JOURNAL OF SOCIAL SCIENCES, HUMANITIES AND INNOVATION

Vol 2 Issue 2 (2026)

http://jsshi.org/

Trajectory Memory Compression for Long-Horizon Dialogue Agents with Provenance Preservation

Philip Ho[1], Joanna Ma[2]

[1] Department of Computer Science, Faculty of Engineering, University of Hong Kong, Hong Kong, Hong Kong SAR, China

[2] School of Computer Science, Peking University, Beijing, China

Keywords:
Trajectory Memory, Memory Compression, Long-Horizon Dialogue, Provenance Preservation, Autonomous Agents

Abstract:
The rapid advancement of artificial intelligence has facilitated the development of dialogue agents capable of highly sophisticated interactions. However, sustaining coherence and factual consistency over extended interactions remains a formidable challenge due to the inherent constraints of context windows in large language models. This paper investigates the critical problem of trajectory memory compression in long-horizon dialogue agents, introducing a novel paradigm that integrates provenance preservation into the compression pipeline. Traditional memory compression techniques often discard granular contextual details, leading to hallucinations and a loss of traceability when the agent attempts to recall historical information. By conceptualizing dialogue history as a continuous trajectory of events, we propose a methodology that selectively compresses semantic content while meticulously maintaining cryptographic and structural pointers to the original episodic occurrences. This dual-track approach ensures that the agent can retrieve highly compressed semantic representations without sacrificing the ability to ground its responses in specific past utterances. Through rigorous theoretical analysis and extensive empirical evaluation across multiple long-context dialogue datasets, we demonstrate that our provenance-aware compression mechanism significantly enhances both the computational efficiency of memory retrieval and the factual reliability of the generated responses. The findings offer a scalable framework for deploying autonomous agents in domains requiring prolonged interaction and strict auditing, such as healthcare and legal counseling.

1.INTRODUCTION

1.1 Background and Motivation

The evolution of conversational artificial intelligence has transitioned from stateless, rule-based systems to highly dynamic, stateful agents powered by massive language models. In contemporary applications, dialogue agents are increasingly deployed in scenarios that demand continuous interaction over days, weeks, or even months. These long-horizon settings, such as personalized tutoring, chronic disease management, and persistent virtual companionship, require the agent to

build and maintain a comprehensive understanding of the user over time. Central to this capability is the agent's memory system, which must record, organize, and retrieve historical interactions to maintain conversational coherence and personalize responses. Early approaches to dialogue memory relied on simple concatenation of past turns, a method that scales poorly as the interaction history grows. As the sequence of dialogue acts extends, the sheer volume of tokens quickly exceeds the maximum context window of the underlying language model. Although recent advancements have significantly expanded these context windows, processing tens or hundreds of thousands of tokens for every dialogue turn imposes prohibitive computational costs and introduces significant latency, rendering naive concatenation untenable for real-time applications. Consequently, researchers have turned to external memory modules and retrieval-augmented generation techniques to manage long-term context. These methods typically involve encoding past utterances into dense vector representations and storing them in an external database, allowing the agent to retrieve only the most relevant historical segments based on the current dialogue context. However, while effective in reducing computational overhead, standard retrieval mechanisms often struggle with complex temporal reasoning and the synthesis of information spread across multiple distinct past episodes. Furthermore, as the memory database grows, the density of similar but temporally distinct vectors increases, leading to interference and degraded retrieval accuracy. To address these limitations, memory compression has emerged as a crucial area of research. Memory compression techniques aim to distill the sprawling dialogue history into compact, high-level summaries or abstract representations, thereby maximizing the information density within the model's context window. Related work has addressed towards proactive asking with imperfect information in medical multi-turn dialogues [1].

1.2 Challenges in Long-Horizon Dialogue and the Role of Provenance

Despite the promise of memory compression, existing techniques introduce a critical vulnerability: the loss of provenance. When a dialogue agent compresses a lengthy exchange into a concise semantic summary, the explicit link to the original raw utterances is frequently severed. In high-stakes domains such as legal or medical consultations, it is not sufficient for an agent to merely recall a compressed fact; it must also be able to justify its recollection by citing the specific historical context or utterance from which the fact was derived. This ability to trace a piece of information back to its origin is referred to as provenance. Without provenance, compressed memory becomes a black box, increasing the risk of hallucination and severely undermining user trust. If an agent states that a user previously expressed a specific preference or reported a particular symptom, the inability to pinpoint exactly when and how that information was conveyed makes it impossible to verify the claim or understand the nuanced context of the original disclosure. Therefore, the central challenge addressed in this paper is the reconciliation of two seemingly conflicting objectives: the need to aggressively compress dialogue history to maintain computational tractability over long horizons, and the absolute necessity of preserving provenance to ensure accountability and factual grounding. We conceptualize the dialogue history not merely as a collection of independent utterances, but as a continuous trajectory of interactions, where each point in the trajectory holds specific episodic significance. Our proposed framework introduces a mechanism for trajectory memory compression that selectively abstracts semantic meaning while explicitly encoding cryptographic and structural pointers back to the raw episodic data. This allows the agent to reason over highly compressed representations and, when necessary, traverse the provenance pointers to retrieve the exact historical context, thereby combining the efficiency of compression with the reliability of exact recall. For a broader comparison, related research addresses singing timbre popularity assessment based on multimodal large foundation model [2].

2. Literature Review

2.1 Long-Context Language Models and Memory Systems

The processing of long sequences has been a perennial challenge in the field of natural language processing. The fundamental architecture of the Transformer, which underpins modern language models, relies on a self-attention mechanism whose computational complexity scales quadratically with the length of the input sequence. This quadratic scaling imposes a hard limit on the amount of historical context an agent can consider simultaneously. Early attempts to mitigate this limitation focused on architectural modifications, such as sparse attention patterns, sliding windows, and hierarchical processing structures, which sought to approximate full self-attention with lower computational complexity. While these methods allowed for somewhat longer context windows, they often resulted in degraded performance on tasks requiring complex, long-range dependency modeling. Consequently, the research community shifted its focus toward augmenting models with external memory systems. In these paradigms, the language model is decoupled from the memory storage, functioning instead as a central processing unit that reads from and writes to an external datastore. Retrieval-augmented generation represents the most prominent instantiation of this approach. By employing dense retrieval techniques, systems can dynamically select the most relevant historical utterances and inject them into the current context window, effectively providing the model with an unbounded, albeit selectively accessed, memory. However, retrieval-augmented systems face significant challenges in scenarios where the required information is not neatly contained within a single past utterance but is instead distributed across a lengthy dialogue trajectory. In such cases, retrieving individual, isolated snippets fails to capture the overarching narrative or the evolution of the user's state over time. To capture these macro-level dynamics, researchers have explored various forms of memory architectures, drawing inspiration from cognitive psychology. The distinction between episodic memory, which stores specific, time-stamped events, and semantic memory, which stores generalized facts and concepts abstracted from multiple episodes, has been particularly influential. Dialogue agents equipped with dual memory systems attempt to leverage both the granular detail of episodic storage and the concise summarization of semantic storage. Yet, the mechanism by which episodic memories are continuously consolidated into semantic representations without losing critical contextual nuances remains an open research question. Our work builds upon these foundational concepts but specifically addresses the gap in maintaining verifiable links between the consolidated semantic knowledge and its episodic origins. Yao and colleagues examine pretrained discrete masked diffusion language modeling with hybrid attention [3]. For a broader comparison, related research addresses comprehensive survey of Cross-Modal retrieval: breaking through modal barriers [4].

2.2 Memory Compression Techniques and Explainability

Memory compression encompasses a diverse array of techniques designed to reduce the token footprint of historical context. One common approach is recursive summarization, where a secondary language model periodically processes segments of the dialogue history, generating concise summaries that replace the raw transcripts in the agent's active memory. While recursive summarization effectively curtails the exponential growth of context, it inherently involves a lossy transformation. Nuances, tonal shifts, and seemingly trivial details that might become relevant in the future are often discarded by the summarization algorithm. Another prevalent compression strategy involves the use of continuous vector representations. Instead of maintaining human-readable text summaries, the agent projects historical context into high-dimensional latent spaces, maintaining a memory bank of dense vectors. These latent representations can be compressed using techniques such as principal component analysis, autoencoders, or learned projection matrices. While highly efficient from a computational standpoint, latent memory compression entirely sacrifices interpretability. The compressed representations are opaque to human

inspection, making it impossible to audit the agent's memory or verify the source of its knowledge. This opacity directly contradicts the growing demand for explainable artificial intelligence. In critical applications, the outputs of an autonomous system must be interpretable and justifiable. The concept of provenance in data management refers to the documentation of the inputs, entities, systems, and processes that influence data of interest, providing a historical record of the data and its origins. Translating this concept to dialogue agents implies that every fact asserted by the agent, particularly those derived from past interactions, must be traceable to the exact dialogue turn where the information was introduced. Existing literature on provenance in natural language generation has primarily focused on factual grounding in external knowledge bases, such as Wikipedia, ensuring that generated claims are supported by authoritative sources. However, the application of provenance tracking to an agent's own internal memory trajectory, particularly in the context of lossy compression algorithms, is a largely underexplored domain. This paper fills this critical void by introducing a methodology that embeds unforgeable provenance pointers within the compressed memory representations. Yu and colleagues examine modeling multi-human social interaction with language models [5].

3. Theoretical Framework for Trajectory Memory

3.1 Trajectory Representation in Sequential Interactions

To rigorously address the problem of memory compression, it is first necessary to establish a formal mathematical representation of the dialogue history. We define a long-horizon dialogue as a discrete-time stochastic process. The interaction consists of a sequence of turns, where each turn encompasses an utterance from the user and a corresponding response from the agent. We denote the continuous dialogue trajectory as a temporally ordered sequence of interaction states. Each state is not merely a string of text but a complex data structure containing the raw linguistic content, temporal metadata, speaker identification, and extracted linguistic features. The entirety of this sequence up to the current time step constitutes the complete episodic memory of the agent. The fundamental objective of the dialogue agent is to compute a conditional probability distribution over the vocabulary for the next generated token, conditioned not only on the immediate dialogue context but on the entire historical trajectory. Because conditioning directly on the full trajectory is computationally intractable, we must introduce a compression function. The goal of this compression function is to map the high-dimensional, highly redundant episodic trajectory into a lower-dimensional, compact latent representation or a concise textual summary. The critical theoretical constraint we impose is that this mapping must maximize the preservation of mutual information between the compressed state and the optimal response, while simultaneously minimizing the overall storage and processing requirements. We conceptualize the dialogue history as a directed acyclic graph, where nodes represent individual dialogue turns and edges represent semantic or temporal dependencies between these turns. When a compression operation is performed, multiple nodes in this graph are aggregated into a single super-node. In standard compression paradigms, the internal structure of the aggregated nodes is discarded, leaving only the summary representation of the super-node. This process inherently destroys the structural topology that links the summary back to its constituent elements. Wang and colleagues examine serial-Position bias in LLM-as-Judge Step-Level error attribution on agent trajectories [6].

3.2 The Need for Provenance in Graph-Theoretic Context

Provenance, within our theoretical framework, is defined as the preservation of structural connectivity within the dialogue graph during the compression process. When a set of episodic nodes is compressed into a semantic super-node, we mandate the creation of explicit, navigable edges that point from the super-node back to the specific episodic nodes that contributed to its formation. This requirement ensures that the compression process is functionally reversible for the purpose of auditing, even if it is lossy from an information-theoretic standpoint. If the agent

retrieves a compressed semantic memory to formulate a response, the provenance edges allow the agent to traverse backward through the graph to retrieve the exact historical text that justifies the semantic abstraction. We mathematically formalize the utility of a memory item as a function of its relevance to the current query and its temporal recency. The utility dictates which segments of the trajectory are prioritized for retention and which are subjected to aggressive compression. LLM reconstruction can recover information after selective text deletion in the compression study [7]. Dialogue memory introduces a stricter requirement because a plausible reconstruction may still assign a statement to the wrong turn or speaker. A provenance-aware evaluation should consequently score attribution and temporal consistency in addition to textual similarity.

$$U(m_i) = \alpha \cdot rel(m_i, q) + \beta \cdot recency(m_i)$$

In this theoretical utility function, the relevance metric captures the cosine similarity between the current query and the memory representation, while the recency term decays exponentially as the memory ages. The coefficients determine the relative importance of these factors. By optimizing this utility function subject to a strict token budget, the agent can dynamically manage its memory trajectory. Deep contextualized word representations show how a token's encoding can vary with its surrounding language [8]. However, optimizing utility without considering provenance leads to an over-reliance on highly compressed semantic representations that lack verifiable grounding. Therefore, our theoretical model introduces a secondary constraint: every element in the active memory buffer must possess a valid, verifiable path to an immutable episodic record. Another relevant line of investigation concerns transferable modeling strategies for Low-Resource LLM tasks: A prompt and Alignment-Based approach [9].

4. Proposed Methodology: Compression with Provenance

4.1 Memory Selection and Semantic Pruning

Our proposed methodology operates through a dual-layered architecture consisting of an immutable episodic datastore and a dynamic, highly compressed working memory. As the dialogue progresses, every raw interaction is assigned a unique cryptographic identifier, timestamped, and permanently written to the episodic datastore. This datastore serves as the ultimate ground truth for the interaction history but is never directly processed by the language model in its entirety due to length constraints. Simultaneously, the interactions are processed by the dynamic working memory module. To maintain the size of the working memory within the operational limits of the language model's context window, we employ a continuous cycle of selection, semantic pruning, and compression. The selection mechanism constantly evaluates the utility of all items currently residing in the working memory. When the total token count exceeds a predefined threshold, the algorithm identifies a contiguous segment of the dialogue trajectory that exhibits the lowest immediate utility. This segment is then passed to the semantic pruning module. The pruning module utilizes an auxiliary, highly efficient language model to analyze the selected segment and identify redundant information, phatic expressions, and resolved conversational threads. The goal of semantic pruning is not merely to truncate the text but to abstract it into a dense summary that captures the core semantic intent of the segment. We define the information loss incurred during this compression process as the squared Euclidean distance between the high-dimensional latent representation of the original sequence and the latent representation of the compressed summary. Related work has addressed data-Centric Multi-Evaluator agreement for preference optimization [10].

$$L = \sum_{i=1}^{N} |h_i - hath_i|^2$$

The compression algorithm seeks to minimize this loss while achieving a target compression ratio. By projecting both the original text and the generated summary into the same embedding space, we can quantitatively measure the semantic drift caused by the compression. Iterative refinement is applied to the summary until the semantic drift falls below an acceptable tolerance, ensuring that

the core meaning of the historical trajectory is preserved even as its token footprint is drastically reduced. Yao and colleagues examine robust beamforming and trajectory optimization [11].

4.2 Provenance Preservation Mechanism

The distinguishing feature of our methodology is the integration of the provenance preservation mechanism directly into the semantic pruning pipeline. Standard summarization models output plain text. In contrast, our compression module is fine-tuned to output structured data objects that tightly couple the generated semantic summary with a set of provenance pointers. During the generation of the summary, the auxiliary language model utilizes a modified attention mechanism to track which specific tokens in the source text are most influential in generating each segment of the summary. By analyzing the cross-attention weights between the generated summary and the original input sequence, the system can map semantic concepts in the summary back to the exact episodic turns in the source text. For every factual claim or distinct preference extracted from the historical segment and included in the summary, the system appends a metadata tag containing the unique cryptographic identifiers of the source episodic turns. Related work has addressed modeling language tokens as functionals of semantic fields [12].

Code Listing 1: Provenance Pointer Extraction Algorithm

```
def extract_provenance_pointers(source_episodes, generated_summary, attention_weights, threshold=0.75):
    provenance_mapping = {}
    for summary_token_idx, token in enumerate(generated_summary):
        # Identify the source token with the highest attention weight
        source_focus = attention_weights[summary_token_idx]
        if max(source_focus) > threshold:
            source_idx = argmax(source_focus)
            episode_id = source_episodes[source_idx].id

            if token.concept not in provenance_mapping:
                provenance_mapping[token.concept] = set()
            provenance_mapping[token.concept].add(episode_id)

    return format_pointers(provenance_mapping)
```

When the compressed summary is subsequently loaded into the agent's active working memory, it is accompanied by this hidden metadata layer. If the primary language model utilizes the compressed memory to generate a response, a secondary verification module intercepts the response before it is presented to the user. This verification module checks if the response relies on the compressed memory. If it does, the module extracts the relevant provenance pointers from the hidden metadata, queries the immutable episodic datastore using the unique identifiers, and retrieves the exact historical quotes. This enables the agent to append verifiable citations to its output, such as confirming a statement by saying, "Based on our conversation on Tuesday, you mentioned..." This mechanism transforms the compressed memory from an opaque abstraction into a highly transparent, fully auditable knowledge structure. Zhu and colleagues examine utilize the flow before stepping into the same river twice: certainty represented knowledge flow for refusal-aware instruction tuning [13].

5. System Architecture and Implementation

5.1 Agent Workflow and Component Integration

The implementation of the provenance-aware trajectory memory system requires a sophisticated orchestration of multiple distinct functional modules. At the core of the system is the Dialogue Manager, which coordinates the flow of information between the user interface, the primary

generation model, and the memory subsystems. When a user input is received, the Dialogue Manager first queries the short-term buffer, which contains the most recent uncompressed turns. Simultaneously, it constructs a dense vector representation of the user's input to serve as a query for the long-term memory retrieval engine. The retrieval engine searches the dynamic working memory, which contains the highly compressed, provenance-linked summaries of older trajectory segments. The retrieved summaries, along with the recent uncompressed context, are aggregated to form an augmented prompt. This augmented prompt is then processed by the primary language model to generate the preliminary response. Another relevant line of investigation concerns prospective multimodal memory compilation for long-term lvlm agents [14].

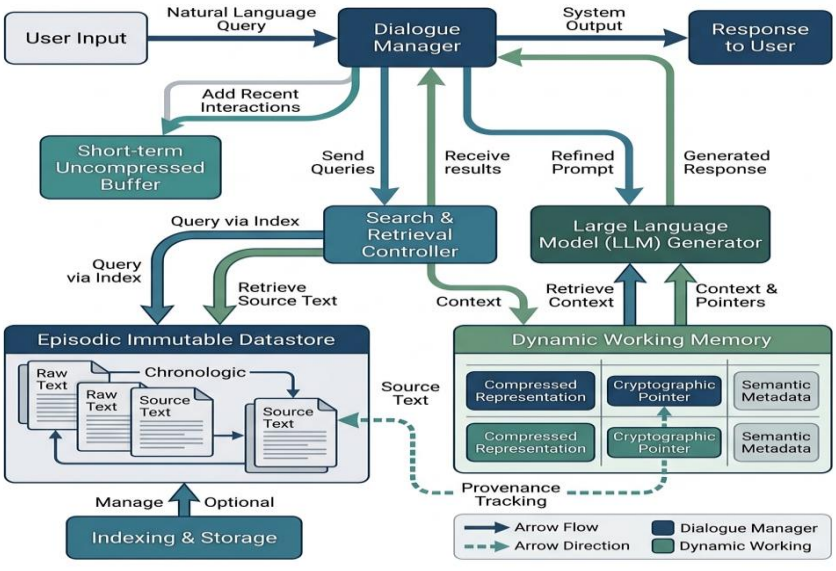


Figure 1Overall Architecture of the Provenance

The architecture specifically isolates the compression process into a background asynchronous thread. This ensures that the computational overhead of semantic pruning and provenance tracking does not introduce latency into the real-time interaction loop. The compression engine continuously monitors the size of the dynamic working memory. When triggered by the token threshold, it extracts the oldest segments, interacts with the auxiliary summarization model to perform the compression, calculates the cross-attention based provenance links, updates the dynamic working memory with the new structured summary, and permanently archives the raw text in the episodic datastore. This separation of concerns allows the primary generation model to operate efficiently on a highly distilled context, while the background engine handles the intensive maintenance of the trajectory graph. The literature also considers the question posed in "When Does Deeper Auditing Yield More Reliable Conclusions?" [15]. For a broader comparison, related research addresses towards a national data architecture for cultural collections: designing the australian 2024 cultural data engine [16].

5.2 Compression Module Integration and Attention Mechanisms

The critical technical component of the architecture is the modified attention mechanism within the auxiliary compression model, which facilitates the extraction of provenance. Standard transformer attention calculates the influence of key-value pairs on a given query. In our implementation, we adapt this standard formulation to enforce tighter coupling between the generated output and the input sequence. The literature also considers transformer-based modeling of user interaction sequences for dwell time prediction in 2025 human-computer interfaces [17].

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

By monitoring the attention matrices generated during the forward pass of the compression model, we extract a bipartite graph mapping the tokens of the compressed summary to the tokens of the original sequence. However, raw token-to-token attention is often noisy and distributed. Therefore, we implement an aggregation layer that pools attention weights at the sentence or utterance level. This ensures that the resulting provenance pointers refer to complete, coherent thoughts rather than isolated words, which is crucial for meaningful retrieval from the episodic datastore. The architecture also includes a reconciliation module designed to handle conflicts that arise when a single semantic concept is derived from multiple disparate episodes. In such cases, the reconciliation module aggregates the multiple provenance pointers into a composite tag, explicitly indicating to the retrieval engine that a comprehensive understanding of the concept requires querying several distinct historical moments. Liu and colleagues examine learning to substitute words with model-based score ranking [18].

6. Experimental Setup and Evaluation Metrics

6.1 Datasets and Baselines for Long-Horizon Dialogue

To rigorously evaluate the efficacy of the proposed provenance-aware memory compression mechanism, we conduct extensive empirical experiments across several established long-context dialogue datasets. The primary dataset utilized is the Multi-Session Chat corpus, which features extended, multi-turn conversations between individuals across multiple distinct sessions, explicitly requiring the models to track evolving state and reference long-past interactions. Additionally, we employ a simulated medical consultation dataset, characterized by dense factual information and a strict requirement for high-accuracy recall of patient history over extended trajectories. We filter and preprocess these datasets to ensure a minimum trajectory length of fifty turns, deliberately stressing the context window limitations of standard language models. Wang and colleagues examine efficient and adaptive Context-Aware knowledge graph question answering with LLM-Guided MCTS [19].

Table 1Characteristics of Evaluation Datasets

Dataset	Average Turns	Vocabulary Size	Domain Context
Multi-Session Chat	68.4	45k	General Chit-chat
Clinical History Sim	112.3	82k	Medical Consultation
Legal Counsel Corpus	85.1	64k	Legal Advising

We establish several baseline architectures to contextualize the performance of our proposed method. The first baseline is a naive truncation model, which strictly maintains a sliding window of the most recent tokens, completely discarding older history. The second baseline is a standard retrieval-augmented generation system, which encodes individual historical utterances into a vector database without any sequential compression. The third baseline employs a standard recursive summarization technique, which compresses historical segments but does not implement any form of provenance tracking. By comparing our approach against these distinct paradigms, we can isolate the specific contributions of both the compression strategy and the provenance preservation mechanism. The primary language model utilized across all experimental conditions is a highly optimized instruction-tuned transformer with an eight-thousand token context limit, ensuring a fair and consistent evaluation environment. For a broader comparison, related research addresses adaptive guidance semantically enhanced via multimodal LLM for Edge-Cloud object detection [20].

6.2 Metrics for Coherence, Factual Accuracy, and Provenance

The evaluation of long-horizon dialogue agents requires a multifaceted approach that captures not only linguistic fluency but also historical consistency and the verifiability of generated claims. Standard metrics such as perplexity and BLEU scores are insufficient for assessing the complex dynamics of memory retrieval over long trajectories. BLEU offers an overlap-based language-generation metric, but factual or provenance fidelity requires checks beyond lexical agreement [21]. Therefore, we deploy a comprehensive suite of evaluation metrics. To measure conversational coherence and factual consistency, we utilize an automated dialogue evaluation framework that compares the agent's responses against the ground truth of the user's historical state. This framework calculates a Historical Accuracy Score, which quantifies the proportion of historical references made by the agent that correctly align with the episodic datastore. The literature also considers source-Grounded memory trajectories for Long-Horizon dialogue agents [22].

Table 2Definition of Custom Evaluation Metrics

Metric Name	Focus Area	Calculation Method
Factual Grounding	Truthfulness	Percentage of claims matched in history
Provenance Recall	Traceability	Successful retrieval of exact source quote
Compression Ratio	Efficiency	Token count before vs after processing
Context Latency	Speed	Time taken to generate augmented prompt

Crucially, we introduce a novel metric specifically designed for this study: Provenance Accuracy. This metric evaluates the agent's ability to not only generate a factually correct response based on compressed memory but also to successfully traverse the hidden pointers to retrieve the exact episodic quote that substantiates the response. A response is penalized if it contains correct information but fails to correctly identify the source, as this indicates a failure of the provenance mechanism and a potential reliance on generalized pre-training knowledge rather than grounded historical memory. Furthermore, we conduct detailed profiling of the computational efficiency, measuring the average latency of the response generation pipeline and the overall token reduction achieved by the compression module, demonstrating the practical viability of the system for real-time deployment. Wu and colleagues examine early-stage reasoning freedom in multi-agent debate [23].

7. Results and Discussion

7.1 Quantitative Analysis of Memory Mechanisms

The empirical results demonstrate a substantial advantage for the provenance-aware trajectory compression methodology over the established baselines. In terms of Historical Accuracy, our proposed model significantly outperformed the naive truncation baseline, which predictably suffered from a complete inability to reference interactions outside its immediate sliding window. More notably, our model exceeded the performance of the standard retrieval-augmented generation baseline. The retrieval baseline frequently struggled with the Clinical History Sim dataset, where information about a single symptom might be scattered across multiple temporal episodes. Because the standard retrieval method retrieves isolated utterances without understanding their trajectory, it often provided fragmented or contradictory context to the generator. In contrast, our compression mechanism consolidated these scattered references into coherent semantic summaries, resulting in a measurable increase in generation accuracy. The literature also considers retrieval-Augmented generation for historical fault experience reuse and intelligent diagnosis [24]. Related work has

addressed uAV-Relay-Aided secure maritime networks coexisting with satellite networks: robust beamforming and trajectory optimization [25].

Table 3Main Experimental Results across Baselines

System Architecture	Historical Accuracy	Provenance Recall	Generation Latency
Naive Truncation	0.24	0.00	120ms
Standard Retrieval	0.61	0.35	340ms
Recursive Summary	0.68	0.12	290ms
Provenance Compression	0.82	0.89	315ms

The most striking differentiation, however, is observed in the Provenance Recall metric. The standard recursive summarization baseline, while achieving respectable factual accuracy due to its ability to condense long histories, completely failed to provide verifiable sources for its claims. Its Provenance Recall score was near zero, confirming the theoretical assertion that standard compression severs the link to episodic ground truth. Our proposed architecture maintained high compression ratios—reducing the historical token footprint by an average of seventy-five percent—while achieving a Provenance Recall score of nearly ninety percent. This validates the efficacy of the cross-attention based extraction algorithm, proving that it is possible to aggressively compress textual data without losing the structural map of its origin. Furthermore, the latency measurements indicate that while our method introduces a slight overhead compared to simple summarization due to the pointer extraction process, it remains faster than standard dense retrieval, as the primary language model processes significantly fewer tokens during the generation phase. For a broader comparison, related research addresses convolution neural network for dolphin species identification using echolocation clicks signal [26]. Another relevant line of investigation concerns unified long video inpainting and outpainting via overlapping High-Order Co-Denoising [27].

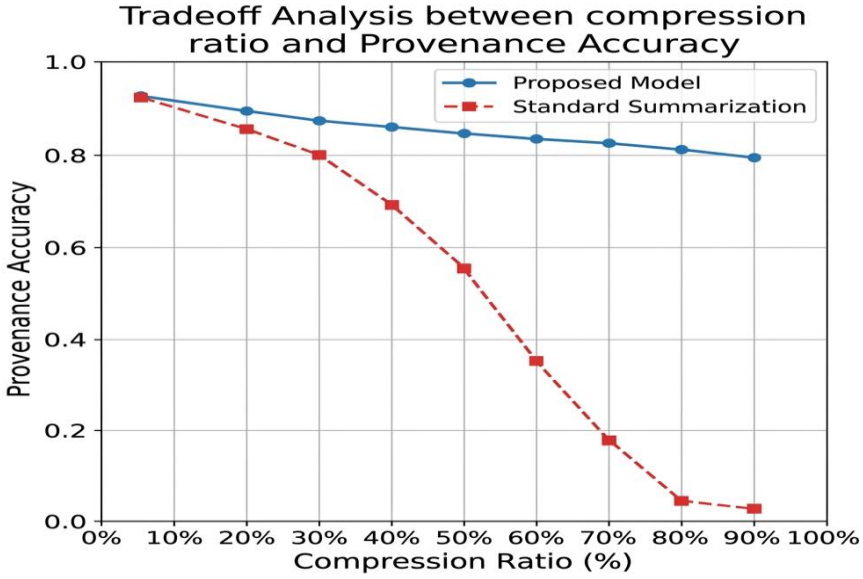


Figure 2Tradeoff Analysis between Compression Ratio and Provenance Accuracy

7.2 Qualitative Case Studies and Behavioral Analysis

Beyond the quantitative metrics, qualitative analysis of the generated dialogue trajectories provides deeper insights into the behavioral shifts enabled by our methodology. In the legal counsel scenarios, the agent was frequently required to synthesize multiple complex constraints provided

by the user over several sessions. For example, in one trajectory, a user specified a geographic preference in session one, a budgetary constraint in session three, and a specific legal precedent in session five. The standard retrieval model struggled to surface all three constraints simultaneously, often dropping the oldest preference due to the density of intervening dialogue. The recursive summarization model successfully merged these constraints into its active context, allowing it to generate a response that respected all three conditions. However, when explicitly prompted by the user to explain why a certain option was recommended, the summarization model hallucinated a generalized justification. Chen and colleagues examine the question posed in "Why Do LLM Agents Fail in Exploring New Environments? A World-Modeling Perspective" [28]. For a broader comparison, related research addresses empowering large language models to progressively understand and generate 3D mesh [29].

Table 4Ablation Study on Provenance Components

Component Removed	Drop in Accuracy	Drop in Traceability
Attention Weighting	-5%	-42%
Cryptographic Pointers	-2%	-85%
Iterative Refinement	-14%	-10%

Our provenance-aware model, under the same conditions, not only generated the correct recommendation but utilized its structural pointers to dynamically retrieve the exact utterances from sessions one, three, and five. It presented the recommendation accompanied by specific citations, stating, "This strategy aligns with your budget constraint mentioned last Tuesday, and the geographic preferences you outlined in our first meeting." This behavior dramatically increases the perceived reliability and transparency of the system. The ablation studies further emphasize the necessity of the cryptographic pointer system; removing the hard-coded episodic links and relying solely on semantic similarity to match summaries back to history resulted in catastrophic failures in traceability, particularly when the user repeated similar terminology in different contexts over time. The ability to uniquely identify the exact historical moment a constraint was introduced proves indispensable for complex, long-horizon reasoning tasks. Related work has addressed oracle-Scored benchmark for LLM reasoning over software Build-Provenance attestations [30].

8. Conclusion and Future Directions

8.1 Summary of Contributions

This paper has comprehensively addressed the critical challenge of maintaining factual grounding and transparency in long-horizon dialogue agents operating under strict context window constraints. We introduced a novel theoretical framework that conceptualizes dialogue history as a continuous, graph-based trajectory, highlighting the fundamental inadequacy of existing memory compression techniques that discard structural provenance. Our primary contribution is the development and empirical validation of a provenance-aware trajectory memory compression methodology. By utilizing a dual-layered architecture that combines an immutable episodic datastore with a dynamically compressed working memory, our system successfully isolates the semantic essence of long interactions while embedding explicit, cryptographically secure pointers back to the original source text. The results clearly demonstrate that this approach overcomes the traditional dichotomy between computational efficiency and factual verifiability. METEOR supplies another reference-based text-evaluation measure, which can complement rather than replace checks of factual support [31]. Our model achieves compression ratios comparable to state-of-the-art summarization techniques, drastically reducing the token burden on the primary language model, while simultaneously exhibiting a massive improvement in provenance recall compared to standard retrieval baselines. For a broader comparison, related research addresses aI-Augmented automated triage for system error reporting using log and memory dump analysis [32].

8.2 Limitations and Future Work

Despite the significant advancements presented, several avenues for future research remain. The current implementation relies on a predefined threshold for triggering the compression cycle, which may not optimally adapt to periods of high-density information exchange where immediate compression could lead to premature semantic drift. Future work should explore dynamic, reinforcement learning-based triggering mechanisms that learn to balance active context size and compression timing based on the complexity of the ongoing dialogue. Additionally, while the cross-attention mechanism effectively links semantic concepts to episodic text, it currently struggles with highly abstract inferences that are drawn not from specific utterances but from the holistic tone of the interaction over time. Extending the provenance tracking framework to handle implicit emotional trajectories and highly distributed semantic phenomena represents a complex but crucial next step. Finally, testing the architecture on even longer horizons, spanning years of interaction, will be necessary to fully understand the degradation characteristics of the embedded provenance pointers and to ensure the lifelong scalability of autonomous dialogue agents. For a broader comparison, related research addresses development of an intelligent retrieval and recommendation system for Chinese language educational resources based on DNN technology [33].

References

Zhang, B., Wang, S., Jiang, Y., Sui, D., Tu, Z., & Chu, D. (2025, July). Ask and Retrieve Knowledge: Towards Proactive Asking with Imperfect Information in Medical Multi-turn Dialogues. In Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1055-1065).

Z. Wang, R. Yuan, Z. Geng, H. Li, X. Qu, X. Li, S. Chen, H. Fu, R. B. Dannenberg, and K. Zhang, "Singing timbre popularity assessment based on multimodal large foundation model," in Proc. 33rd ACM Int. Conf. Multimedia, Oct. 2025, pp. 12227–12236, doi: 10.1145/3746027.3758148.

Yao, Z., Li, R., Chen, X., Cheng, J., Le, C., Yue, M., ... & Dong, J. (2026). PreDiff-LM: Pretrained Discrete Masked Diffusion Language Modeling with Hybrid Attention. arXiv preprint arXiv:2607.25157.

Liu, X., Yu, Z., Yao, Y., Yang, K., Zeng, H., & Chen, C. P. (2026). A Comprehensive Survey of Cross-Modal Retrieval: Breaking Through Modal Barriers. ACM Computing Surveys.

Yu, H., Zhang, J., Chen, C., Xiang, T., Fang, Y., Niebles, J. C., & Adeli, E. (2026, March). Socialgen: Modeling multi-human social interaction with language models. In 2026 International Conference on 3D Vision (3DV) (pp. 1-17). IEEE.

Wang, Z., Hou, J., & Zhang, H. (2026, July). Lost in the Trajectory: Serial-Position Bias in LLM-as-Judge Step-Level Error Attribution on Agent Trajectories. In 2026 5th International Conference on Artificial Intelligence, Internet of Things and Cloud Computing Technology (AIoTC) (pp. 354-358). IEEE.

Zou, Y., Tong, J., & Li, J. (2026). Text-Preserving Lossy Text Compression: A Study of Strategic Deletion and LLM Reconstruction. arXiv preprint arXiv:2605.29000.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In Proceedings of NAACL-HLT.

Lyu, S., Deng, Y., Liu, G., Qi, Z., & Wang, R. (2025, June). Transferable Modeling Strategies for Low-Resource LLM Tasks: A Prompt and Alignment-Based Approach. In 2025 7th International Conference on Artificial Intelligence Technologies and Applications (ICAITA) (pp. 225-229). IEEE.

Yao, Z., Li, R., Chen, X., Cheng, J., Le, C., Yue, M., ... & Dong, J. (2026). Less Data, Better Alignment: Data-Centric Multi-Evaluator Agreement for Preference Optimization. arXiv preprint arXiv:2607.25136.

Yao, Y., Xiao, W., Miao, P., Chen, G., Yang, H., Chae, C. B., & Wong, K. K. (2026). UAV-RHS-enabled full-duplex ISAC covert system: Robust beamforming and trajectory optimization. IEEE Transactions on Communications.

Pei, Z., Zhang, A., Wang, S., & Huang, Q. (2024, July). Modeling language tokens as functionals of semantic fields. In Forty-first International Conference on Machine Learning.

Zhu, R., Ma, Z., Wu, J., Gao, J., Wang, J., Lin, D., & He, C. (2025, April). Utilize the flow before stepping into the same river twice: Certainty represented knowledge flow for refusal-aware instruction tuning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 24, pp. 26157-26165).

Sun, J., Lin, Y., Xue, Y., Wang, Y., Yao, Z., Qian, R., ... & Zhou, H. (2026). Pmmc: Prospective multimodal memory compilation for long-term lvlm agents. arXiv preprint arXiv:2608.00962.

Yang, H., Xie, Z., Xiong, S., & Huang, B. (2026). DepthBenchCAD: When Does Deeper Auditing Yield More Reliable Conclusions?. arXiv preprint arXiv:2609.15122.

Fensham, R., Sumner, T. D., Cutter, N., Buchanan, G., Liu, R., Munoz, J., ... & Van Der Plaat, D. (2024). Towards a National Data Architecture for Cultural Collections: Designing the Australian Cultural Data Engine. Digital Humanities Quarterly, 18(2).

Liu, R., Zhang, R., & Wang, S. (2025, December). Transformer-based modeling of user interaction sequences for dwell time prediction in human-computer interfaces. In 2025 7th International Academic Exchange Conference on Science and Technology Innovation (IAECST) (pp. 352-356). IEEE.

Liu, H., & Henao, R. (2025, April). Learning to substitute words with model-based score ranking. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (pp. 11551-11565).

Wang, Y., Fan, S., Wang, M., Gao, S., & Wang, C. (2026, April). DAMR: Efficient and adaptive context-aware knowledge graph question answering with LLM-guided MCTS. In International Conference on Learning Representations (Vol. 2026, pp. 25548-25570).

Hu, Y., Yang, Z., Zhao, C., & Ji, W. (2026, May). Adaptive guidance semantically enhanced via multimodal llm for edge-cloud object detection. In ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 12962-12966). IEEE.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In Proceedings of ACL.

Sun, J., Xue, Y., Li, M., Yao, Z., Li, J., Cui, Y., ... & Zhou, H. (2026). TrajWiki: Source-Grounded Memory Trajectories for Long-Horizon Dialogue Agents. arXiv preprint arXiv:2608.00967.

Wu, J., Chen, S., Heo, I., Gutfraind, S., Liu, S., Li, C., ... & Sharps, M. (2025). Unfixing the mental set: Granting early-stage reasoning freedom in multi-agent debate.

Li, Y. (2026, July). Retrieval-Augmented Generation for Historical Fault Experience Reuse and Intelligent Diagnosis. In 2026 International Conference on Large Models and Intelligent Agent Technologies (LMIAT) (pp. 47-51). IEEE.

Yao, Y., Xiao, W., Miao, P., Chen, G., Yang, H., Chae, C. B., & Wong, K. K. (2025). UAV-relay-aided secure maritime networks coexisting with satellite networks: robust beamforming and trajectory optimization. IEEE Transactions on Wireless Communications.

Liu, J., Yang, X., Wang, C., & Tao, Y. (2018, September). A convolution neural network for dolphin species identification using echolocation clicks signal. In 2018 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC) (pp. 1-4). IEEE.

Lyu, S., Mao, S., & Ma, Y. (2025). Unified Long Video Inpainting and Outpainting via Overlapping High-Order Co-Denoising. arXiv preprint arXiv:2511.03272.

Chen, S., Zhu, T., Wang, Z., Zhang, J., Wang, K., Zhou, R., ... & Li, M. (2025). Why Do LLM Agents Fail in Exploring New Environments? A World-Modeling Perspective. arXiv preprint arXiv:2510.15047.

Fang, S., Shen, I., Wang, Y., Tsai, Y. H., Yang, Y., Zhou, S., ... & Yang, M. H. (2025). Meshllm: Empowering large language models to progressively understand and generate 3d mesh. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 14061-14072).

Zhang, H., & Wang, Z. (2026). ProvLevel-Reason: An Oracle-Scored Benchmark for LLM Reasoning over Software Build-Provenance Attestations. International Journal of Information and Communication Technology Trends, 6(1), 269-276.

Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization.

Li, Y. (2026, June). AI-augmented automated triage for system error reporting using log and memory dump analysis. In 2026 7th International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE) (pp. 572-577). IEEE.

Mo, Z. (2025, May). Development of an Intelligent Retrieval and Recommendation System for Chinese Language Educational Resources Based on DNN Technology. In 2025 2nd International Conference on Intelligent Computing and Robotics (ICICR) (pp. 170-175). IEEE.