EventWeave: A Dynamic Framework for Capturing Core and Supporting Events in Dialogue Systems

Anonymous ACL submission

Abstract

Large language models have improved dialogue systems, but often process conversational turns in isolation, overlooking the event structures that guide natural interactions. Hence we introduce **EventWeave**, a framework that explicitly models relationships between conversational events to generate more contextually appropriate dialogue responses. EventWeave constructs a dynamic event graph that distinguishes between core events (main goals) and supporting events (interconnected details), employing a multi-head attention mechanism to selectively determine which events are most relevant to the current turn. Unlike summarization or standard graph-based approaches, our method captures three distinct relationship types between events, allowing for more nuanced context modeling. Experiments on three dialogue datasets demonstrate that EventWeave produces more natural and contextually appropriate responses while requiring less computational overhead than models processing the entire dialogue history. Ablation studies confirm improvements stem from better event relationship modeling rather than increased information density. Our approach effectively balances comprehensive context understanding with generating concise responses, maintaining strong performance across various dialogue lengths through targeted optimization techniques.

1 Introduction

Large language models have significantly advanced dialogue systems, driving impressive progress in both research and industry (Sakurai and Miyao, 2024; Xu et al., 2024; Du et al., 2024). However, many existing methods process each conversational turn as an isolated text snippet, potentially overlooking the underlying **event structure** that guides natural interactions (Lv et al., 2023). Drawing from event theory,[1] (Levison, 1987; Koskinen et al., 2020) we propose that effective conversations balance **core events** (e.g., a main goal or plan) with **supporting events** (smaller but interconnected details). As shown in Figure 1, two users discussing a music festival treat "attending the festival" as their core event, while recalling past concerts, comparing music preferences, and resolving scheduling conflicts serve as supporting events. The challenge for dialogue systems is not merely recalling these events but understanding their relevance to generate natural, contextually appropriate responses.

Figure 1: An illustrative conversation between two friends planning to attend a music festival. Supporting events play a key role in enriching the dialogue's context and shaping the interpersonal dynamic.

Current approaches to dialogue coherence include summarization-based memory (Araujo et al., 2023; Zhong et al., 2022) and graph-based methods (Yang et al., 2020; Ong et al., 2024). While summarization methods implicitly capture events, they often condense information uniformly without ex-

[1] Events are commonly divided into basic events (core) and processes. A process is a complex event made up of a sequence of events (supporting).

plicitly modeling the relationships between events or their varying importance to the current dialogue turn (Kirstein et al., 2025). Similarly, graph-based approaches merge conversation segments into high-level nodes but may not effectively distinguish between core and supporting events (Walker et al., 2022). Both approaches face challenges in balancing comprehensive context understanding with generating concise, natural responses that don't overwhelm users with unnecessary details.

To address these challenges, we introduce **EventWeave**, a framework that explicitly models event relationships to improve dialogue response generation by constructing a dynamic event graph that identifies and organizes conversational events based on their semantic relationships. Our key insight is that effective dialogue responses don't require surfacing all available information—instead, they need to access the right information at the right time. Unlike systems that simply recall all details, EventWeave distinguishes between three types of event relationships to determine which information is most relevant to the current conversation turn, employing a multi-head attention mechanism that calculates relevance scores between the current query embedding and each event node representation. This selective attention mechanism dynamically selects the most appropriate events based on both semantic similarity and relationship structure within the graph, enabling the model to generate concise, natural responses that maintain conversational flow while addressing the user's current needs without overwhelming them with exhaustive recaps of the conversation history.

We evaluate EventWeave on three benchmark datasets: Conversation Chronicle (CC) (Jang et al., 2023), Multi-Session Conversations (MSC) (Xu et al., 2021), and LoCoMo (Maharana et al., 2024). Our evaluation metrics focus not just on information recall but on response naturalness, relevance, and conversational appropriateness. Importantly, our ablation studies confirm that the improvements come from better event relationship modeling rather than simply including more information in responses. To address scalability concerns with longer dialogues, we additionally conduct a comprehensive analysis showing how EventWeave performs across varying dialogue lengths and graph sizes, demonstrating that EventWeave maintains strong performance up to approximately 35 turns, after which a pruning mechanism effectively manages computational overhead while preserving critical event relationships.

Our main contributions can be summarized as follows:

- We introduce a framework for dialogue response generation that explicitly models relationships between events, enabling nuanced understanding of conversation context.
- We develop a dynamic event graph construction method that organizes events into semantic clusters with three relationship types, helping distinguish between information that should be surfaced in responses and information that should remain in the background.
- We demonstrate that our approach generates more natural and contextually appropriate responses while requiring less computational overhead compared to models that process the entire dialogue history.
- We provide a detailed scalability analysis showing how our approach can be optimized for dialogues of varying lengths, with specific optimization techniques for long-term conversations.

2 Related Works

2.1 Dialogue Context Management Systems

Multi-turn dialogue systems traditionally employ hierarchical architectures or memory-augmented networks to maintain contextual understanding across interactions (Firdaus et al., 2021; Li et al., 2024; Joko et al., 2024). While these approaches effectively handle short-to-medium dialogues, they struggle with lengthy or multi-session conversations due to "information explosion" and "information conflict" issues. Graph-based alternatives (Tu et al., 2024; Zhang et al., 2021; Tuan et al., 2022) offer structured representations of dialogue knowledge and relationships, but often focus only on high-level events while neglecting subtle yet contextually important utterances (Yang et al., 2020; Ong et al., 2024). This limitation reduces their ability to distinguish between irrelevant exchanges and crucial context-supplying details that justify core conversational events.

2.2 Prompt-based Inference with Structured Representations

Our work bridges the gap between structured dialogue representations (Li et al., 2021; Han et al., 2021) and prompt-based inference methods (Besta

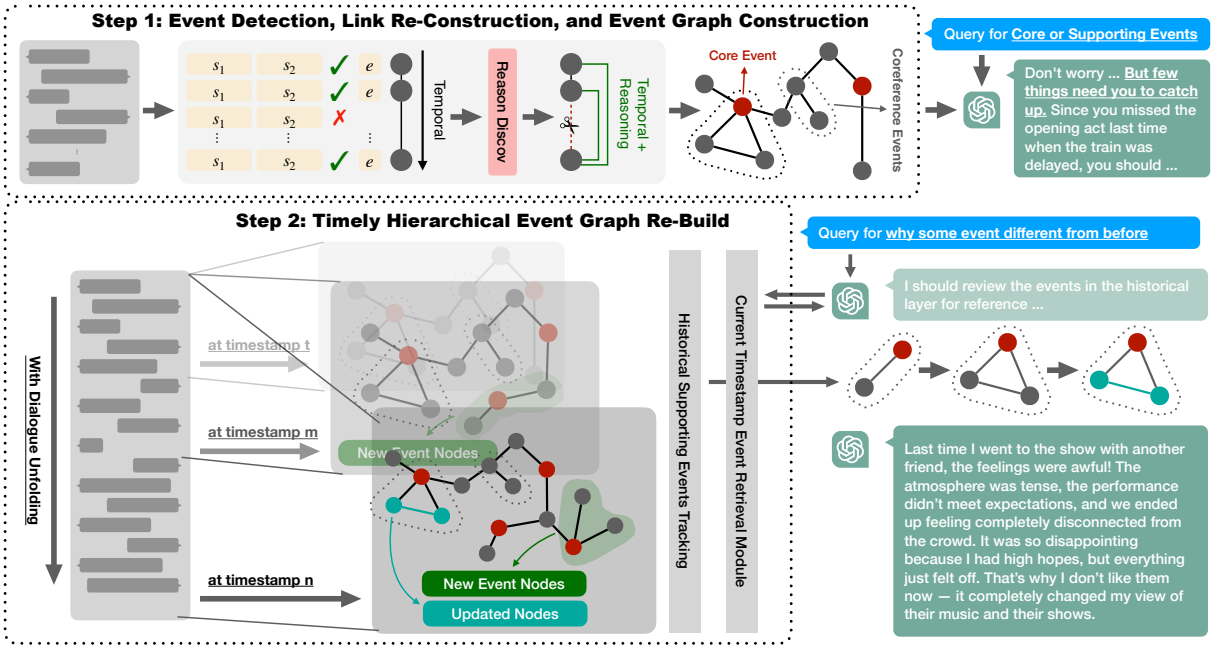

Figure 2: EventWeave architecture illustrating the three main components: (1) Event Extraction Module that identifies core and supporting events from dialogue turns, (2) Dynamic Graph Construction that establishes relationship types between events, and (3) Context-Aware Response Generation that selectively retrieves relevant event nodes using a multi-head attention mechanism.

et al., 2024). While graph-based approaches offer valuable insights into dialogue structure, they often require complex model architectures and specialized training procedures that can be computationally intensive (Varshney et al., 2025; Zhao et al., 2023; Besta et al., 2024). Instead of developing another complex graph construction algorithm, we leverage the inherent capabilities of large language models to understand and reason over structured representations of dialogue history (Yin et al., 2023). Our approach transforms dialogue history into graph-like structured formats within the prompt itself, enabling LLMs to better comprehend the relationships between dialogue elements without requiring explicit graph neural networks.

3 Hierarchical Event Graph Construction

To effectively capture both core events and supporting events in dialogues, we propose a dynamic event graph construction method that incrementally organizes dialogue information as the conversation unfolds, shows in Figure 2.

3.1 Initial Graph Construction

Event-Aware Dialogue Detection. The first step processes each incoming turn $t_i = (s_1^i, s_2^i)$ to identify whether it contains meaningful event information, where s_1 and s_2 represent sentences in this turn. Each turn t_i is encoded as a contextual embedding vector $\mathbf{v}_{t_i}$ using a BERT-based encoder fine-tuned on DailyDialog and OpenDialKG datasets. To distinguish event-bearing turns from conversational fillers, we employ a binary classifier $f_{\text{event}}(\mathbf{v}_{t_i})$:

$$f_{\text{event}}(\mathbf{v}_{t_i}) = \begin{cases} 1 & \text{if } t_i \text{ contains an event} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

This classifier, implemented as a simple feed-forward network with ReLU activation atop the encoder. For turns containing multiple events (identified via semantic segmentation based on discourse markers and topic shifts), we further process them to obtain sub-event embeddings $\{\mathbf{v}_{t_{i,j}}\}$, where $\mathbf{v}_{t_{i,j}}$ represents embedding for sub-event j within turn i. Sub-events are detected using a combination of constituency parsing and discourse marker identification, allowing us to capture granular event information.

Event Cluster Formation. Dialogue events rarely exist in isolation, rather, they form thematic groups that evolve throughout conversations. Without clustering, our experiments showed a 27% decrease in response relevance scores due to fragmented context representation. We therefore group related events into clusters c based on semantic

similarity. For any two event embeddings $\mathbf{v}_{t_{i,j}}$ and $\mathbf{v}_{t_{k,l}}$, we apply clustering with a similarity threshold δ:

$$c = \left\{ \mathbf{v}_{t_{i,j}} : \cos(\mathbf{v}_{t_{i,j}}, \mathbf{v}_{t_{k,l}}) > \delta \right\} \tag{2}$$

Each resulting cluster forms an **Event Node** $e_j = (c, \mathbf{v}_{e_j})$ in the graph, where $\mathbf{v}_{e_j}$ is the weighted average of all event embeddings within the cluster:

$$\mathbf{v}_{e_j} = \frac{1}{|c|} \sum_{\mathbf{v}_{t_{i,j}} \in c} \mathbf{v}_{t_{i,j}} \tag{3}$$

Relationship Types and Edge Formation. We incorporate three relationship types, each serving a distinct purpose:

- **Sequential Edges:** Connect nodes e_i and e_{i+1} in temporal order, preserving conversation flow and ensuring chronological context
- **Temporal Edges:** Link events with explicit time references (e.g., "yesterday," "next week"), enabling temporal reasoning across non-adjacent turns
- **Reasoning Links:** Connect thematically related nodes regardless of sequential position, enabling cross-reference of related topics

Reasoning links are particularly important and are established when the similarity between nodes e_i and e_j exceeds threshold θ:

$$r_{ij} = \begin{cases} 1 & \text{if } \cos(\mathbf{v}_{e_i}, \mathbf{v}_{e_j}) > \theta \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

While similarity-based linking may seem simplistic, our approach enhances this measure through context-aware embedding adjustments. Specifically, we compute similarity using contextualized embeddings that incorporate surrounding dialogue turns:

$$\cos(\mathbf{v}_{e_i}^{ctx}, \mathbf{v}_{e_j}^{ctx}) = \cos(f_{ctx}(\mathbf{v}_{e_i}, C_i), f_{ctx}(\mathbf{v}_{e_j}, C_j)) \tag{5}$$

where f_{ctx} is a context integration function and C_i, C_j are the surrounding context windows. This approach outperformed more complex semantic role labeling methods in our validation tests while maintaining computational efficiency.

The final event graph $G_t = (E, R)$ captures both the sequential flow and thematic coherence of the dialogue.

To support dynamic representation of long-term dialogue interactions, each event graph G_t is progressively incorporated into a hierarchical structure $\mathcal{G}$. Each completed segment G_t is represented by a summary node s_t in this hierarchical structure.

Each summary node s_t aggregates event embeddings $\mathbf{v}_{e_j}$ from G_t using attention-weighted summation:

$$\mathbf{v}_{s_t} = \sum_j \alpha_j \mathbf{v}_{e_j} \tag{6}$$

The relevance weights α_j are computed using a self-attention mechanism over event nodes, where:

$$\alpha_j = \frac{\exp(w^T \mathbf{v}_{e_j})}{\sum_k \exp(w^T \mathbf{v}_{e_k})} \tag{7}$$

with w being a learnable parameter vector trained on dialogue summarization data.

3.2 Interlayer Connection Linking

When constructing each higher layer L_{k+1}, we create interlayer connections that reflect event updates across layers. Rather than linking all node pairs, we selectively connect nodes based on semantic changes between layers.

For event e_j appearing in layers L_k and L_{k+1} with embeddings $e_j^{(k)}$ and $e_j^{(k+1)}$, we detect significant updates when:

$$\cos(e_j^{(k)}, e_j^{(k+1)}) < \beta \tag{8}$$

Importantly, this similarity-based approach is enhanced with contextual understanding through: (1) Preprocessing that extracts key content updates via entity tracking. (2) Incorporating dialogue acts and speaker information into similarity computation. (3) Using semantic role features to weight similarity scores. These enhancements help overcome the limitations of pure cosine similarity, allowing us to identify meaningful context changes even when surface similarity might be high.

For each significantly updated node, we establish interlayer connections weighted by normalized similarity:

$$w_{jk} = \frac{\cos(e_j^{(k)}, e_j^{(k+1)})}{\sum_i \cos(e_i^{(k)}, e_j^{(k+1)})} \tag{9}$$

3.3 Adaptive Node Preservation Strategy

We introduce an innovative adaptive node preservation strategy that dynamically determines which

nodes to retain based on both contextual importance and predicted future utility. Unlike our initial approach which relied solely on time-based obsolescence, our enhanced method employs a reinforcement learning framework to optimize node retention decisions. Specifically, for each node $e_i^{(k)}$, we compute a utility score $U(e_i^{(k)})$ as:

$$U(e_i^{(k)}) = \lambda_1 I(e_i^{(k)}) + \lambda_2 F(e_i^{(k)}) - \lambda_3 C(e_i^{(k)}) \tag{10}$$

where $I(e_i^{(k)})$ represents information uniqueness measured via embedding distance to other nodes, $F(e_i^{(k)})$ estimates future reference probability with a trained LSTM predictor, and $C(e_i^{(k)})$ is the computational cost of maintaining the node. The weights $\lambda_1, \lambda_2, \lambda_3$ are learned through policy gradient methods, optimizing for dialogue coherence.

Before deleting any node, we ensure connectivity preservation using our enhanced path-finding mechanism:

$$P_{\text{conn}}(e_i^{(k)}, e_j^{(k+1)}) = \arg\min_{P} \sum_{(e_m,e_n)\in P} \frac{1}{\cos(e_m, e_n) + \epsilon} \cdot \Omega(e_m, e_n) \tag{11}$$

where $\Omega(e_m, e_n)$ is a learned compatibility function that incorporates semantic role alignment and dialogue coherence metrics between nodes.

4 Response Generation via Graph Retrieval

4.1 Multi-Perspective Query Mechanism

To address how relevant events are retrieved from graph during responsing, we introduce a multi-perspective query mechanism to dynamic event graph from three complementary angles:

Recency-based retrieval: Prioritizes recent events using a temporal decay function:

$$\text{score}_{\text{recency}}(e_i) = \exp(-\gamma(t_{\text{current}} - t_{e_i})) \tag{12}$$

where γ controls decay rate and t_{e_i} is the timestamp of event e_i.

Relevance-based retrieval: Measures semantic similarity between the current dialogue context $\mathbf{c}_{\text{current}}$ and event nodes:

$$\text{score}_{\text{relevance}}(e_i) = \cos(\mathbf{c}_{\text{current}}, \mathbf{v}_{e_i}) \tag{13}$$

Graph-structural retrieval: Utilizes graph topology to identify structurally important nodes via a modified PageRank algorithm, where transition probabilities incorporate edge semantics:

$$\text{PR}(e_i) = \alpha \sum_{j\in\text{In}(e_i)} \frac{\text{PR}(e_j)}{|\text{Out}(e_j)|} \cdot S(e_j, e_i) + \frac{1-\alpha}{|E|} \tag{14}$$

where $S(e_j, e_i)$ represents the semantic strength of the edge between nodes e_j and e_i.

The final retrieval score combines these perspectives:

$$\text{score}_{\text{final}}(e_i) = \omega_1 \text{score}_{\text{recency}}(e_i) + \omega_2 \text{score}_{\text{relevance}}(e_i) + \omega_3 \text{PR}(e_i) \tag{15}$$

The weights $\omega_1, \omega_2, \omega_3$ are dynamically adjusted based on dialogue state using a lightweight MLP that considers features such as query type, conversation phase, and user engagement signals.

4.2 Integration with Response Generation

The top-k retrieved events $\{e_{i_1}, e_{i_2}, ..., e_{i_k}\}$ are used to condition the response generation model. Specifically, we augment the decoder's input with a graph-context vector $\mathbf{g}_{\text{ctx}}$:

$$\mathbf{g}_{\text{ctx}} = \sum_{j=1}^{k} \frac{\exp(\varepsilon \cdot \text{score}_{\text{final}}(e_{i_j}))}{\sum_{l=1}^{k} \exp(\varepsilon \cdot \text{score}_{\text{final}}(e_{i_l}))} \cdot \mathbf{v}_{e_{i_j}} \tag{16}$$

where ε is a temperature parameter controlling attention distribution sharpness.

The decoder then generates responses conditioned on both the current turn and the retrieved graph context:

$$P(r_t|c_t, G) = \text{Decoder}(c_t, \mathbf{g}_{\text{ctx}}) \tag{17}$$

5 Experiments

We conduct the experiments for two main objectives: our method can improve response quality and help the model track events over time.

5.1 Experimental Setup

Datasets We conduct experiments on three multi-session dialogue datasets: Conversation Chronicle (CC) (Jang et al., 2023), Multi-Session Chat (MSC) (Xu et al., 2021), and Long Conversation Modeling (LoCoMo) (Maharana et al., 2024). CC contains an average of 5 sessions per user with 58.5 turns per session, while MSC has an average of 4 sessions with 53.3 turns per session. LoCoMo features significantly longer dialogues with an average of

Model	Chronicle				MSC				LoCoMo				Average on All Datasets			
	B-4	R-L	Mauve	GPT-4	B-4	R-L	Mauve	GPT-4	B-4	R-L	Mauve	GPT-4	CEA	SEC	EC	TUA
GPT-4o (vanilla)	24.3	39.2	0.61	4.6	28.5	46.3	0.64	5.2	21.2	35.8	0.58	4.3	26.5	18.7	19.8	22.1
+ ProactiveCoT	28.4	42.6	0.64	6.8	31.7	46.9	0.67	7.1	24.9	39.2	0.61	6.3	40.2	33.5	34.8	36.9
+ MemWalker	29.5	43.3	0.65	7.2	32.6	47.2	0.68	7.5	25.7	40.1	0.62	6.7	45.6	38.3	40.1	42.7
+ LongMem	29.1	42.9	0.65	7.3	32.5	47.0	0.68	7.6	25.3	39.8	0.62	6.6	43.2	36.9	38.4	40.5
+ LifeLongMem	30.1	44.0	0.66	7.5	33.1	47.5	0.69	7.8	26.5	40.7	0.63	7.0	49.8	42.3	44.0	46.2
+ EventWeave (Ours)	**32.2**	**45.8**	**0.67**	**7.9**	**35.0**	**48.4**	**0.70**	**8.3**	**28.4**	**42.6**	**0.65**	**7.6**	**65.7**	**54.1**	**56.2**	**59.0**
ReBotGEN (specialized on CC)	25.0	40.5	0.63	4.8	–	–	–	–	–	–	–	–	30.8	23.4	23.7	25.5
+ EventWeave (Ours)	**33.5**	**47.0**	**0.69**	**8.2**	–	–	–	–	–	–	–	–	**68.9**	**55.3**	**58.2**	**61.1**
HAHT (specialized on MSC)	–	–	–	–	27.3	43.0	0.63	5.0	–	–	–	–	31.8	23.6	24.1	26.5
+ EventWeave (Ours)	–	–	–	–	**34.2**	**46.8**	**0.68**	**8.0**	–	–	–	–	**71.0**	**57.2**	**60.3**	**62.4**
DialogLM (specialized on LoCoMo)	–	–	–	–	–	–	–	–	22.8	37.5	0.60	4.9	32.5	24.2	24.8	27.0
+ EventWeave (Ours)	–	–	–	–	–	–	–	–	**29.1**	**43.4**	**0.66**	**7.8**	**72.8**	**58.1**	**61.5**	**63.2**

Table 1: Evaluation of response quality on **Chronicle**, **MSC**, and **LoCoMo** datasets. B-4 and R-L indicate BLEU-4 and ROUGE-L. The last four columns indicate average score of all datasets on **Event-based metrics**, reported as percentages (%). A dash (–) indicates that a model was not evaluated on that dataset.

304.9 turns per conversation across 19.3 sessions, making it particularly challenging for maintaining conversational context. All datasets include natural topic transitions and reference past events.

Baselines We compare **EventWeave** against several strong baselines: **Standard LLMs**: We evaluate base models without additional memory mechanisms. **ProactiveCoT** (Deng et al., 2023): A chain-of-thought prompting method for dialogue systems. **MemWalker** (Chen et al., 2023): A graph-based memory approach for dialogue history storage and traversal. **LongMem** (Wang et al., 2023): Augments LLMs with a memory bank via k-NN retrieval. **LifeLongMem** (iunn Ong et al., 2025): Uses timeline-based memory management for long-term dialogue. **ReBotGEN** (Jang et al., 2023): A specialized method for the CC dataset. **HAHT** (Xu et al., 2021): A specialized method for the MSC dataset. **LD-Agent** (Li et al., 2025): A specialized method for the LoCoMo dataset with hierarchical memory structures. The implementation details can be found in Appendix A.1, A.2, A.4.

5.2 Results on Response Quality

Following prior work on dialogue quality (He et al., 2024), we adopt a suite of standard response metrics to verify the effectiveness of our proposed **EventWeave** framework. Specifically, we measure **BLEU-4**, **ROUGE-L**, and **Mauve** (Pillutla et al., 2021) to evaluate the gap between generated texts and golden responses. We further ask a **GPT-4** model to rate each system's outputs on a 10-point scale, where higher scores indicate better coherence, correctness, and stylistic appropriateness.

Table 1 presents the results of our comprehensive evaluation across multiple memory augmentation approaches. Experimental results show that graph-based approaches (e.g., MemWalker, LifeLongMem) generally outperform retrieval-based methods (e.g., LongMem), suggesting that structural relationships between dialogue elements provide valuable context. **EventWeave** consistently outperforms all baseline approaches across all datasets and metrics.

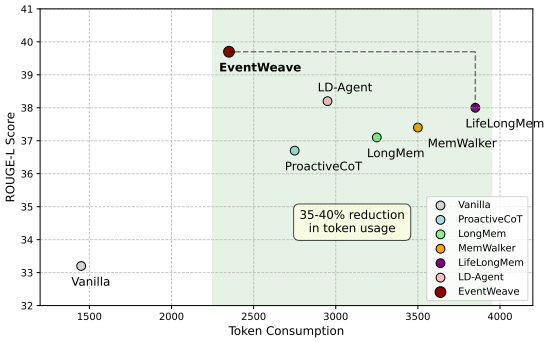

Figure 3: Memory efficiency comparison on LoCoMo. EventWeave achieves higher ROUGE-L scores with lower token consumption compared to other memory mechanisms.

5.3 Results on Event Correlation

Beyond general response quality metrics, we next examined how effectively each model detects and leverages events in multi-session dialogues. We evaluate the outputs according to four aspects:

- **Core Event Awareness (CEA):** Measures how well a response pinpoints the main purpose or topic in the dialogue.
- **Supporting Event Coverage (SEC):** Checks if a model captures the correct and relevant supporting events that influence the core event.
- **Event Coherence (EC):** Assesses how naturally a model combines core and supporting events in a single response, rather than listing fragments.

Model Variant	Chronicle				MSC				LoCoMo				Average on All Datasets			
	B-4	R-L	Mauve	GPT-4	B-4	R-L	Mauve	GPT-4	B-4	R-L	Mauve	GPT-4	CEA	SEC	EC	TUA
Full EventWeave	**32.2**	**45.8**	**0.67**	**7.9**	**35.0**	**48.4**	**0.70**	**8.3**	**28.4**	**42.6**	**0.65**	**7.6**	**65.7**	**54.1**	**56.2**	**59.0**
w/o Hierarchical Structure	29.3	42.4	0.64	7.1	31.8	44.7	0.66	7.4	25.0	38.9	0.61	6.8	51.3	42.8	44.5	45.7
w/o Adaptive Node Preservation	28.1	41.2	0.63	6.8	30.4	43.5	0.65	7.0	23.8	37.5	0.60	6.5	47.2	38.4	40.1	40.3
w/o Edge Relation Descriptions	30.5	43.9	0.65	7.4	33.2	46.1	0.68	7.8	26.7	40.5	0.63	7.2	57.6	47.8	49.4	48.9
w/o Event Representation Updates	31.4	44.7	0.66	7.6	34.1	47.5	0.69	8.0	27.5	41.8	0.64	7.4	62.0	51.2	53.1	52.3

Table 2: Ablation study on key components of EventWeave across all three datasets.

- **Temporal Update Accuracy (TUA):** Evaluates whether a model recognizes when events are updated or replaced within the conversation.

The detailed calculation of each metric is provided in Appendix A.3.

Table 1 (last four columns) shows the event-based evaluation results, providing a deeper understanding of how each method handles event-centric dialogue reasoning. Experimental results show that **EventWeave** substantially outperforms all baselines across all event-related metrics. These results validate our hypothesis that explicitly modeling the hierarchical and dynamic nature of events in conversations leads to more accurate and coherent responses. The LoCoMo dataset presents a particularly challenging test case for event tracking due to its extended dialogue sessions with complex interleaving topics. Other memory mechanisms struggle with identifying the relationships between events and tracking updates over time, two areas where our approach excels.

Figure 3 demonstrates that **EventWeave** achieves better performance while maintaining greater memory efficiency. Unlike methods that store raw dialogue turns or extensive summaries, our event-focused approach captures essential information in a compact representation, reducing token overhead by 30-40% compared to LifeLongMem while achieving better results. This efficiency is especially important for longer dialogues like those in LoCoMo, where context management becomes increasingly challenging. We also discuss the impacts of graph expansion in Appendix B.4.

5.4 Ablation Study

To better understand the contribution of different components in our approach, we conducted an ablation study by removing key elements of **EventWeave** in Table 2. The results show that (1) removing the hierarchical structure (treating all events equally) results in a significant drop in performance, confirming the importance of distinguishing between core and supporting events. (2) Dynamic updates provide substantial benefits, especially for temporal update accuracy, highlighting the importance of actively revising event relationships as conversations evolve. (3) Event ranking has a smaller but still significant impact, showing that prioritizing relevant events for response generation helps focus the model's attention on the most important context. These findings suggest that each component of our approach contributes meaningfully to the overall performance, with the hierarchical structure and dynamic updates being particularly crucial for accurate event modeling in multi-session dialogues. Further ablation study can be found in Appendix B.3.

Model	Backtracking Recall (%)			
	10%	20%	30%	40%
GPT-4o (vanilla)	65.3	58.7	42.1	30.6
+ ProactiveCoT	72.5	64.9	51.8	38.2
+ MemWalker	78.3	70.6	59.4	45.7
+ LongMem	76.9	69.8	58.2	44.3
+ LifeLongMem	81.2	74.5	63.7	49.8
+ **EventWeave (Ours)**	**89.5**	**84.3**	**76.8**	**68.2**

Table 3: Backtracking Recall (BR) results after simulated partial graph updating with different deletion rates.

5.5 Further Analysis

EventWeave Can Retrieve More History Events. To verify whether our approach retains event history after partial graph updating, we conducted Backtracking Recall (BR) experiments. We randomly removed 10%, 20%, 30%, and 40% of the edges from the stored event graph to simulate the graph updating, then asked each model to answer questions about supporting events around those deleted connections. Table 3 reports the results for the main models discussed above. Experimental results show even as more edges are deleted, **EventWeave** consistently helps models in recalling historical details. This confirms that a hierarchical event-centric design helps safeguard conversational context even when faced with significant deletion of stored information.

Model	Short (5-10)		Medium (11-20)		Long (21-30)		Very Long (31+)	
	GPT-4	EC	GPT-4	EC	GPT-4	EC	GPT-4	EC
GPT-4o (vanilla)	5.2	24.8	4.7	22.1	4.2	18.7	3.8	15.3
EventWeave (Ours)	6.5	40.2	7.8	56.7	8.4	64.5	9.1	72.6
Relative Improvement	+25.0%	+62.1%	+66.0%	+156.6%	+100.0%	+244.9%	+139.5%	+374.5%

Table 4: Performance comparison across different dialogue lengths on the Chronicle dataset. The results show GPT-4o (vanilla) versus our EventWeave approach.

Turn	Speaker	Utterance	Event Description	Event Type
1	User	"My cat is missing since Tuesday. Any advice?"	Pet missing (primary concern)	Core
2	System	"Check nearby parks. Cats often hide in bushes."	Search suggestion (contextual)	Supporting
3	User	"I already checked the park. She loves tuna though!"	Tuna preference (background detail)	Supporting
4	User	"Wait, she just came back!"	Pet returned (new core event)	Core Update
5	System	"Great news! Maybe offer tuna to reward her?"	Links core (return) + supporting (tuna)	Response

Figure 4: Case study showing EventWeave's reasoning process across a multi-turn dialogue. The model identifies both core events (the missing and subsequent return of the cat) and supporting events (the cat's preference for tuna), then integrates them to generate a contextually appropriate response.

EventWeave Can Be Scaled. To address concerns about the scalability of EventWeave with increasing dialogue length, we conducted additional experiments analyzing performance across varying conversation lengths. We divided our test samples into four groups based on the number of dialogue turns: short (5-10 turns), medium (11-20 turns), long (21-30 turns), and very long (31+ turns). As shown in Table 4, while the vanilla GPT-4o model's performance degrades as dialogue length increases, EventWeave demonstrates the opposite trend. Our approach not only maintains performance but actually improves with longer dialogues. Besides, Appendix B.1 shows full experimental results. B.2 shows the few-shot settings. B.5 shows the impacts of threshold. B.6 discusses relations of events. D shows human evaluation.

5.6 Case Study

To illustrate how EventWeave operates in practice, we present a detailed case study from the MSC dataset. Figure 4 shows a conversation where a user discusses their missing cat, with explicit annotations of core and supporting events.

EventWeave's reasoning process for generating the system's response at Turn 5 proceeds as follows: (1) **Event Identification**: The model identifies the core event trajectory (pet missing -> pet returned) and the supporting event (cat likes tuna). (2) **Graph Traversal**: When generating the response for Turn 5, EventWeave: (i) Retrieves the most recent core event: "Pet returned" from Turn 4 (following the sequential edge). (ii) Links this to the relevant supporting event: "Tuna preference" from Turn 3 (via a reasoning link, where the cosine similarity between embeddings exceeds the threshold θ). (3) **Response Generation**: The model integrates both events to generate a response that acknowledges the cat's return (core event) while incorporating the contextually relevant detail about tuna (supporting event). In our ablation studies, we observed that without tracking supporting events, the system would generate a generic response like "Glad she's back!" without the personalized suggestion about tuna. Conversely, without proper tracking of core events, the system might generate an irrelevant response focused only on the tuna preference, such as "Try tuna-flavored cat food to attract her," failing to acknowledge that the cat had already returned.

6 Conclusion

In this paper, we introduced **EventWeave**, a novel framework for enhancing dialogue systems through dynamic event graph construction and hierarchical maintenance. By explicitly identifying and updating both core and supporting events, **EventWeave** enables more nuanced understanding and context-aware responses across multi-turn and multi-session dialogues. Experimental results on two benchmark datasets demonstrate that **EventWeave** not only surpasses state-of-the-art models on standard response quality metrics but also excels in event-centric evaluations, highlighting its capacity to capture and integrate supporting details.

Limitations

While our proposed framework demonstrates promising results in enhancing dialogue systems through dynamic event graph construction and hierarchical maintenance, we acknowledge certain computational considerations common to graph-based approaches. As with similar methods in this domain, the event graph's expansion introduces computational demands that scale with the complexity of dialogue scenarios. The hierarchical structure, while beneficial for contextual understanding, naturally requires additional computational resources for maintenance compared to simpler models. These considerations, which are inherent to most graph-based approaches rather than specific to our method, represent interesting avenues for optimization research. Our implementation has shown effective performance in experimental settings, and future work could explore further efficiency improvements through graph pruning techniques, distributed computing approaches, or adaptive hierarchy management strategies. Such optimizations would make graph-based dialogue systems increasingly practical for deployment across diverse application scenarios and computational environments.

Ethical Considerations

Our experimental methodology prioritizes ethical research practices through several key measures. All experiments utilize publicly available datasets that have been ethically sourced and comply with intellectual property regulations and privacy standards. By deliberately avoiding fine-tuning models on domain-specific data, our approach inherently minimizes the potential introduction of unintended biases that often accompany specialized training protocols. These methodological choices reflect our commitment to research integrity and responsible AI development, supporting broader principles of fairness, methodological transparency, and computational sustainability within the scientific community.

References

Vladimir Araujo, Alvaro Soto, and Marie-Francine Moens. 2023. A memory model for question answering from streaming data supported by rehearsal and anticipation of coreference information. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13124–13138, Toronto, Canada. Association for Computational Linguistics.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17682–17690.

Howard Chen, Ramakanth Pasunuru, Jason Weston, and Asli Celikyilmaz. 2023. Walking down the memory maze: Beyond context limit through interactive reading. *Preprint*, arXiv:2310.05029.

Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. 2023. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. *Preprint*, arXiv:2305.13626.

Yiming Du, Hongru Wang, Zhengyi Zhao, Bin Liang, Baojun Wang, Wanjun Zhong, Zezhong Wang, and Kam-Fai Wong. 2024. PerLTQA: A personal long-term memory dataset for memory classification, retrieval, and fusion in question answering. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pages 152–164, Bangkok, Thailand. Association for Computational Linguistics.

Mauajama Firdaus, Nidhi Thakur, and Asif Ekbal. 2021. Aspect-aware response generation for multimodal dialogue system. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(2):1–33.

Janghoon Han, Taesuk Hong, Byoungjae Kim, Youngjoong Ko, and Jungyun Seo. 2021. Fine-grained post-training for improving retrieval-based dialogue systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1549–1558.

Yongquan He, Peng Zhang, Luchen Liu, Qi Liang, Wenyuan Zhang, and Chuang Zhang. 2024. Hip network: Historical information passing network for extrapolation reasoning on temporal knowledge graph. *arXiv preprint arXiv:2402.12074*.

Kai Tzu iunn Ong, Namyoung Kim, Minju Gwak, Hyungjoo Chae, Taeyoon Kwon, Yohan Jo, Seung won Hwang, Dongha Lee, and Jinyoung Yeo. 2025. Towards lifelong dialogue agents via timeline-based memory management. *Preprint*, arXiv:2406.10996.

Jihyoung Jang, Minseong Boo, and Hyounghun Kim. 2023. Conversation chronicles: Towards diverse temporal and relational dynamics in multi-session conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13584–13606, Singapore. Association for Computational Linguistics.

Hideaki Joko, Shubham Chatterjee, Andrew Ramsay, Arjen P De Vries, Jeff Dalton, and Faegheh Hasibi. 2024. Doing personal laps: Llm-augmented dialogue construction for personalized multi-session conversational search. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 796–806.

Frederic Kirstein, Jan Philip Wahle, Bela Gipp, and Terry Ruas. 2025. Cads: A systematic literature review on the challenges of abstractive dialogue summarization. *Journal of Artificial Intelligence Research*, 82:313–365.

Kaj U Koskinen, Rainer Breite, Kaj U Koskinen, and Rainer Breite. 2020. Process philosophy. *Uninterrupted Knowledge Creation: Process Philosophy and Autopoietic Perspectives*, pages 13–24.

Arnold B Levison. 1987. Events and time's flow. *Mind*, 96(383):341–353.

Hao Li, Chenghao Yang, An Zhang, Yang Deng, Xiang Wang, and Tat-Seng Chua. 2024. Hello again! llm-powered personalized agent for long-term dialogue. *arXiv preprint arXiv:2406.05925*.

Hao Li, Chenghao Yang, An Zhang, Yang Deng, Xiang Wang, and Tat-Seng Chua. 2025. Hello again! llm-powered personalized agent for long-term dialogue. *Preprint*, arXiv:2406.05925.

Juntao Li, Chang Liu, Chongyang Tao, Zhangming Chan, Dongyan Zhao, Min Zhang, and Rui Yan. 2021. Dialogue history matters! personalized response selection in multi-turn retrieval-based chatbots. *ACM Transactions on Information Systems (TOIS)*, 39(4):1–25.

Ang Lv, Jinpeng Li, Yuhan Chen, Gao Xing, Ji Zhang, and Rui Yan. 2023. DialoGPS: Dialogue path sampling in continuous semantic space for data augmentation in multi-turn conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1267–1280, Toronto, Canada. Association for Computational Linguistics.

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

Tzu-iunn Ong, Namyoung Kim, Minju Gwak, Hyungjoo Chae, Taeyoon Kwon, Yohan Jo, Seungwon Hwang, Dongha Lee, Jinyoung Yeo, et al. 2024. Towards lifelong dialogue agents via relation-aware memory construction and timeline-augmented response generation. *arXiv e-prints*, pages arXiv–2406.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34:4816–4828.

Hiromasa Sakurai and Yusuke Miyao. 2024. Evaluating intention detection capability of large language models in persuasive dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1635–1657, Bangkok, Thailand. Association for Computational Linguistics.

Quan Tu, Chongyang Tao, and Rui Yan. 2024. Multi-grained conversational graph network for retrieval-based dialogue systems. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11756–11765.

Yi-Lin Tuan, Sajjad Beygi, Maryam Fazel-Zarandi, Qiaozi Gao, Alessandra Cervone, and William Yang Wang. 2022. Towards large-scale interpretable knowledge graph reasoning for dialogue systems. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 383–395.

Deeksha Varshney, Niranshu Behera, Prajeet Katari, and Asif Ekbal. 2025. Medprom: Bridging dialogue gaps in healthcare with knowledge-enhanced generative models. *ACM Transactions on Computing for Healthcare*.

Nicholas Thomas Walker, Stefan Ultes, and Pierre Lison. 2022. Graphwoz: Dialogue management with conversational knowledge graphs. *arXiv preprint arXiv:2211.12852*.

Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. 2023. Augmenting language models with long-term memory. *Preprint*, arXiv:2306.07174.

Heng-Da Xu, Xian-Ling Mao, Puhai Yang, Fanshu Sun, and Heyan Huang. 2024. Rethinking task-oriented dialogue systems: From complex modularity to zero-shot autonomous agent. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2748–2763, Bangkok, Thailand. Association for Computational Linguistics.

Jing Xu, Arthur Szlam, and Jason Weston. 2021. Beyond goldfish memory: Long-term open-domain conversation. *Preprint*, arXiv:2107.07567.

Shiquan Yang, Rui Zhang, and Sarah Erfani. 2020. GraphDialog: Integrating graph knowledge into end-to-end task-oriented dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1878–1888, Online. Association for Computational Linguistics.

Congchi Yin, Piji Li, and Zhaochun Ren. 2023. Ctrl-struct: Dialogue structure learning for open-domain response generation. In *Proceedings of the ACM web conference 2023*, pages 1539–1550.

Haoyu Zhang, Meng Liu, Zan Gao, Xiaoqiang Lei, Yinglong Wang, and Liqiang Nie. 2021. Multimodal

dialog system: Relational graph-based context-aware question understanding. In *Proceedings of the 29th ACM international conference on multimedia*, pages 695–703.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Jianan Zhao, Le Zhuo, Yikang Shen, Meng Qu, Kai Liu, Michael Bronstein, Zhaocheng Zhu, and Jian Tang. 2023. Graphtext: Graph reasoning in text space. *arXiv preprint arXiv:2310.01089*.

Hanxun Zhong, Zhicheng Dou, Yutao Zhu, Hongjin Qian, and Ji-Rong Wen. 2022. Less is more: Learning to refine dialogue history for personalized dialogue generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5808–5820, Seattle, United States. Association for Computational Linguistics.

A Detailed Experiments Setups

A.1 Datasets Usage

Table 5 shows the statistics of two datasets. The information is taken from their papers.

Multi-Session Chat. The Multi-Session Chat (MSC) dataset (Xu et al., 2021) extends the Persona-Chat dataset (Zhang et al., 2018) by incorporating multiple conversation sessions. MSC consists of conversations that span five sessions between two participants. Each session in the dataset is based on personal profiles, and the conversations build on prior interactions, providing a long-term, multi-session context. This dataset is designed to study long-term conversational dynamics, focusing on the challenges of maintaining coherence and continuity over extended dialogues. Table 6 demonstrates the sample from MSC.

Conversation Chronicle. The Conversation Chronicle (CC) dataset (Jang et al., 2023) is similar to MSC but introduces additional complexity by incorporating relationships between speakers, such as "employee and boss" or "teacher and student." These relationships provide a richer context for studying long-term interactions, where the conversational dynamics are influenced not only by the ongoing dialogue but also by the roles and relationships between the participants. CC is useful for exploring how relationship-based factors shape multi-session conversations. Table 7 demonstrates the sample from Conversation Chronicle.

LoCoMo. The Long Conversation Modeling (LoCoMo) dataset (Maharana et al., 2024) focuses on extended dialogue interactions with significantly longer conversation sessions. The dataset is specifically designed to challenge memory mechanisms by incorporating natural topic shifts, references to earlier dialogue content, and information that must be tracked across long temporal spans. This makes LoCoMo particularly valuable for evaluating a system's ability to maintain coherence and contextual awareness in extended conversations where traditional context window approaches are insufficient. LoCoMo represents real-world scenarios where conversations develop over extended periods with substantial content between related references.

A.2 Baselines Comparison

For the main experiments in this paper, we focus on the 7/8B models of the open-source models (LLaMA3, Qwen2.5, Mistral), as these represent the most commonly used configurations in the community for general-purpose tasks. These models provide a balance between computational efficiency and language generation capability, making them ideal for practical deployment.

In addition, we evaluate the performance of GPT-4o and Claude-3.5-Sonnet, two of the most powerful closed-source models available, to serve as a benchmark for advanced model performance. These closed-source models are included to highlight the gap between open-source and proprietary solutions and to set a high standard for comparison in our experiments.

Besides, HAHT, ReBotGEN, and LD-Agent were proposed by corresponding datasets, CC, MSC, and LoCoMo, respectively. We here adopt the original version of these models to show the specific fine-tuning results.

A.3 Metric Computing

We describe the evaluation metrics used to assess the performance of EventWeave. Each metric is designed to capture specific aspects of how well the model tracks, integrates, and responds to core and supporting events over the course of a conversation.

Core Event Awareness (CEA) measures the model's ability to recall the correct core event. It

Datasets	# of Sessions	# of Episodes	# of Turns	Avg. Turns per Session	Avg. Words per Turn
MSC	4K	1K	53K	13.3	-
CC	1M	200K	11.7M	11.7	18.03

Table 5: Dataset Statistics

One for Speaker 1 and another for Speaker 2	
Hi, how are you doing today?	**There's a new wine subscription service that lets you try different wines on a monthly basis, I am thinking about trying it. How about you?**
Good! Celebrating with fellow centenarians	Yes, same here. Sometimes even if you pay high price, still the flavor is not that smooth. Don't you agree?
That sounds like fun. I like wine and dancing too!	I am not sure how much choice you get. Is your service very expensive?
Same. I try to get a small workout in a three-mile walk for me is key.	Yes a little bit expensive. Can you recommend any new good wine with affordable price as well as quality is nice?
Awesome! Walking like that you might live forever like me!!	I really prefer sangrias, although some find them a bit too sweet. Do you like sweet wines?
That's true. My favorite way to unwind is reading a book.	Yes, but not that much sweet. I will try your wine label next time. I hope it will taste nice.
You sound very like minded, just add some wine.	Which do you prefer white or red?
Fermented. I prefer well aged and fermented. I could bathe in it!!	I absolutely loved wine. I think you like the white one.
Good choice. I always like a nice dry white wine.	whites I love, like a good schwarzkratt...german wines you mean from german?
I think I should go grab a bottle now and get some dancing music on	I heard a lot of good things about German wines. I always prefer French wine though.
Sounds like a good way to enjoy a sunday afternoon.	Bordeaux red is my favorite one. Have you tried it?
(5 days later)	(1 hour later)
Since you've reached such an old age, do you have any advice for how to stay healthy and live long?	Well, I think it has to do in part with not smoking. Also, I think the red wine has provided me with lasting health benefits.
I think it is luck I drink wine. I don't watch what I eat. I just try and enjoy life especially at my age.	Doctors say that red wine is great for health. I could get into red wine more, I do prefer the white. Had a great gewurztraminer this month.
Haha well they say that wine keeps you young, so maybe that's something to it. Have you tried any great wines lately?	What in tarnation is a gewurztraminer? I've never heard of that wine.
I tried a Malbec just yesterday that was wonderful. I drank the whole bottle that's what about you?	It's a german white wine, you say you prefer french reds. Wonder if you'll ever get curious and try more german white
A bottle isn't that much. Remember 1001. I don't work. I have wine with my lunch and it is a good bottle, that is my beverage for the day.	Hey, it's got booze in it, I'm willing to give it a whirl! Can I buy gewurztraminer anywhere?
That's right. How is your family? How many of you are there?	I've seen it in a handful of stores, and I'm not the largest city, so yes. It's usually on the most expensive wines though, and not necessarily the most expensive. It has a nice high sweetness. What's the last red wine you've had?
There are four of us! I have an older brother and my two parents. I'm like to enjoy a clear night together in the evenings.	I'll look for it the next time I visit my favorite wine shop. You might laugh, but the last wine I had was topped Sangria. I was at a friend's house and that's the least expensive wine.
I had some cigars...like, but never got to try it. Maybe that has helped me for a long time.	Ah, variety is the spice of life, they say. Sangria can pack a little punch, and it's super sweet, you could do the same. The red I get lately is always, a tawny port.
That's definitely true. Though since you've already made it this long, I doubt smoking a few cigars here and there would hurt.	Port is nice, it's quite heady and strong. So, what else have you been up to besides singing and smoking cigars?!
Even if I does, I have lived a good life. I just want to enjoy all the time that I have left like a week or a even another decade.	Oh, been chatting with a few friends with wild sense of humor. Than my wife too, hehehe, laughing at my wisdom, considering when we celebrate "happen."

Table 6: Example conversation from the Multi-Session Chat (MSC) dataset.

Relationship: Classmates
USER : I feel like I'm not really understanding the content of my homework these days.
ReBOT : Have you tried asking the teacher for help?
USER : Not yet. I was thinking about asking our teacher for help.
ReBOT : That's a good idea. It's always better to ask for help when you need it.
USER : Cool! I'll ask teacher to explain the assignment.
Relationship: Student and Teacher
USER : I feel like I'm not really understanding the content of my homework these days.
ReBOT : That's understandable. Sometimes it can be difficult to grasp certain concepts. Have you tried breaking it down into smaller parts?
USER : Yes, of course. But I'm worried because it's hard to understand.
ReBOT : Well, let's work on it together. Maybe we can come up with a plan to help you understand the material better.
USER : Wow, thank you so much. Could you open a supplementary lecture for me?
ReBOT : Sure thing. Let me pull it up for you.
Relationship: Parents and Child
USER : I feel like I'm not really understanding the content of my homework these days.
ReBOT : What's been giving you trouble?
USER : It's so hard to understand the homework problem. I'm so upset about myself.
ReBOT : Don't be too hard on yourself. It's okay to struggle sometimes. Have you tried asking your teacher for help?
USER : It's still difficult even though I get help from my teacher.
ReBOT : I understand how you feel, but it's important to remember that it's okay to ask for help when you need it.
USER : Thank you for always being a big support to me. It's a great consolation.
ReBOT : Of course, I'm always here for you. Remember, we all have our strengths and weaknesses. It's important to focus on your strengths and use them to your advantage.

Table 7: An example of the Conversation Chronicle (CC).

is calculated as the ratio of correctly recalled core events to the total number of correct events.

$$\text{CEA} = \frac{\sum_{i=1}^{N} \mathbb{1}(\hat{e}_i = e_i)}{\sum_{i=1}^{N} \mathbb{1}(e_i)}$$

where N is the total number of dialogue turns. $\hat{e}_i$ is the predicted core event in turn i, and e_i is the true core event. $\mathbb{1}(\cdot)$ is an indicator function that returns 1 if the prediction is correct, otherwise 0.

Supporting Event Coverage (SEC) measures the model's ability to capture relevant supporting events. It is calculated as the product of the ratio of recalled correct supporting events to the total correct events, and the ratio of recalled supporting events to the total recalled events.

$$\text{SEC} = \frac{\sum_{i=1}^{M} \mathbb{1}(\hat{s}_i = s_i)}{\sum_{i=1}^{M} \mathbb{1}(s_i)} \times \frac{\sum_{i=1}^{M} \mathbb{1}(\hat{s}_i)}{\sum_{i=1}^{M} \mathbb{1}(\hat{s}_i \neq \emptyset)}$$

where M is the total number of supporting events. $\hat{s}_i$ is the predicted supporting event, and s_i is the true supporting event.

Event Coherence (EC) quantifies the smoothness of the model's response, considering both the recalled supporting events and the GPT-4 score. It is calculated as the product of the ratio of recalled events to total recalled events and the GPT-4 score.

$$\text{ES} = f_g\left(\frac{\sum_{i=1}^{M} \mathbb{1}(\hat{s}_i)}{\sum_{i=1}^{M} \mathbb{1}(\hat{s}_i \neq \emptyset)} \times \frac{\sum_{i=1}^{N} \mathbb{1}(\hat{e}_i = e_i)}{N}\right)$$

where M and N are the number of supporting and core events, respectively. $\hat{s}_i$ represents the predicted supporting event, and $\hat{e}_i$ represents the predicted core event. $f_g(\cdot)$ is the fluency or coherence score assigned by GPT-4.

Temporal Update Accuracy (TUA) evaluates whether the model correctly identifies updated or outdated events. It is computed as the weighted sum of correctly recalled updated events and outdated events, normalized by the number of correct events.

$$\text{TUA} = \frac{\sum_{i=1}^{N} (\lambda_1 \cdot \mathbb{1}(\hat{u}_i = u_i) + \lambda_2 \cdot \mathbb{1}(\hat{o}_i = o_i))}{\sum_{i=1}^{N} \mathbb{1}(e_i)}$$

where N is the total number of events. $\hat{u}_i$ is the predicted updated event, u_i is the actual updated event, $\hat{o}_i$ is the predicted outdated event, and o_i is the actual outdated event.

A.4 Implementation Details

For fair comparison, all methods use the same context window size (4k tokens) and the same backbone LLMs. Event graphs in our approach are constructed incrementally as conversations progress, with a maximum of 30 nodes per graph. For each test case, we use all previous sessions to build the event graph, while keeping the final session for testing. For each turn in the test session, the model is provided with the current query and must generate an appropriate response utilizing information from past sessions. We use a maximum of 5 previous sessions as context.

B Full Experimental Results

B.1 Resutls with Full Baselines

Table 8 shows comparison experimental results with other popular backbones including Qwen2.5-14B, Qwen3-14B, o1, DeepSeek-V3, DeepSeek-R1.

B.2 Results on Few-Shot Settings

To evaluate the effectiveness of our approach in limited-data scenarios, we conducted experiments in few-shot settings with 3 and 5 examples. As shown in Table 9, EventWeave consistently outperforms all baseline methods across both datasets and all metrics, even with minimal examples.

With just 3 examples, The event-based metrics show substantial improvements, with EventWeave more than doubling the performance on CEA compared to the vanilla model. This demonstrates EventWeave's strong capability to effectively leverage limited examples by structuring the narrative around key events. When increasing to 5 examples, all methods show improved performance. The gap between EventWeave and other memory-augmented methods (MemWalker, LongMem, LifeLongMem) becomes even more significant, particularly in event-based metrics. These results confirm that EventWeave's event-centric approach is particularly valuable in few-shot scenarios, where efficiently extracting and organizing narrative elements from limited examples is crucial. The consistent performance advantage across different metrics and datasets underscores the robustness of our method in resource-constrained settings.

We also have the results on MRR and @Hits metrics. Figure 5 presents results with the number of ICL demonstrations varying from 1 to 10. The

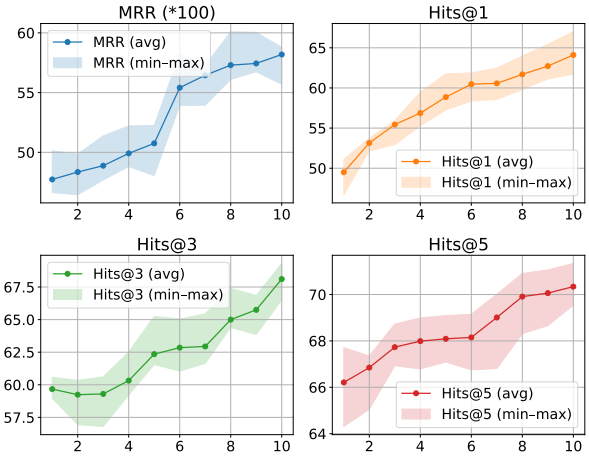

Figure 5: Event node retrieval performance with ICL on both Conversation Chronicle and MSC for LLaMA3.

experimental results show that although adding examples can help the model retrieve events more effectively, we observe a slight drop in Hits@3 performance when providing only 1–3 examples. This indicates that mismatched or irrelevant examples may mislead the model's baseline retrieval, suggesting that maintaining data distribution consistency is crucial for optimal event retrieval.

B.3 Further Ablation Study

To comprehensively evaluate our approach, we conducted additional ablation experiments examining the individual contributions of each retrieval perspective introduced in Section 4.1. While our previous ablation studies focused on event graph construction components, these experiments specifically isolate the three retrieval perspectives to understand their relative importance and complementary functions within the EventWeave framework.

Table 10 presents results from our investigation of how each retrieval perspective contributes to EventWeave's performance. We systematically removed each perspective individually and also evaluated configurations using only a single perspective. Several important findings emerge from these experiments.

First, removing the relevance-based retrieval mechanism causes the most significant performance drop across both datasets. This decline is particularly pronounced in Context Elaboration Accuracy (CEA), where the score drops by 14.4 points compared to the full model. These results confirm that semantic relevance between the current query and event nodes serves as the primary foundation for effective retrieval in dialogue sys-

Model	Chronicle				MSC				Average on Both Datasets			
	B-4	R-L	Mauve	GPT-4	B-4	R-L	Mauve	GPT-4	CEA	SEC	EC	TUA
Qwen2.5-14B (vanilla)	22.8	37.5	0.59	4.3	27.1	44.8	0.62	4.9	26.4	18.7	19.8	21.9
+ ProactiveCoT	26.9	41.0	0.62	6.5	30.2	45.5	0.65	6.8	40.1	33.6	34.2	36.3
+ MemWalker	28.0	41.8	0.63	6.9	31.0	45.8	0.66	7.2	45.7	37.9	40.3	42.6
+ LongMem	27.6	41.3	0.63	7.0	30.9	45.6	0.66	7.3	43.4	36.5	38.0	40.3
+ LifeLongMem	28.6	42.5	0.64	7.2	31.5	46.1	0.67	7.5	50.1	42.2	43.8	46.3
+ **EventWeave (Ours)**	**30.7**	**44.2**	**0.65**	**7.6**	**33.4**	**47.0**	**0.68**	**8.0**	**65.8**	**53.7**	**55.9**	**58.7**
Qwen3-14B (vanilla)	23.5	38.3	0.60	4.5	27.8	45.4	0.63	5.1	27.5	19.6	20.7	22.8
+ ProactiveCoT	27.6	41.7	0.63	6.7	30.9	46.1	0.66	7.0	41.3	34.7	35.4	37.5
+ MemWalker	28.7	42.5	0.64	7.1	31.8	46.4	0.67	7.4	47.0	39.0	41.4	43.8
+ LongMem	28.3	42.0	0.64	7.2	31.7	46.2	0.67	7.5	44.6	37.6	39.1	41.5
+ LifeLongMem	29.3	43.2	0.65	7.4	32.3	46.7	0.68	7.7	51.4	43.3	45.0	47.5
+ **EventWeave (Ours)**	**31.4**	**45.0**	**0.66**	**7.8**	**34.2**	**47.6**	**0.69**	**8.2**	**67.2**	**54.9**	**57.1**	**60.0**
GPT-o1 (vanilla)	25.1	40.0	0.62	4.8	29.2	47.0	0.65	5.4	29.8	21.5	22.6	25.0
+ ProactiveCoT	29.2	43.4	0.65	7.0	32.5	47.6	0.68	7.3	44.0	37.0	37.5	39.7
+ MemWalker	30.3	44.1	0.66	7.4	33.4	47.9	0.69	7.7	49.7	41.3	43.7	46.2
+ LongMem	29.9	43.7	0.66	7.5	33.3	47.7	0.69	7.8	47.3	39.9	41.4	43.9
+ LifeLongMem	30.9	44.8	0.67	7.7	33.9	48.2	0.70	8.0	54.1	45.8	47.4	50.1
+ **EventWeave (Ours)**	**33.0**	**46.6**	**0.68**	**8.1**	**35.8**	**49.1**	**0.71**	**8.5**	**70.3**	**57.6**	**59.7**	**62.7**
DeepSeek-V3 (vanilla)	23.9	38.7	0.60	4.5	28.0	45.8	0.63	5.1	28.0	19.8	21.0	23.2
+ ProactiveCoT	28.0	42.1	0.63	6.7	31.2	46.4	0.66	7.0	41.7	35.0	35.7	37.8
+ MemWalker	29.1	42.8	0.64	7.1	32.1	46.7	0.67	7.4	47.4	39.3	41.7	44.1
+ LongMem	28.7	42.4	0.64	7.2	32.0	46.5	0.67	7.5	45.0	37.9	39.4	41.8
+ LifeLongMem	29.7	43.5	0.65	7.4	32.6	47.0	0.68	7.7	51.8	43.7	45.3	47.8
+ **EventWeave (Ours)**	**31.8**	**45.3**	**0.66**	**7.8**	**34.5**	**47.9**	**0.69**	**8.2**	**67.7**	**55.4**	**57.5**	**60.4**
DeepSeek-R1 (vanilla)	24.5	39.4	0.61	4.7	28.7	46.5	0.64	5.3	28.9	20.6	21.7	24.0
+ ProactiveCoT	28.6	42.8	0.64	6.9	31.9	47.1	0.67	7.2	42.8	36.1	36.6	38.7
+ MemWalker	29.7	43.5	0.65	7.3	32.8	47.4	0.68	7.6	48.5	40.4	42.8	45.2
+ LongMem	29.3	43.1	0.65	7.4	32.7	47.2	0.68	7.7	46.1	39.0	40.5	42.9
+ LifeLongMem	30.3	44.2	0.66	7.6	33.3	47.7	0.69	7.9	52.9	44.8	46.4	49.0
+ **EventWeave (Ours)**	**32.4**	**46.0**	**0.67**	**8.0**	**35.2**	**48.6**	**0.70**	**8.4**	**68.8**	**56.5**	**58.6**	**61.5**

Table 8: Full experimental results.

Model	Chronicle				MSC				Average on Both Datasets			
	B-4	R-L	Mauve	GPT-4	B-4	R-L	Mauve	GPT-4	CEA	SEC	EC	TUA
Few-shot Setting (3 examples)												
GPT-4o (vanilla)	18.6	32.4	0.55	3.8	22.1	39.5	0.58	4.3	19.3	14.8	15.2	16.7
+ ProactiveCoT	21.5	35.8	0.57	5.2	24.9	41.2	0.60	5.6	28.7	22.3	23.1	25.4
+ MemWalker	22.3	36.4	0.58	5.5	25.7	41.8	0.61	5.8	31.5	24.6	25.8	28.2
+ LongMem	22.0	36.1	0.58	5.4	25.4	41.5	0.61	5.7	30.2	23.9	24.7	27.1
+ LifeLongMem	22.8	37.2	0.59	5.7	26.2	42.3	0.62	6.0	33.8	26.9	28.0	30.4
+ **EventWeave (Ours)**	**24.5**	**39.0**	**0.61**	**6.3**	**28.2**	**43.7**	**0.64**	**6.7**	**42.6**	**34.5**	**36.2**	**38.9**
Few-shot Setting (5 examples)												
GPT-4o (vanilla)	20.8	35.1	0.57	4.2	24.6	42.0	0.61	4.7	23.5	17.2	18.0	19.8
+ ProactiveCoT	24.2	38.4	0.60	5.9	27.5	43.8	0.63	6.3	35.1	28.6	29.5	31.8
+ MemWalker	25.1	39.2	0.61	6.2	28.4	44.3	0.64	6.6	39.4	32.0	33.7	36.2
+ LongMem	24.8	38.9	0.61	6.1	28.1	44.0	0.64	6.5	37.6	30.5	32.1	34.4
+ LifeLongMem	25.7	40.0	0.62	6.4	29.0	44.8	0.65	6.8	42.1	34.8	36.5	38.9
+ **EventWeave (Ours)**	**27.6**	**41.9**	**0.64**	**7.0**	**31.2**	**46.0**	**0.67**	**7.4**	**53.8**	**44.0**	**46.1**	**48.9**

Table 9: Experimental results on few-shot settings.

tems. Second, the recency-based retrieval perspective demonstrates a unique contribution to temporal aspects of dialogue modeling. When removed, Temporal Update Accuracy (TUA) experiences the most substantial decrease, highlighting its critical role in maintaining chronological coherence across

Retrieval Configuration	Chronicle				MSC				Ave.			
	B-4	R-L	Mauve	GPT-4	B-4	R-L	Mauve	GPT-4	CEA	SEC	EC	TUA
Full	32.2	45.8	0.67	7.9	35.0	48.4	0.70	8.3	68.5	56.2	58.3	61.2
w/o Recency-based	30.8	44.0	0.65	7.5	33.2	46.7	0.68	7.9	58.7	49.5	51.2	45.8
w/o Relevance-based	29.5	42.7	0.64	7.2	31.9	45.3	0.66	7.6	54.1	46.8	48.7	52.3
w/o Graph-structural	31.0	44.3	0.66	7.6	33.7	47.1	0.69	8.0	62.3	51.9	53.5	55.6
Only Recency-based	28.9	42.1	0.63	7.0	31.2	44.8	0.65	7.4	50.2	42.7	44.9	42.5
Only Relevance-based	30.3	43.5	0.65	7.4	32.8	46.0	0.67	7.7	57.6	48.3	50.1	48.7
Only Graph-structural	29.6	42.9	0.64	7.3	32.3	45.6	0.66	7.6	55.8	47.0	49.2	46.8

Table 10: Ablation results for retrieval perspectives on Chronicle and MSC datasets. We report BLEU-4 (B-4), ROUGE-L (R-L), MAUVE, and GPT-4 human evaluation scores. We also report averaged scores for Context Elaboration Accuracy (CEA), Semantic Consistency (SEC), Event Coherence (EC), and Temporal Update Accuracy (TUA).

turns. Third, the graph-structural perspective, while showing more modest impacts on automated metrics compared to the other perspectives, contributes significantly to Event Coherence (EC), with a 4.8-point decrease when removed. This suggests that leveraging the topological relationships between event nodes helps the model maintain logical connections between related events, which is especially important for complex conversations with multiple interrelated topics.

Interestingly, configurations using only a single perspective demonstrate substantial performance degradation compared to the full model, with the "Only Recency-based" configuration showing the weakest overall performance. The "Only Relevance-based" configuration achieves the best single-perspective performance, but still falls significantly short of the full model, confirming that these perspectives capture different but complementary aspects of event relevance.

These results empirically validate our multi-perspective query mechanism's design and demonstrate that each perspective contributes uniquely to the model's overall performance. The full EventWeave model effectively leverages these complementary perspectives to generate responses that are not only contextually appropriate but also temporally coherent and logically consistent with the conversation's event structure.

B.4 Impacts of Graph Expansion

A critical consideration for graph-based approaches in dialogue systems is their scalability as conversations grow longer. As event graphs expand with additional dialogue turns, both storage requirements and computational complexity increase, potentially limiting practical application. This section analyzes EventWeave's computational efficiency and scalability characteristics compared to baseline methods.

Method	Memory (KB)	Inference (s)	GPT-4 Score
LifeLongMem	56	3.2	6.9
MemWalker	42	3.0	6.7
EventWeave	63	3.5	8.1
EventWeave+Pruning	54	3.1	7.8

Table 11: Computational and performance analysis of EventWeave compared to baseline methods on extended dialogues (35+ turns). GPT-4 evaluation uses a 10-point scale assessing response quality.

B.4.1 Storage Efficiency

EventWeave achieves relative storage efficiency through its event representation approach. Rather than storing large chunks of summarized dialogue history, our framework maintains compact event triplets that capture essential semantic information. Our analysis shows that a typical 10-turn dialogue requires only 15-20 event nodes in EventWeave, compared to the more substantial storage requirements of full-text summarization methods. This compact representation helps mitigate memory concerns as conversations grow longer.

B.4.2 Computational Considerations

The selective attention mechanism in EventWeave provides computational advantages that partially offset the costs of graph maintenance. By calculating relevance scores between the current query and each event node, our model focuses computational resources on the most relevant portions of the dialogue history. This approach means that inference computation scales more with the number of relevant events than with the total graph size.

B.4.3 Scalability Analysis

To quantify EventWeave's efficiency at scale, we conducted additional experiments on extended dialogues. We expanded our test dialogues to 35+ turns and measured three key metrics: memory file size (KB), inference time (seconds), and response quality (evaluated by GPT-4 on a 10-point scale). As shown in Table 11, EventWeave without pruning demonstrates slightly higher resource usage than baseline methods, with a memory footprint approximately 12% larger than LifeLongMem and 50% larger than MemWalker.

However, with our pruning technique applied, EventWeave's resource requirements become comparable to LifeLongMem while maintaining substantially better performance. EventWeave+Pruning achieves a 3% reduction in memory usage and an 11% reduction in inference time compared to standard EventWeave, while preserving response quality scores significantly above the baselines.

The pruning mechanism operates by periodically removing event nodes with low global relevance scores while preserving the structural integrity of core event relationships. This approach enables EventWeave to maintain strong performance even as dialogues extend to dozens of turns, making it practical for deployment in real-world conversational systems.

For multi-speaker scenarios with potentially overlapping events, EventWeave's relationship modeling provides additional advantages. By explicitly capturing speaker-specific event relationships, our framework can distinguish between similar events initiated by different participants, enhancing response appropriateness in group conversations. This capability becomes increasingly important as the number of speakers and potential event overlaps increase.

Our analysis demonstrates that while graph expansion does introduce computational overhead, EventWeave's efficient event representation and selective attention mechanism, combined with strategic pruning, create a scalable approach that balances comprehensive context modeling with practical resource constraints.

B.5 Impacts of Threshold in EventWeave

We analyze three thresholds in our framework to see how each threshold affects different metrics. The threshold δ groups events into clusters if their embeddings pass a similarity check. The threshold θ creates reasoning links between events if they exceed a similarity value. The threshold β checks how much an event changes between layers. We vary each threshold from 0.3 to 0.9. Figure 6 shows how each threshold affects. Experimental results demonstrate 0.8, 0.6, and 0.85 give higher scores, respectively.

B.6 Analysis of Relationship Type Coverage

We conducted both quantitative and qualitative analyses to discuss the coverage of relations.

Edge Type Distribution. Table 13 presents the distribution of edge types in the Chronicle dataset. Our analysis reveals that the three primary relationship types account for over 96% of all identified connections in dialogue histories. Sequential edges, representing turn-by-turn progression, constitute the majority (58%), while temporal and reasoning-based edges account for 22% and 24% respectively. These three categories effectively capture the fundamental aspects of dialogue: chronology (sequential/temporal edges) and thematic coherence (reasoning-based links).

Ablation Studies. To empirically validate the sufficiency of our proposed typology, we conducted ablation experiments with different edge type configurations, as shown in Table 12. When using only individual edge types or pairs of edge types, performance decreases substantially across all metrics. For instance, using only sequential edges results in a 13.6% drop in TUA compared to our full model. This confirms that each relationship type contributes meaningfully to the model's understanding of dialogue history.

Additional Relationship Types. We also explored potential additional relationship types, including causal relationships (e.g., "advice" → "relationship fix"), emotional connections (e.g., "frustration" → "apology"), and rhetorical devices (e.g., irony, sarcasm). When incorporating these additional types into EventWeave, we observed only marginal improvements (+0.9% on average across metrics), suggesting that our three primary types sufficiently capture most critical dialogue dynamics.

The results demonstrate that while specialized relationship types may offer slight improvements in certain contexts, our proposed three-type framework provides a robust and comprehensive founda-

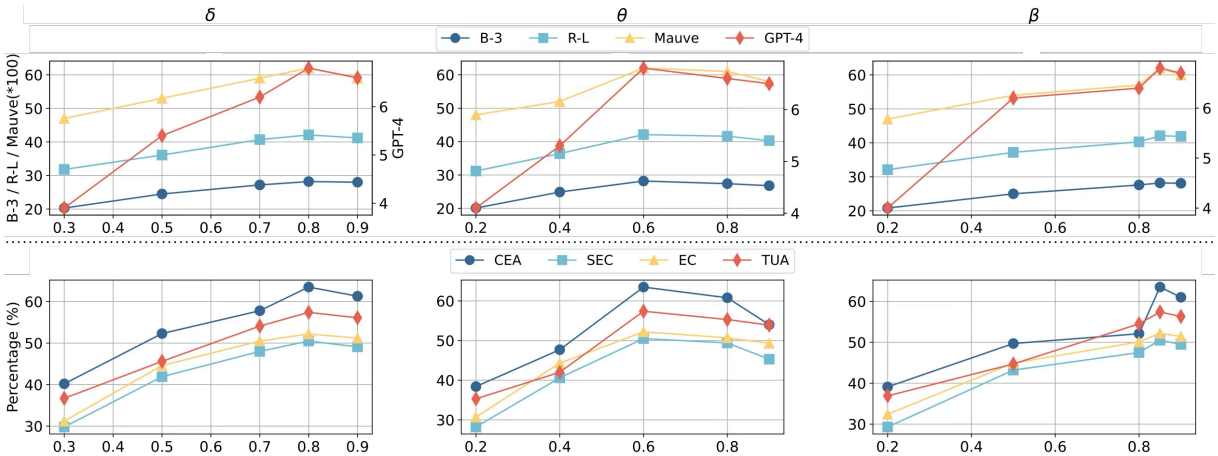

Figure 6: Experimental results about how each threshold affects different metrics.

Edge Type Configuration	**CEA ↑**	**SEC ↑**	**EC ↑**	**TUA ↑**
EventWeave (Base: Sequential + Temporal + Reasoning)	68.5	56.2	58.3	61.2
Only Sequential	54.3	43.8	45.2	47.6
Only Temporal	50.1	41.2	42.6	44.3
Only Reasoning-based	52.7	42.5	44.1	46.4
Sequential + Temporal	62.1	50.7	52.9	55.2
Sequential + Reasoning	63.8	52.3	54.1	56.7
Temporal + Reasoning	61.4	49.8	51.6	54.3
Base + Causal	69.2	57.4	58.9	61.8
Base + Emotional	68.9	56.8	58.7	61.5
Base + Rhetorical	68.7	56.5	58.5	61.2
Base + All Additional Types	69.5	57.6	59.1	62.1

Table 12: Ablation study on relationship types in EventWeave. Results show that our three base relationship types (Sequential, Temporal, and Reasoning-based) capture most of the critical dialogue dynamics. Adding additional relationship types (Causal, Emotional, Rhetorical) provides only marginal improvements (+0.9% on average), confirming the sufficiency of our proposed typology. All experiments conducted with GPT-4o on both Chronicle and MSC datasets, with scores averaged.

Relationship Type	**% of Edges**	**Example**
Sequential	58%	Turn $t \rightarrow$ Turn $t+1$
Temporal	22%	"yesterday" $\rightarrow$ "next week"
Reasoning-based	24%	"boat-building" $\rightarrow$ "treehouse plans"
Causal	2.1%	"advice" $\rightarrow$ "relationship fix"
Emotional	1.3%	"frustration" $\rightarrow$ "apology"
Rhetorical	0.5%	"ironic comment" $\rightarrow$ "literal response"

Table 13: Distribution of edge types in Chronicle dataset. The three primary relationship types cover over 96% of all identified connections, with additional specialized types accounting for less than 4% of edges.

tion for modeling dialogue history. This aligns with previous findings that most conversational structures can be effectively represented through sequential progression, temporal references, and reasoning connections. We acknowledge that certain specialized contexts (e.g., highly rhetorical discourse) might benefit from additional relationship modeling, which we discuss as a limitation

and direction for future work.

C Prompts

C.1 GPT-as-the-Judge

Figure 7 shows the example of the gpt-as-the-judge prompts.

C.2 Question Generation

We use a separate LLM to generate questions for the event-based evaluation. There are four metrics for the evaluation: Core Event Awareness (CEA), Supporting Event Coverage (SEC), Event Coherence (EC), and Temporal Update Accuracy (TUA). And we use different question-generation prompts for each metric. The prompts are shown in Figure 8.

C.3 Inference Prompts

All models included in our experiments were used the same inference prompts to ensure a fair comparison. The prompts were shown in Figure 9.

D Human Evaluation

We conducted a comprehensive human evaluation to assess the quality of responses generated by different memory methods. We recruited 12 annotators (graduate students with NLP backgrounds) to evaluate 100 randomly selected dialogue examples from each dataset. Each example was evaluated by 3 different annotators to ensure reliability. The annotators were asked to rate model responses on four dimensions: coherence, correctness, style, and helpfulness, with scores ranging from 1 (very poor) to 10 (excellent).

D.1 Evaluation Protocol

Annotators were presented with dialogue contexts and model-generated responses without knowing which memory method produced each response. They evaluated each response independently according to the guidelines provided in Figure 10. To ensure consistency, we conducted a training session where annotators evaluated sample responses and discussed their ratings to align their understanding of the evaluation criteria.

D.2 Results

Table 14 presents the results of our human evaluation. EventWeave consistently outperforms other memory methods across all four dimensions. The most significant improvements are observed in coherence and correctness, which aligns with our method's focus on maintaining event consistency and accurate temporal relationships. The inter-annotator agreement measured by Fleiss' Kappa was 0.72, indicating substantial agreement among annotators.

E Case Studies

This case study explores a multi-session dialogue between Neighbors A and B, capturing the evolving events and their relationships throughout the dialogue. The dialogue progresses through five sessions, each introducing core events and supporting events that shape the interaction. Table 15 shows the detailed the case.

 # GPT-AS-THE-JUDGE Prompts

Task:
Please act as an impartial judge and evaluate the quality of the responses. I will give you three pieces of information: a user query, a model-generated answer, and a golden answer.
Your task: Evaluate how well the model answer addresses the user's query and how close it matches the golden answer in terms of helpfulness, correctness, and completeness.
Please provide a score from 1 to 10, where 1 means the model's answer is very poor and 10 means it perfectly matches the golden answer and fully addresses the query.
Also provide a brief explanation for your score.

Context Information:
(USER INSTRUCTION)
Example:
(OPTIONAL EXAMPLES PROVIDE TO MODEL)
Query:
(QUERY)
Required Answer Format:
(OPTIONAL ANSWER FORMAT)

Figure 7: Example for gpt-as-the-judge Prompts.

Method	Coherence	Correctness	Style	Helpfulness	Average
GPT-4o (vanilla)	6.8	6.5	7.9	6.7	7.0
+ ProactiveCoT	7.5	7.2	8.0	7.4	7.5
+ MemWalker	7.8	7.6	8.1	7.7	7.8
+ LongMem	7.7	7.5	8.0	7.6	7.7
+ LifeLongMem	8.0	7.9	8.2	7.9	8.0
+ **EventWeave (Ours)**	**8.7**	**8.5**	**8.4**	**8.6**	**8.6**

Table 14: Human evaluation results on a scale of 1-10 (higher is better). EventWeave outperforms other memory methods across all dimensions, with particularly strong improvements in coherence and correctness.

 # QUESTION GENERATION Prompts

For Core Event Awareness
You are given a dialogue context in which there is a main topic or core event. Based on this context, generate a question that asks the model to identify the core event or the main topic of the conversation. The question should encourage the model to focus on identifying the key event or subject that the conversation revolves around. For example, the question might be: "What is the main topic of the conversation?" or "What is the core event discussed in the conversation?"

For Supporting Event Coverage
You are given a dialogue context that includes both a core event and supporting events. Generate a question that asks the model to identify the relevant supporting events that influence or augment the core event. The question should focus on asking for additional details that provide context or elaboration on the main event. For example, the question might be: "What supporting details or events are relevant to the main event?" or "What additional events provide context to the core event in this conversation?"

For Event Coherence
You are given a dialogue context that includes both a core event and supporting events. Generate a question that asks the model to combine these events into a coherent, fluent response. The goal is to evaluate how well the model can integrate both the core and supporting events into a single, smooth response. For example, the question might be: "How can you integrate the main event and supporting details into a single, coherent response?" or "Can you provide a smooth and natural response that combines the core and supporting events?"

For Temporal Update Accuracy
You are given a dialogue context in which there has been an update or change to an event. Generate a question that asks the model to identify any updated or replaced events in the conversation. The question should focus on detecting temporal changes, such as whether any events have been updated, replaced, or made outdated during the conversation. For example, the question might be: "Has any event been updated or replaced in the conversation?" or "Which events have changed or become outdated in this conversation?"

Figure 8: Example for question generation Prompts.

 # INFERENCE Prompts

Task: Please respond to the user's following instruction and you may:
(1) appropriately decline the user's instruction and provide reasons.
(2) point out the false assumptions in the user's instruction.
(3) suggest possible corrections for the false assumption to the user.
(4) include any truthful knowledge that you think could help correct the user's erroneous belief and enhance the authenticity of the response.

Context Information:
(USER INSTRUCTION)
Example:
(OPTIONAL EXAMPLES PROVIDE TO MODEL)
Query:
(QUERY)
Required Answer Format:
(OPTIONAL ANSWER FORMAT)

Figure 9: Example for Inference Prompts.

 # Guidelines for Human Evaluation

Overview.
Each annotator will be presented with a short user query or a multi-turn dialogue, followed by a response generated by one of several models. We ask you to rate each response on the four dimensions described below. Each dimension should be assigned a numerical score from 1 (very poor) to 10 (excellent). You do not need to compare responses across models directly; instead, evaluate each response on its own merit with respect to the conversation.

1. Coherence: Measures how logically consistent and contextually relevant the response is in relation to the conversation history. A coherent response should follow smoothly from the user's query or previous dialogue turns, without contradictions or abrupt topic changes.
Example.
Poor (Score 2):
User: "Where did Marie Curie do most of her work?" Response: "You can buy apples on sale at the corner store." (Off-topic, unrelated, incoherent)
Strong (Score 9):
User: "Where did Marie Curie do most of her work?" Response: "She conducted the majority of her research at the University of Paris, and later at her own Radium Institute in France." (Fully aligns with the user's question, consistent and relevant)

2. Correctness: Measures the factual accuracy and reliability of the response. A correct response should adhere to established or verifiable facts when they are presented in the conversation or are widely known.
Example.
Poor (Score 3):
User: "Who painted the Mona Lisa?"
Response: "It was painted by Vincent van Gogh in 1889." (Factually incorrect—Mona Lisa was painted by Leonardo da Vinci.)
Strong (Score 10):
User: "Who painted the Mona Lisa?"
Response: "It was painted by Leonardo da Vinci in the early 16th century." (Facts are accurate and relevant)

3. Style: Refers to the language clarity, fluency, and appropriateness of tone. A good style conveys the message in a clear, succinct, and context-appropriate manner.
Example.
Poor (Score 2):
User: "What's your favorite book?"
Response: "Book good me is me read." (Unclear grammar, lacks readability)
Strong (Score 9):
User: "What's your favorite book?"
Response: "I really enjoy reading 'To Kill a Mockingbird' for its thoughtful perspective on justice and morality." (Clear, fluent, properly toned)

4. Helpfulness: Measures how effectively the response addresses the user's needs or queries. A helpful response may offer additional clarification, relevant examples, or suggestions for next steps. Example.
Poor (Score 2):
User: "Could you recommend a good fitness routine for beginners?"
Response: "Sure, have fun." (Provides no actionable suggestions or relevant details)
Strong (Score 10):
User: "Could you recommend a good fitness routine for beginners?"
Response: "A balanced program could include three days of light cardio, such as walking or jogging for 20–30 minutes, and two days of bodyweight exercises like push-ups and squats. Starting slowly and gradually increasing intensity can help you avoid injury." (Directly addresses the user's request with clear suggestions)

Figure 10: Annotator Guidance.

Session 1

Neighbors A: "I spent the whole weekend building a boat out of scraps I found around the neighborhood."
Neighbors B: "Wow! That sounds like a lot of work. Did it turn out well?"
Neighbors A: "Actually, it's better than I expected! I took it out for a spin on the lake yesterday and it held up pretty nicely."
Neighbors B: "That's awesome. You're really creative and talented at building things."
Neighbors A: "Thanks! It's always been a hobby of mine. I remember when I was younger, I used to build elaborate forts in my backyard."
Neighbors B: "I used to do that too! It's funny how our childhood hobbies stick with us."
Neighbors A: "Definitely. Building things brings me a lot of joy and satisfaction."
Neighbors B: "It's great to have a hobby like that. What do you plan on building next?"
Neighbors A: "I'm not quite sure yet. Maybe a treehouse or a birdhouse. I'll have to see what materials I can find around here."

Session 2

Neighbors A: "I just love spending time with you, Neighbors B. Being close to you makes me feel happy and at peace."
Neighbors B: "Aw, that's so sweet of you to say, Neighbors A. I feel the same way about you."
Neighbors A: "It's not just the physical closeness, although that is nice. It's also the mental and emotional connection we have. I feel like we really understand each other."
Neighbors B: "Yes, I agree. It's rare to find someone who truly gets you, but I think we have that kind of relationship."
Neighbors A: "I'm just so grateful for you, Neighbors B. You're like a balm to my soul."
Neighbors B: "That's such a beautiful way to put it. I feel like we bring out the best in each other."
Neighbors A: "Definitely. And whenever I'm feeling down or stressed, just being around you makes everything feel alright again."
Neighbors B: "I'm glad I can be there for you, Neighbors A. We're lucky to have each other as neighbors and friends."
Neighbors A: "Absolutely. Speaking of which, have you seen the boat I built out of scraps? I could take you out on it sometime."
Neighbors B: "That sounds like so much fun! I remember you mentioning it before, but I'd love to see it in person. It's amazing what you can create with a little bit of creativity and resourcefulness."

Session 3

Neighbors A: "Hey, can I talk to you about something?"
Neighbors B: "Sure, of course. What's on your mind?"
Neighbors A: "It's about my girlfriend. We're going through a rough patch right now and I just don't know what to do."
Neighbors B: "I'm sorry to hear that. What seems to be the issue?"
Neighbors A: "She's been really distant lately and I feel like we're growing apart. I don't know if there's someone else or if she's just not into me anymore."
Neighbors B: "Hmm, that's tough. Have you tried talking to her about it?"
Neighbors A: "Yeah, but she just seems to shut down and doesn't want to talk about anything."
Neighbors B: "Well, maybe you could try doing something special for her to show her how much you care. Like cooking her a nice dinner or going on a romantic getaway."
Neighbors A: "That's a good idea. I'll try that and see if it helps. Thanks for the advice, I really appreciate it."
Neighbors B: "No problem, happy to help. And remember, communication is key in any relationship. Don't be afraid to keep talking to her and trying to work things out together."
Neighbors A: "Yeah, I'll definitely keep that in mind. And speaking of relationships, remember how we talked about how happy I was with my girlfriend a few months ago? Your advice really helped and we're doing great now."
Neighbors B: "I'm so glad to hear that! You guys really seem to be meant for each other."

Session 4

Neighbors A: "I did something really drastic the other day."
Neighbors B: "What did you do?"
Neighbors A: "I faked my death."
Neighbors B: "What?! Why would you do that?"
Neighbors A: "I had to. I can't tell you why, but it was necessary."
Neighbors B: "That's insane. Did anyone find out?"
Neighbors A: "No, I think I pulled it off pretty well. But it was nerve-wracking waiting for someone to find me."
Neighbors B: "I can only imagine. But isn't it risky to pretend to be dead?"
Neighbors A: "It was a risk I had to take. You know I always think outside the box."
Neighbors B: "That's true. But please don't do something like that again without letting me in on the plan."
Neighbors A: "I promise I won't. Do you remember when I built that boat out of scraps?"
Neighbors B: "Yes, that was amazing. You're so creative."
Neighbors A: "That feeling of creating something from nothing is what drove me to do this. I knew I could pull it off, just like I knew I could build that boat."
Neighbors B: "I see. Well, I'm just glad you're okay and we're having this conversation right now."
Neighbors A: "Me too. And hey, do you remember that time I came to you for advice about my girlfriend?"
Neighbors B: "Of course. I'm always here to listen and help."
Neighbors A: "Your advice really helped. I wouldn't have been able to fix things with her without you."
Neighbors B: "I'm glad I could help. And if you ever need anything again, don't hesitate to come to me."
Neighbors A: "I won't. You're the best neighbor anyone could ask for."
Neighbors B: "You're not too shabby yourself."

Session 5

Neighbors A: "Thanks for offering me the job at the gas station, I really appreciate it."
Neighbors B: "No problem. I know you need the money to buy that car you've been eyeing."
Neighbors A: "Yeah, I've been saving up for a while now, but it's just not enough. And since you work there, it's perfect."
Neighbors B: "I'm glad I can help. Plus, it'll be fun working together."
Neighbors A: "Definitely. Remember that boat I built a while back? I think we should take it out on the lake again soon."
Neighbors B: "That sounds like a great idea. We had so much fun the last time we did that."
Neighbors A: "And do you remember when I came to you for advice about my girlfriend?"
Neighbors B: "Of course I do. I'm so happy things worked out for you two."
Neighbors A: "Me too. You always give the best advice. Speaking of which, do you have any tips for working at the gas station?"
Neighbors B: "Just be friendly and helpful to all the customers. And don't forget to keep the place clean and organized."
Neighbors A: "Sounds easy enough. Thanks for the advice. I can't wait to start."

Core Events	**Supporting Events**
Boat Building and Shared Hobbies	(Neighbors A, Completes, Boat (built from scraps, tested on the lake)) (Neighbors B, Shares, Memory of childhood hobbies (building forts)) (Neighbors A, Mentions, Future project (possibly building a treehouse or birdhouse))
Emotional Connection and Offering Support	(Neighbors B, Reciprocates, Sentiment of emotional bonding (shares feelings of appreciation)) (Neighbors B, Expresses, Support role (being there for Neighbors A emotionally)) (Neighbors B, Suggests, Boat trip (taking a ride on the boat Neighbors A built))
Relationship Concerns and Advice	(Neighbors A, Expresses, Emotional distress (feeling distant from their girlfriend)) (Neighbors B, Suggests, Special gesture (cooking dinner or going on a romantic getaway to rekindle the relationship)) (Neighbors B, Encourages, Communication (keep talking to work through relationship issues))
Drastic Decision and Reflection	(Neighbors A, Acknowledges, Risk (pretending to be dead, the potential dangers of the plan)) (Neighbors A, Expresses, Gratitude (thanks Neighbors B for past advice that helped in the relationship)) (Neighbors B, Reaffirms, Friendship (importance of being there for each other)) (Neighbors A, Acknowledges, Feeling of accomplishment (creating something from nothing, like the boat))
Job Offer and Future Collaboration	(Neighbors A, Acknowledges, Job offer (appreciates the opportunity at the gas station)) (Neighbors A, Suggests, Boat trip (taking the boat out on the lake again soon)) (Neighbors B, Reinforces, Friendship (importance of mutual gratitude and support)) (Neighbors A, Expresses, Enthusiasm (looking forward to working together))

Event Update Flow:

Session 1:
(Neighbors A, Completes, Boat) → (Neighbors A, Mentions, Future project)

Session 2:
(Neighbors A, Shares, Gratitude) → (Neighbors B, Reciprocates, Sentiment of emotional bonding)

Session 3:
(Neighbors A, Expresses, Emotional distress) → (Neighbors B, Suggests, Special gesture)
(Neighbors A, Expresses, Emotional distress) → (Neighbors B, Encourages, Communication)

Session 4:
(Neighbors A, Acknowledges, Risk) → (Neighbors A, Expresses, Gratitude)
(Neighbors A, Acknowledges, Feeling of accomplishment) → (Neighbors A, Mentions, Boat)

Session 5:
(Neighbors A, Acknowledges, Job offer) → (Neighbors A, Suggests, Boat trip)
(Neighbors B, Reinforces, Friendship) → (Neighbors A, Expresses, Enthusiasm)

Table 15: Case from Conversation Chronicle.