonstrates that lexical retention and downstream utility are not necessarily aligned. Although Chunk Summary achieves higher ROUGE recall than CORAL, its human QA accuracy decreases to 10%, compared with 50% for CORAL, which matches the uncompressed baseline. LLMLingua-2 exhibits both lower event-summary recall and lower QA accuracy. These results underscore the need to evaluate compressed memory along complementary dimensions: ROUGE recall measures broad information retention, whereas downstream QA assesses whether the retained information remains usable for retrieval and reasoning.

Compression Behavior Depends on the Model Family. Experiment 3 shows that CORAL can be instantiated with different model families, but the resulting memory quality varies substantially. Qwen3-8B achieves the strongest downstream QA, while Llama-3.1-8B achieves the highest ROUGE recall. DeepSeek-R1-Distill-8B performs worse on both measures. The OpenAI condition reaches a much lower actual compression fraction than the local models despite the same target fraction, so its results are not directly comparable under an identical memory budget. Taken together, these findings show that the compression framework is transferable across model families, but its performance remains sensitive to the compressor used.

Overall, the experiments point to a broader principle: long-term memory compression should be treated as a task-aware memory construction problem rather than as text shortening alone. A good memory is not necessarily the one that retains the most words, but the one that preserves the information most useful for future retrieval and reasoning under a constrained memory budget.

6 LIMITATIONS

Our work has several limitations. First, all experiments are conducted on LoCoMo, so the results may depend on its data characteristics and may not generalize to other conversational settings. Second, evaluation focuses mainly on question answering and ROUGE recall against session summaries, which do not fully capture broader aspects of long-term memory such as personalization, consistency, planning, or open-ended dialogue quality. Third, we use a fixed retrieval and QA pipeline, and performance may vary with different downstream components. In addition, target compression fractions are not matched perfectly across compressor families, and the human evaluation is based on a relatively small sample. Finally, our offline evaluation does not model dynamic memory processes such as updating, forgetting, conflict resolution, or consolidation.

Memory compression can reduce storage and inference costs, but lossy selection may discard information that becomes relevant later or distort retained content. Persistent conversational memory also raises privacy concerns. Practical systems should therefore support memory correction, deletion, and appropriate data protection mechanisms.

7 FUTURE WORK

Future work can extend CORAL in several directions. First, the proposed selective compression framework could be applied to a broader range of modalities and tasks, including long documents, images, audio, video, multi-modal inputs, and non-conversational reasoning settings, to test whether the compression-utility trade-off generalizes beyond dialogue. Second, compression and retrieval could be optimized jointly rather than treated as separate stages, allowing the memory representation to adapt directly to retrieval behavior. Finally, CORAL could be evaluated in online settings where memory grows continuously, previously stored information must be revised or consolidated, and compression decisions must be made incrementally over long interaction histories.

8 CONCLUSION

We introduced CORAL, a multi-agent system for constructing compact long-term conversational memory under an explicit token budget. Evaluated on LoCoMo, CORAL substantially reduces memory size while maintaining competitive downstream QA performance and preserving a meaningful portion of the reference conversational information. Across compression fractions, compressor sizes, compression baselines, and model families, the results consistently show that neither retaining more text nor using a larger compressor guarantees better downstream utility. In particular, CORAL can match or exceed uncompressed memory on human-evaluated QA at a small fraction of the original memory size, while outperforming alternative compression baselines on downstream task performance. These findings support treating conversational memory compression as selective memory construction and underscore the importance of evaluating compressed memory jointly in terms of memory efficiency, information preservation, and downstream task utility.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on research methodology and experimental design, implement methods, and assist with the interpretation of results. We have not used generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for or assist in the writing of proofs, propose or refine hypotheses, assist with translation, clean or reformat datasets, or support qualitative or thematic data analysis, and the remaining required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for brainstorming, creating or editing software code, drafting parts of the research paper, sourcing and searching for information, editing the paper to improve readability, and suggesting manuscript structure. We have reviewed all AI-assisted work. LLM-generated code was manually inspected and tested for correctness by the author, and factual claims and references were independently verified. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the CORAL framework, compression procedures, retrieval and question-answering pipeline, evaluation metrics, and experimental settings in the main paper and appendix. The experiments use the publicly available LoCoMo dataset and publicly documented language models. Model configurations, prompts, compression ratios, retrieval settings, and evaluation procedures are reported to facilitate reproduction. Additional implementation details and extended results are provided in the appendix. The implementation code and evaluation scripts are also released anonymously.

REFERENCES

Mihail Eric and Christopher D. Manning. Key-value retrieval networks for task-oriented dialogue, 2017. URL https://arxiv.org/abs/1705.05414.

Silin Gao, Beatriz Borges, Soyoung Oh, Deniz Bayazit, Saya Kanno, Hiromi Wakaki, Yuki Mitsufuji, and Antoine Bosselut. PeaCoK: Persona commonsense knowledge for consistent and engaging narratives. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6569–6591, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.362. URL https://aclanthology.org/2023.acl-long.362/.

Stratos Idreos, Raghav Kaushik, Vivek Narasayya, and Ravishankar Ramamurthy. Estimating the compression fraction of an index using sampling. In *2010 IEEE 26th International Conference on Data Engineering (ICDE 2010)*, pp. 441–444. IEEE, 2010.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LLMLingua: Compressing prompts for accelerated inference of large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13358–13376, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.825. URL https://aclanthology.org/2023.emnlp-main.825/.

Hung Le, Doyen Sahoo, Chenghao Liu, Nancy Chen, and Steven CH Hoi. Uniconv: A unified conversational neural architecture for multi-domain task-oriented dialogues. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1860–1877, 2020.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Yifei Li, Weidong Guo, Lingling Zhang, Rongman Xu, Muye Huang, Hui Liu, Lijiao Xu, Yu Xu, and Jun Liu. Locomo-plus: Beyond-factual cognitive memory evaluation framework for llm agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25085–25100. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1150.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of llm agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13851–13870. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.747.

Shikib Mehri, Tejas Srinivasan, and Maxine Eskenazi. Structured fusion networks for dialog, 2019. URL https://arxiv.org/abs/1907.10016.

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

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Xufang Luo, Hao Cheng, Dongsheng Li, Yuqing Yang, Chin-Yew Lin, H Vicky Zhao, Lili Qiu, et al. Secom: On memory construction and retrieval for personalized conversational agents. In *International Conference on Learning Representations*, volume 2025, pp. 91851–91885, 2025.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior, 2023. URL https://arxiv.org/abs/2304.03442.

Libo Qin, Wenbo Pan, Qiguang Chen, Lizi Liao, Zhou Yu, Yue Zhang, Wanxiang Che, and Min Li. End-to-end task-oriented dialogue: A survey of tasks, methods, and future directions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5925–5941, 2023.

Ye Shen, Dun Pei, Yiqiu Guo, Junying Wang, Yijin Guo, Zicheng Zhang, Qi Jia, Jun Zhou, and Guangtao Zhai. Evolmem: A cognitive-driven benchmark for multi-session dialogue memory. *arXiv preprint arXiv:2601.03543*, 2026.

Haoran Sun, Shaoning Zeng, and Bob Zhang. H-MEM: Hierarchical memory for high-efficiency long-term reasoning in LLM agents. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 341–350, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.15. URL https://aclanthology.org/2026.eacl-long.15/.

Xiaoqiang Wang, Suyuchen Wang, Yun Zhu, and Bang Liu. R^3Mem: Bridging memory retention and retrieval via reversible compression. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 4541–4557, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.235. URL https://aclanthology.org/2025.findings-acl.235/.

Yuyao Wang, Zhongjian Zhang, Mo Chi, Kaichi Yu, Yuhan Li, Miao Peng, Bing Tong, Chen Zhang, Yan Zhou, and Jia Li. Evomembench: Benchmarking agent memory from a self-evolving perspective. *arXiv preprint arXiv:2605.18421*, 2026.

Chien-Sheng Wu, Richard Socher, and Caiming Xiong. Global-to-local memory pointer networks for task-oriented dialogue, 2019. URL https://arxiv.org/abs/1901.04713.

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

Buqiang Xu, Yijun Chen, Jizhan Fang, Ruobin Zhong, Yunzhi Yao, Yuqi Zhu, Lun Du, and Shumin Deng. Structmem: Structured memory for long-horizon behavior in llms, 2026. URL https://arxiv.org/abs/2604.21748.

Jing Xu, Arthur Szlam, and Jason Weston. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2022.

Zihao Yi, Jiarui Ouyang, Zhe Xu, Yuwen Liu, Tianhao Liao, Haohao Luo, and Ying Shen. A survey on recent advances in llm-based multi-turn dialogue systems, 2025. URL https://arxiv.org/abs/2402.18013.

Urchade Zaratiana, Gil Pasternak, Oliver Boyd, George Hurn-Maloney, and Ash Lewis. GLiNER2: Schema-driven multi-task learning for structured information extraction. In Ivan Habernal, Peter Schulam, and Jörg Tiedemann (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 130–140, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-334-0. doi: 10.18653/v1/2025.emnlp-demos.10. URL https://aclanthology.org/2025.emnlp-demos.10/.

A Statistical Significance Analysis

To assess whether the observed differences in downstream QA performance are statistically meaningful, we conduct paired statistical tests over the same 10 QA questions used across all conditions. Because each question is evaluated under every experimental condition, the resulting observations are paired rather than independent.

For token-level QA F1, we use the Friedman test, a non-parametric test for comparing multiple paired conditions. For human accuracy, where each answer is labeled as correct or incorrect, we use Cochran's Q test for paired binary outcomes. Statistical significance is assessed at $p < 0.05$.

Table 5: Statistical significance tests for downstream QA performance.

Experiment	QA F1	Human Acc.
Exp. 1A: Compression Fraction	$p = 0.245$	$p = 0.300$
Exp. 1B: Compressor Size	$p = 0.649$	$p = 0.651$
Exp. 2: Compression Methods	$p = 0.742$	$\mathbf{p = 0.041}$
Exp. 3: Model Families	$\mathbf{p = 0.029}$	$\mathbf{p = 0.017}$

As shown in Table 5, Experiments 1A and 1B do not show statistically significant overall differences in either QA F1 or human accuracy. This indicates that the present evaluation does not provide sufficient evidence that compression fraction or compressor size alone systematically affects downstream QA performance.

Experiment 2 shows a significant overall difference in human accuracy across compression methods ($p = 0.041$), although the corresponding QA F1 differences are not significant. Experiment 3 shows significant overall differences across model families for both QA F1 ($p = 0.029$) and human accuracy ($p = 0.017$), providing stronger evidence that compressor family influences downstream memory utility.

Because the evaluation contains only 10 QA questions, the statistical power of these tests is limited. Non-significant results should therefore not be interpreted as evidence that no differences exist between conditions. Instead, the statistical analysis should be viewed together with the observed performance trends reported in the main experiments.