

# 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,<sup>1</sup> (Levison, 1987; Koskinen et al.,

<sup>1</sup>Events are commonly divided into basic events (core) and processes. A process is a complex event made up of a

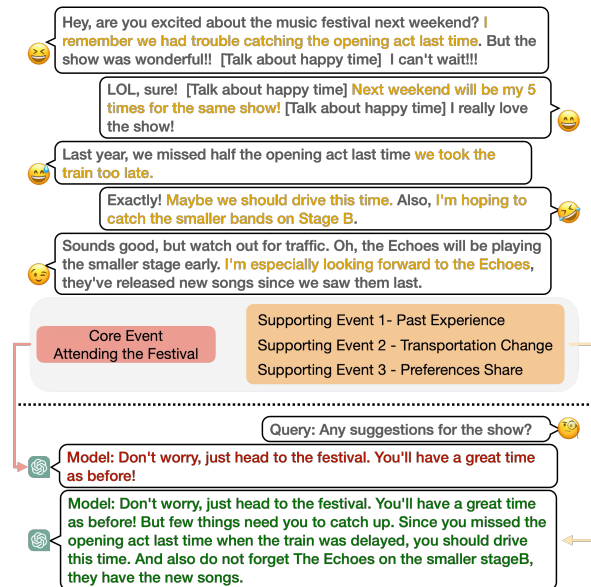


Figure 1: An illustrative conversation between two friends planning to attend a music festival. Supporting events play a key role in enriching the dialogue’s context and shaping the interpersonal dynamic.

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.

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-

058	explicitly modeling the relationships between events	110
059	or their varying importance to the current dialogue	111
060	turn (Kirstein et al., 2025). Similarly, graph-based	112
061	approaches merge conversation segments into high-	
062	level nodes but may not effectively distinguish be-	
063	tween core and supporting events (Walker et al.,	
064	2022). Both approaches face challenges in bal-	
065	ancing comprehensive context understanding with	
066	generating concise, natural responses that don't	
067	overwhelm users with unnecessary details.	
068	To address these challenges, we introduce <b>Event-</b>	
069	<b>Weave</b> , a framework that explicitly models event	
070	relationships to improve dialogue response gen-	
071	eration by constructing a dynamic event graph	
072	that identifies and organizes conversational events	
073	based on their semantic relationships. Our key	
074	insight is that effective dialogue responses don't	
075	require surfacing all available information—instead,	
076	they need to access the right information at the right	
077	time. Unlike systems that simply recall all details,	
078	EventWeave distinguishes between three types of	
079	event relationships to determine which informa-	
080	tion is most relevant to the current conversation	
081	turn, employing a multi-head attention mechanism	
082	that calculates relevance scores between the current	
083	query embedding and each event node representa-	
084	tion. This selective attention mechanism dynam-	
085	ically selects the most appropriate events based on	
086	both semantic similarity and relationship structure	
087	within the graph, enabling the model to generate	
088	concise, natural responses that maintain conversa-	
089	tional flow while addressing the user's current	
090	needs without overwhelming them with exhaustive	
091	recaps of the conversation history.	
092	We evaluate EventWeave on three benchmark	
093	datasets: Conversation Chronicle (CC) (Jang et al.,	
094	2023), Multi-Session Conversations (MSC) (Xu	
095	et al., 2021), and LoCoMo (Maharana et al., 2024).	
096	Our evaluation metrics focus not just on informa-	
097	tion recall but on response naturalness, rele-	
098	levance, and conversational appropriateness. Import-	
099	antly, our ablation studies confirm that the im-	
100	provements come from better event relationship	
101	modeling rather than simply including more infor-	
102	mation in responses. To address scalability con-	
103	cerns with longer dialogues, we additionally con-	
104	duct a comprehensive analysis showing how Event-	
105	Weave performs across varying dialogue lengths	
106	and graph sizes, demonstrating that EventWeave	
107	maintains strong performance up to approximately	
108	35 turns, after which a pruning mechanism effec-	
109	tively manages computational overhead while pre-	
	serving critical event relationships.	110
	Our main contributions can be summarized as	111
	follows:	112
	• We introduce a framework for dialogue response	113
	generation that explicitly models relationships	114
	between events, enabling nuanced understanding	115
	of conversation context.	116
	• We develop a dynamic event graph construction	117
	method that organizes events into semantic clus-	118
	ters with three relationship types, helping distin-	119
	guish between information that should be sur-	120
	facied in responses and information that should	121
	remain in the background.	122
	• We demonstrate that our approach generates	123
	more natural and contextually appropriate re-	124
	sponses while requiring less computational over-	125
	head compared to models that process the entire	126
	dialogue history.	127
	• We provide a detailed scalability analysis show-	128
	ing how our approach can be optimized for dia-	129
	logues of varying lengths, with specific optimiza-	130
	tion techniques for long-term conversations.	131
	<b>2 Related Works</b>	132
	<b>2.1 Dialogue Context Management Systems</b>	133
	Multi-turn dialogue systems traditionally employ	134
	hierarchical architectures or memory-augmented	135
	networks to maintain contextual understanding	136
	across interactions (Firdaus et al., 2021; Li et al.,	137
	2024; Joko et al., 2024). While these approaches	138
	effectively handle short-to-medium dialogues, they	139
	struggle with lengthy or multi-session conversa-	140
	tions due to “information explosion” and “informa-	141
	tion conflict” issues. Graph-based alternatives (Tu	142
	et al., 2024; Zhang et al., 2021; Tuan et al., 2022)	143
	offer structured representations of dialogue knowl-	144
	edge and relationships, but often focus only on	145
	high-level events while neglecting subtle yet con-	146
	textually important utterances (Yang et al., 2020;	147
	Ong et al., 2024). This limitation reduces their abil-	148
	ity to distinguish between irrelevant exchanges and	149
	crucial context-supplying details that justify core	150
	conversational events.	151
	<b>2.2 Prompt-based Inference with Structured</b>	152
	<b>Representations</b>	153
	Our work bridges the gap between structured dia-	154
	logue representations (Li et al., 2021; Han et al.,	155
	2021) and prompt-based inference methods (Besta	156

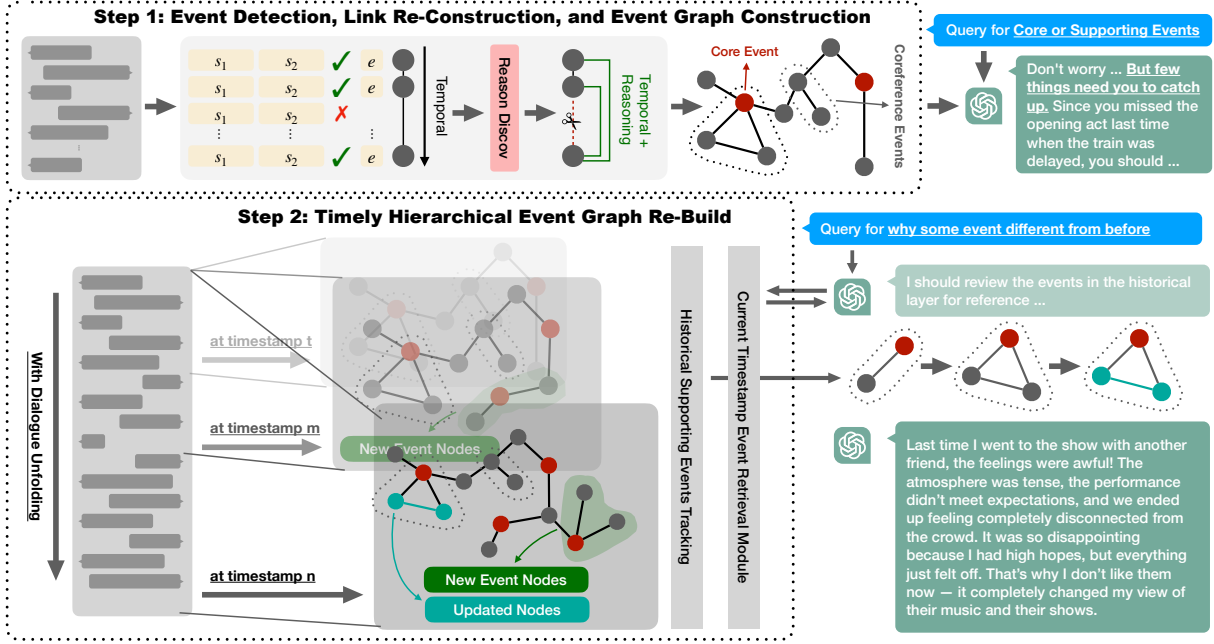


Figure 2: EventWeave architecture illustrating the three main components: (1) Event Extraction Module that identifies core and supporting events from dialogue turns, (2) Dynamic Graph Construction that establishes relationship types between events, and (3) Context-Aware Response Generation that selectively 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, shown 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 OpenDialog 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  $\delta$ :

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

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:

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

**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  $\theta$ :

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

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:

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

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:

$$\mathbf{v}_{s_t} = \sum_j \alpha_j \mathbf{v}_{e_j} \quad (6)$$

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

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

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:

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

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:

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

### 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:

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

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:

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

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 responding, 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:

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

where  $\gamma$  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:

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

**Graph-structural retrieval:** Utilizes graph topology to identify structurally important nodes

via a modified PageRank algorithm, where transition probabilities incorporate edge semantics:

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

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:

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

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

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

where  $\varepsilon$  is a temperature parameter controlling attention distribution sharpness.

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

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

## 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
+ <b>EventWeave (Ours)</b>	<b>32.2</b>	<b>45.8</b>	<b>0.67</b>	<b>7.9</b>	<b>35.0</b>	<b>48.4</b>	<b>0.70</b>	<b>8.3</b>	<b>28.4</b>	<b>42.6</b>	<b>0.65</b>	<b>7.6</b>	<b>65.7</b>	<b>54.1</b>	<b>56.2</b>	<b>59.0</b>
ReBotGEN (specialized on CC)	25.0	40.5	0.63	4.8	–	–	–	–	–	–	–	–	30.8	23.4	23.7	25.5
+ <b>EventWeave (Ours)</b>	<b>33.5</b>	<b>47.0</b>	<b>0.69</b>	<b>8.2</b>	–	–	–	–	–	–	–	–	<b>68.9</b>	<b>55.3</b>	<b>58.2</b>	<b>61.1</b>
HAHT (specialized on MSC)	–	–	–	–	27.3	43.0	0.63	5.0	–	–	–	–	31.8	23.6	24.1	26.5
+ <b>EventWeave (Ours)</b>	–	–	–	–	<b>34.2</b>	<b>46.8</b>	<b>0.68</b>	<b>8.0</b>	–	–	–	–	<b>71.0</b>	<b>57.2</b>	<b>60.3</b>	<b>62.4</b>
DialogLM (specialized on LoCoMo)	–	–	–	–	–	–	–	–	22.8	37.5	0.60	4.9	32.5	24.2	24.8	27.0
+ <b>EventWeave (Ours)</b>	–	–	–	–	–	–	–	–	<b>29.1</b>	<b>43.4</b>	<b>0.66</b>	<b>7.8</b>	<b>72.8</b>	<b>58.1</b>	<b>61.5</b>	<b>63.2</b>

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-

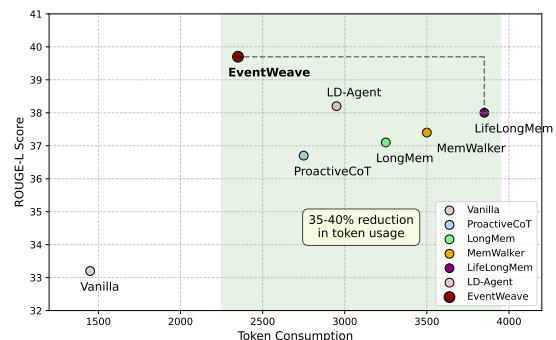


Figure 3: Memory efficiency comparison on LoCoMo. **EventWeave** achieves higher ROUGE-L scores with lower token consumption compared to other memory mechanisms.

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.

## 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	<b>32.2</b>	<b>45.8</b>	<b>0.67</b>	<b>7.9</b>	<b>35.0</b>	<b>48.4</b>	<b>0.70</b>	<b>8.3</b>	<b>28.4</b>	<b>42.6</b>	<b>0.65</b>	<b>7.6</b>	<b>65.7</b>	<b>54.1</b>	<b>56.2</b>	<b>59.0</b>
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.5.

## 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)

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
+ <b>EventWeave (Ours)</b>	<b>89.5</b>	<b>84.3</b>	<b>76.8</b>	<b>68.2</b>

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

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.4.

## 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. Details can be found in Appendix B.1. Besides, Appendix B.2 shows full experimental results. B.3 shows the few-shot settings. B.6 shows the impacts of threshold. B.7 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. Further study cases are in Appendix E.

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  $\theta$ ). (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.

## 562 Limitations

563 While our proposed framework demonstrates  
564 promising results in enhancing dialogue systems  
565 through dynamic event graph construction and hi-  
566 erarchical maintenance, we acknowledge certain  
567 computational considerations common to graph-  
568 based approaches. As with similar methods in this  
569 domain, the event graph’s expansion introduces  
570 computational demands that scale with the com-  
571 plexity of dialogue scenarios. The hierarchical  
572 structure, while beneficial for contextual under-  
573 standing, naturally requires additional computa-  
574 tional resources for maintenance compared to sim-  
575 pler models. These considerations, which are inher-  
576 ent to most graph-based approaches rather than spe-  
577 cific to our method, represent interesting avenues  
578 for optimization research. Our implementation has  
579 shown effective performance in experimental set-  
580 tings, and future work could explore further effi-  
581 ciency improvements through graph pruning tech-  
582 niques, distributed computing approaches, or adap-  
583 tive hierarchy management strategies. Such opti-  
584 mizations would make graph-based dialogue sys-  
585 tems increasingly practical for deployment across  
586 diverse application scenarios and computational  
587 environments.

## 588 Ethical Considerations

589 Our experimental methodology prioritizes ethical  
590 research practices through several key measures.  
591 All experiments utilize publicly available datasets  
592 that have been ethically sourced and comply with  
593 intellectual property regulations and privacy stan-  
594 dards. By deliberately avoiding fine-tuning models  
595 on domain-specific data, our approach inherently  
596 minimizes the potential introduction of unintended  
597 biases that often accompany specialized training  
598 protocols. These methodological choices reflect  
599 our commitment to research integrity and respon-  
600 sible AI development, supporting broader princi-  
601 ples of fairness, methodological transparency, and  
602 computational sustainability within the scientific  
603 community.

## 604 References

605 Vladimir Araujo, Alvaro Soto, and Marie-Francine  
606 Moens. 2023. A memory model for question an-  
607 swering from streaming data supported by rehearsal  
608 and anticipation of coreference information. In *Find-  
609 ings of the Association for Computational Linguistics:*

- ACL 2023*, pages 13124–13138, Toronto, Canada. As- 610  
sociation for Computational Linguistics. 611
- Maciej Besta, Nils Blach, Ales Kubicek, Robert Gersten- 612  
berger, Michal Podstawski, Lukas Gianinazzi, Joanna 613  
Gajda, Tomasz Lehmann, Hubert Niewiadomski, Pi- 614  
otr Nyczyk, et al. 2024. Graph of thoughts: Solving 615  
elaborate problems with large language models. In 616  
*Proceedings of the AAAI Conference on Artificial 617  
Intelligence*, volume 38, pages 17682–17690. 618
- Howard Chen, Ramakanth Pasunuru, Jason Weston, and 619  
Asli Celikyilmaz. 2023. [Walking down the mem- 620  
ory maze: Beyond context limit through interactive 621  
reading](#). *Preprint*, arXiv:2310.05029. 622
- Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, 623  
Wenqiang Lei, and Tat-Seng Chua. 2023. [Prompt- 624  
ing and evaluating large language models for proac- 625  
tive dialogues: Clarification, target-guided, and non- 626  
collaboration](#). *Preprint*, arXiv:2305.13626. 627
- Yiming Du, Hongru Wang, Zhengyi Zhao, Bin Liang, 628  
Baojun Wang, Wanjun Zhong, Zezhong Wang, and 629  
Kam-Fai Wong. 2024. [PerLTQA: A personal long- 630  
term memory dataset for memory classification, re- 631  
trieval, and fusion in question answering](#). In *Pro- 632  
ceedings of the 10th SIGHAN Workshop on Chinese 633  
Language Processing (SIGHAN-10)*, pages 152–164, 634  
Bangkok, Thailand. Association for Computational 635  
Linguistics. 636
- Mauajama Firdaus, Nidhi Thakur, and Asif Ekbal. 2021. 637  
Aspect-aware response generation for multimodal 638  
dialogue system. *ACM Transactions on Intelligent 639  
Systems and Technology (TIST)*, 12(2):1–33. 640
- Janghoon Han, Taesuk Hong, Byoungjae Kim, 641  
Youngjoong Ko, and Jungyun Seo. 2021. Fine- 642  
grained post-training for improving retrieval-based 643  
dialogue systems. In *Proceedings of the 2021 Con- 644  
ference of the North American Chapter of the Asso- 645  
ciation for Computational Linguistics: Human Lan- 646  
guage Technologies*, pages 1549–1558. 647
- Yongquan He, Peng Zhang, Luchen Liu, Qi Liang, 648  
Wenyuan Zhang, and Chuang Zhang. 2024. Hip net- 649  
work: Historical information passing network for ex- 650  
trapolation reasoning on temporal knowledge graph. 651  
*arXiv preprint arXiv:2402.12074*. 652
- Kai Tzu iunn Ong, Namyoun Kim, Minju Gwak, 653  
Hyungjoo Chae, Taeyoon Kwon, Yohan Jo, Seung 654  
won Hwang, Dongha Lee, and Jinyoung Yeo. 2025. 655  
[Towards lifelong dialogue agents via timeline-based 656  
memory management](#). *Preprint*, arXiv:2406.10996. 657
- Jihyoung Jang, Minseong Boo, and Hyounghun Kim. 658  
2023. [Conversation chronicles: Towards diverse tem- 659  
poral and relational dynamics in multi-session con- 660  
versations](#). In *Proceedings of the 2023 Conference 661  
on Empirical Methods in Natural Language Process- 662  
ing*, pages 13584–13606, Singapore. Association for 663  
Computational Linguistics. 664

665	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 <i>Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval</i> , pages 796–806.	
673	Frederic Kirstein, Jan Philip Wahle, Bela Gipp, and Terry Ruas. 2025. Cads: A systematic literature review on the challenges of abstractive dialogue summarization. <i>Journal of Artificial Intelligence Research</i> , 82:313–365.	
678	Kaj U Koskinen, Rainer Breite, Kaj U Koskinen, and Rainer Breite. 2020. Process philosophy. <i>Uninterrupted Knowledge Creation: Process Philosophy and Autopoietic Perspectives</i> , pages 13–24.	
682	Arnold B Levison. 1987. Events and time’s flow. <i>Mind</i> , 96(383):341–353.	
684	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. <i>arXiv preprint arXiv:2406.05925</i> .	
688	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. <i>Preprint</i> , arXiv:2406.05925.	
692	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. <i>ACM Transactions on Information Systems (TOIS)</i> , 39(4):1–25.	
698	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 <i>Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 1267–1280, Toronto, Canada. Association for Computational Linguistics.	
706	Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of llm agents. <i>Preprint</i> , arXiv:2402.17753.	
710	Tzu-iunn Ong, Namyoun 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. <i>arXiv e-prints</i> , pages arXiv–2406.	
716	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. <i>Advances in Neural Information Processing Systems</i> , 34:4816–4828.	
	Hiromasa Sakurai and Yusuke Miyao. 2024. Evaluating intention detection capability of large language models in persuasive dialogues. In <i>Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 1635–1657, Bangkok, Thailand. Association for Computational Linguistics.	722 723 724 725 726 727 728
	Quan Tu, Chongyang Tao, and Rui Yan. 2024. Multi-grained conversational graph network for retrieval-based dialogue systems. In <i>Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)</i> , pages 11756–11765.	729 730 731 732 733 734
	Yi-Lin Tuan, Sajjad Beygi, Maryam Fazel-Zarandi, Qiaozhi Gao, Alessandra Cervone, and William Yang Wang. 2022. Towards large-scale interpretable knowledge graph reasoning for dialogue systems. In <i>Findings of the Association for Computational Linguistics: ACL 2022</i> , pages 383–395.	735 736 737 738 739 740
	Deeksha Varshney, Niranshu Behera, Prajeet Katari, and Asif Ekbal. 2025. Medprom: Bridging dialogue gaps in healthcare with knowledge-enhanced generative models. <i>ACM Transactions on Computing for Healthcare</i> .	741 742 743 744 745
	Nicholas Thomas Walker, Stefan Ultes, and Pierre Lison. 2022. Graphwoz: Dialogue management with conversational knowledge graphs. <i>arXiv preprint arXiv:2211.12852</i> .	746 747 748 749
	Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. 2023. Augmenting language models with long-term memory. <i>Preprint</i> , arXiv:2306.07174.	750 751 752 753
	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 <i>Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 2748–2763, Bangkok, Thailand. Association for Computational Linguistics.	754 755 756 757 758 759 760 761
	Jing Xu, Arthur Szlam, and Jason Weston. 2021. Beyond goldfish memory: Long-term open-domain conversation. <i>Preprint</i> , arXiv:2107.07567.	762 763 764
	Shiquan Yang, Rui Zhang, and Sarah Erfani. 2020. GraphDialog: Integrating graph knowledge into end-to-end task-oriented dialogue systems. In <i>Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)</i> , pages 1878–1888, Online. Association for Computational Linguistics.	765 766 767 768 769 770 771
	Congchi Yin, Piji Li, and Zhaochun Ren. 2023. Ctrl-struct: Dialogue structure learning for open-domain response generation. In <i>Proceedings of the ACM web conference 2023</i> , pages 1539–1550.	772 773 774 775
	Haoyu Zhang, Meng Liu, Zan Gao, Xiaoqiang Lei, Yinglong Wang, and Liqiang Nie. 2021. Multimodal	776 777

778	dialog system: Relational graph-based context-aware question understanding. In <i>Proceedings of the 29th ACM international conference on multimedia</i> , pages 695–703.	830
779		831
780		832
781		
782	Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. <i>Personalizing dialogue agents: I have a dog, do you have pets too?</i> In <i>Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.	833
783		834
784		835
785		836
786		837
787		838
788		839
789		840
790	Jianan Zhao, Le Zhuo, Yikang Shen, Meng Qu, Kai Liu, Michael Bronstein, Zhaocheng Zhu, and Jian Tang. 2023. GraphText: Graph reasoning in text space. <i>arXiv preprint arXiv:2310.01089</i> .	841
791		842
792		843
793		844
794	Hanxun Zhong, Zhicheng Dou, Yutao Zhu, Hongjin Qian, and Ji-Rong Wen. 2022. <i>Less is more: Learning to refine dialogue history for personalized dialogue generation</i> . In <i>Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies</i> , pages 5808–5820, Seattle, United States. Association for Computational Linguistics.	845
795		846
796		847
797		848
798		
799		
800		
801		
802		
803	<b>A Detailed Experiments Setups</b>	
804	<b>A.1 Datasets Usage</b>	
805	Table 5 shows the statistics of two datasets. The information is taken from their papers.	850
806		851
807	<b>Multi-Session Chat.</b> 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.	852
808		853
809		854
810		855
811		856
812		857
813		858
814		859
815		860
816		861
817		862
818		863
819		864
820		865
821	<b>Conversation Chronicle.</b> 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	866
822		867
823		868
824		869
825		870
826		
827		
828		
829		
	for exploring how relationship-based factors shape multi-session conversations. Table 7 demonstrates the sample from Conversation Chronicle.	830
		831
		832
	<b>LoCoMo.</b> 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.	833
		834
		835
		836
		837
		838
		839
		840
		841
		842
		843
		844
		845
		846
		847
		848
	<b>A.2 Baselines Comparison</b>	849
	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.	850
		851
		852
		853
		854
		855
		856
		857
		858
		859
		860
		861
		862
		863
		864
		865
		866
		867
		868
		869
		870
	<b>A.3 Metric Computing</b>	871
	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.	872
		873
		874
		875
		876
	<b>Core Event Awareness (CEA)</b> measures the model’s ability to recall the correct core event. It	877
		878

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

<b>One for Speaker 1 and another for Speaker 2</b>	
<b>Hi, how are you doing today?</b>	<b>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?</b>
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 schwarzkraat...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.

<b>Relationship: Classmates</b>
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.
<b>Relationship: Student and Teacher</b>
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.
<b>Relationship: Parents and Child</b>
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.

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

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.

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

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.

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

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.

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

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.

918	<b>A.4 Implementation Details</b>	
919	For fair comparison, all methods use the same con-	
920	text window size (4k tokens) and the same back-	
921	bone LLMs. Event graphs in our approach are con-	
922	structed incrementally as conversations progress,	
923	with a maximum of 30 nodes per graph. For each	
924	test case, we use all previous sessions to build the	
925	event graph, while keeping the final session for test-	
926	ing. For each turn in the test session, the model is	
927	provided with the current query and must generate	
928	an appropriate response utilizing information from	
929	past sessions. We use a maximum of 5 previous	
930	sessions as context.	
931	<b>B Full Experimental Results</b>	
932	<b>B.1 Detailed Settings and Results on</b>	
933	<b>Scalability</b>	
934	In this section, we provide a comprehensive analy-	
935	sis of how EventWeave scales with increasing dia-	
936	logue length compared to vanilla LLM approaches.	
937	This analysis addresses an important question in	
938	dialogue systems research: how effectively can	
939	models maintain coherent context understanding	
940	as conversations become longer?	
941	<b>B.1.1 Dataset Distribution by Dialogue</b>	
942	<b>Length</b>	
943	To conduct a thorough scalability analysis, we cat-	
944	egorized dialogues from the Chronicle dataset into	
945	four length categories. Table 8 shows the distribu-	
946	tion of dialogues across these categories.	
947	<b>B.1.2 Experimental Setup</b>	
948	We evaluated both GPT-4o (vanilla) and Even-	
949	tWeave across all four length categories. For each	
950	category, we measured performance using:	
951	• <b>GPT-4 Rating:</b> Evaluation scores (1-10) pro-	
952	vided by GPT-4 based on coherence, contex-	
953	tual appropriateness, and response quality	
954	• <b>Event Coherence (EC):</b> Percentage of re-	
955	sponses that correctly maintain event relation-	
956	ships and references across the dialogue his-	
957	tory	
958	To ensure statistical robustness, we conducted	
959	paired t-tests to determine if the performance dif-	
960	ferences between EventWeave and vanilla GPT-4o	
961	were statistically significant.	
	<b>B.1.3 Detailed Results</b>	962
	Table 9 presents the detailed performance compar-	963
	ison between GPT-4o (vanilla) and EventWeave	964
	across different dialogue lengths, including abso-	965
	lute differences and statistical significance.	966
	<b>B.1.4 Analysis of Scalability Performance</b>	967
	The results in Table 9 demonstrate a clear pattern:	968
	while vanilla GPT-4o’s performance degrades with	969
	increasing dialogue length by dropping from 5.2 to	970
	3.8 representing a 27% decrease, EventWeave’s	971
	performance actually has a 40% improvement).	972
	The Event Coherence (EC) metric shows an even	973
	more pronounced divergence, with vanilla GPT-4o	974
	declining from 24.8% to 15.3% as conversations	975
	lengthen, while EventWeave improves dramatically	976
	from 40.2% to 72.6%. This stark difference is at-	977
	tributed to the structure of EventWeave. Rather	978
	than processing the entire dialogue history, Even-	979
	tWeave’s graph structure enables selective attention	980
	to the most relevant events for the current dialogue	981
	turn, reducing the cognitive load and information	982
	overload that affects vanilla approaches. Besides,	983
	the hierarchical event graph preserves critical re-	984
	lationships between core and supporting events	985
	even as the dialogue extends, allowing the model	986
	to maintain coherence across longer time spans.	987
	<b>B.1.5 Qualitative Analysis of Very Long</b>	988
	<b>Dialogues</b>	989
	For very long dialogues (31+ turns), we ob-	990
	served that EventWeave responses frequently con-	991
	tained contextually appropriate references to ear-	992
	lier events. For example, in a 42-turn dialogue	993
	about travel planning, EventWeave correctly re-	994
	called and integrated a previously mentioned food	995
	allergy from turn 7 when discussing restaurant op-	996
	tions in turn 39, whereas the vanilla model com-	997
	pletely lost this context.	998
	This ability to selectively retrieve and integrate	999
	relevant historical events becomes increasingly	1000
	valuable as dialogue length increases. In dialogues	1001
	exceeding 30 turns, we found that EventWeave’s	1002
	structured event representation led to a 57.3 per-	1003
	centage point improvement in Event Coherence	1004
	compared to the vanilla approach.	1005
	<b>B.2 Results with Full Baselines</b>	1006
	Table 10 shows comparison experimental results	1007
	with other popular backbones including Qwen2.5-	1008
	14B, Qwen3-14B, o1, DeepSeek-V3, DeepSeek-	1009
	R1.	1010

Length Category	Number of Dialogues	Percentage
Short (5-10 turns)	358	42.3%
Medium (11-20 turns)	289	34.2%
Long (21-30 turns)	156	18.4%
Very Long (31+ turns)	43	5.1%

Table 8: Distribution of dialogues by length category in the Chronicle dataset.

Length Category	GPT-4o	EventWeave	Absolute Diff.	p-value
<i>GPT-4 Rating (1-10 scale)</i>				
Short (5-10)	5.2	6.5	+1.3	0.008*
Medium (11-20)	4.7	7.8	+3.1	<0.001*
Long (21-30)	4.2	8.4	+4.2	<0.001*
Very Long (31+)	3.8	9.1	+5.3	<0.001*
<i>Event Coherence (EC) Score (%)</i>				
Short (5-10)	24.8	40.2	+15.4	0.003*
Medium (11-20)	22.1	56.7	+34.6	<0.001*
Long (21-30)	18.7	64.5	+45.8	<0.001*
Very Long (31+)	15.3	72.6	+57.3	<0.001*

\* indicates statistical significance ( $p < 0.05$ )

Table 9: Performance comparison between GPT-4o (vanilla) and EventWeave across different dialogue lengths, showing absolute differences and statistical significance.

### B.3 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 11, 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

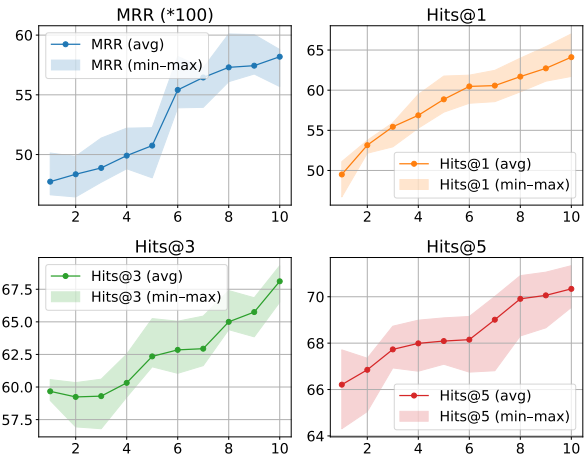


Figure 5: Event node retrieval performance with ICL on both Conversation Chronicle and MSC for LLaMA3.

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 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

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)	<b>30.7</b>	<b>44.2</b>	<b>0.65</b>	<b>7.6</b>	<b>33.4</b>	<b>47.0</b>	<b>0.68</b>	<b>8.0</b>	<b>65.8</b>	<b>53.7</b>	<b>55.9</b>	<b>58.7</b>
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)	<b>31.4</b>	<b>45.0</b>	<b>0.66</b>	<b>7.8</b>	<b>34.2</b>	<b>47.6</b>	<b>0.69</b>	<b>8.2</b>	<b>67.2</b>	<b>54.9</b>	<b>57.1</b>	<b>60.0</b>
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)	<b>33.0</b>	<b>46.6</b>	<b>0.68</b>	<b>8.1</b>	<b>35.8</b>	<b>49.1</b>	<b>0.71</b>	<b>8.5</b>	<b>70.3</b>	<b>57.6</b>	<b>59.7</b>	<b>62.7</b>
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)	<b>31.8</b>	<b>45.3</b>	<b>0.66</b>	<b>7.8</b>	<b>34.5</b>	<b>47.9</b>	<b>0.69</b>	<b>8.2</b>	<b>67.7</b>	<b>55.4</b>	<b>57.5</b>	<b>60.4</b>
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)	<b>32.4</b>	<b>46.0</b>	<b>0.67</b>	<b>8.0</b>	<b>35.2</b>	<b>48.6</b>	<b>0.70</b>	<b>8.4</b>	<b>68.8</b>	<b>56.5</b>	<b>58.6</b>	<b>61.5</b>

Table 10: 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
<b>Few-shot Setting (3 examples)</b>												
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)	<b>24.5</b>	<b>39.0</b>	<b>0.61</b>	<b>6.3</b>	<b>28.2</b>	<b>43.7</b>	<b>0.64</b>	<b>6.7</b>	<b>42.6</b>	<b>34.5</b>	<b>36.2</b>	<b>38.9</b>
<b>Few-shot Setting (5 examples)</b>												
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)	<b>27.6</b>	<b>41.9</b>	<b>0.64</b>	<b>7.0</b>	<b>31.2</b>	<b>46.0</b>	<b>0.67</b>	<b>7.4</b>	<b>53.8</b>	<b>44.0</b>	<b>46.1</b>	<b>48.9</b>

Table 11: Experimental results on few-shot settings.

may mislead the model’s baseline retrieval, suggesting that maintaining data distribution consistency is crucial for optimal event retrieval.

## B.4 Further Ablation Study

To comprehensively evaluate our approach, we conducted additional ablation experiments examining

1045  
1046  
1047

1048  
1049  
1050

1051	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.	1103
1052		1104
1053		1105
1054		1106
1055		1107
1056		1108
1057		
1058	Table 12 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.	1109
1059		1110
1060		1111
1061		1112
1062		1113
1063		1114
1064		1115
1065	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 systems. 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 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.	1116
1066		1117
1067		1118
1068		
1069		1119
1070		1120
1071		1121
1072		1122
1073		1123
1074		1124
1075		1125
1076		1126
1077		1127
1078		1128
1079		1129
1080		1130
1081		
1082		1131
1083		1132
1084		1133
1085		1134
1086		1135
1087		1136
1088		1137
1089		1138
1090	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.	1139
1091		1140
1092		
1093		1141
1094		1142
1095		1143
1096		1144
1097		1145
1098		1146
1099		1147
1100	These results empirically validate our multi-perspective query mechanism’s design and demonstrate that each perspective contributes uniquely to	1148
1101		1149
1102		1150
	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>B.5 Impacts of Graph Expansion</b>	
	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.	
	<b>B.5.1 Storage Efficiency</b>	
	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>B.5.2 Computational Considerations</b>	
	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>B.5.3 Scalability Analysis</b>	
	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 13, EventWeave without pruning demonstrates slightly higher resource usage than baseline methods, with a memory footprint	

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 12: 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).

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 13: 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.

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.6 Impacts of Threshold in EventWeave

We analyze three thresholds in our framework to see how each threshold affects different metrics. The threshold  $\delta$  groups events into clusters if their embeddings pass a similarity check. The threshold  $\theta$  creates reasoning links between events if they exceed a similarity value. The threshold  $\beta$  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.7 Analysis of Relationship Type Coverage

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

**Edge Type Distribution.** Table 15 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).

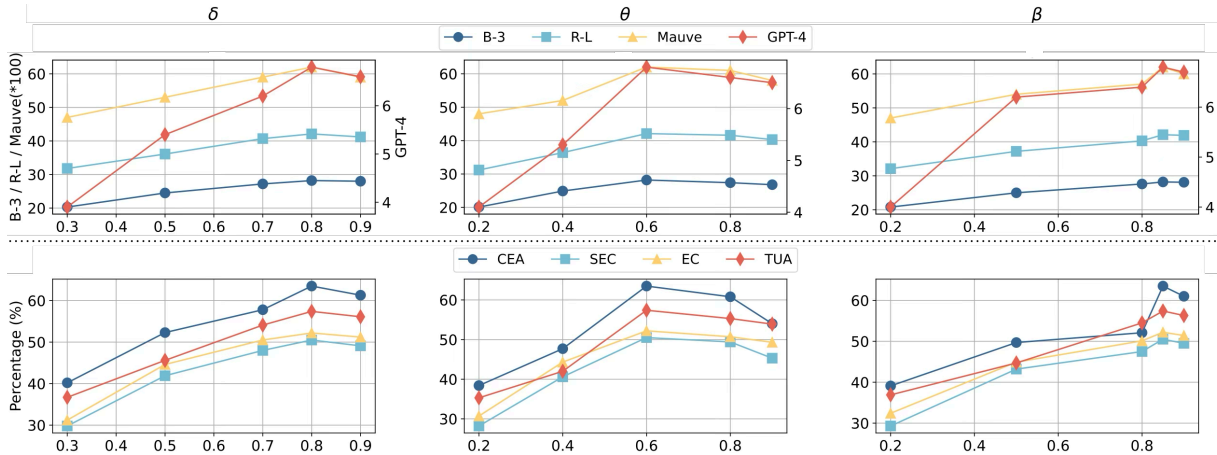


Figure 6: Experimental results about how each threshold affects different metrics.

Edge Type Configuration	CEA $\uparrow$	SEC $\uparrow$	EC $\uparrow$	TUA $\uparrow$
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 14: 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 15: 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.

**Ablation Studies.** To empirically validate the sufficiency of our proposed typology, we conducted ablation experiments with different edge type configurations, as shown in Table 14. 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

1218  
1219  
1220  
1221

1222	full model. This confirms that each relationship	1269
1223	type contributes meaningfully to the model’s un-	1270
1224	derstanding of dialogue history.	1271
1225	<b>Additional Relationship Types.</b> We also ex-	1272
1226	plored potential additional relationship types, in-	1273
1227	cluding causal relationships (e.g., “advice” → “re-	1274
1228	lationship fix”), emotional connections (e.g., “frus-	1275
1229	tration” → “apology”), and rhetorical devices (e.g.,	1276
1230	irony, sarcasm). When incorporating these addi-	1277
1231	tional types into EventWeave, we observed only	
1232	marginal improvements (+0.9% on average across	<b>D.1 Evaluation Protocol</b>
1233	metrics), suggesting that our three primary types	1278
1234	sufficiently capture most critical dialogue dynam-	1279
1235	ics.	1280
1236	The results demonstrate that while specialized	1281
1237	relationship types may offer slight improvements	1282
1238	in certain contexts, our proposed three-type frame-	1283
1239	work provides a robust and comprehensive founda-	1284
1240	tion for modeling dialogue history. This aligns	1285
1241	with previous findings that most conversational	1286
1242	structures can be effectively represented through	1287
1243	sequential progression, temporal references, and	
1244	reasoning connections. We acknowledge that cer-	<b>D.2 Results</b>
1245	tain specialized contexts (e.g., highly rhetorical	1288
1246	discourse) might benefit from additional relation-	1289
1247	ship modeling, which we discuss as a limitation	1290
1248	and direction for future work.	1291
1249		1292
1250	<b>C Prompts</b>	1293
1251	<b>C.1 GPT-as-the-Judge</b>	1294
1252	Figure 7 shows the example of the gpt-as-the-judge	1295
1253	prompts.	1296
1254		1297
1255	<b>C.2 Question Generation</b>	1298
1256	We use a separate LLM to generate questions for	
1257	the event-based evaluation. There are four metrics	<b>E Case Studies</b>
1258	for the evaluation: Core Event Awareness (CEA),	1299
1259	Supporting Event Coverage (SEC), Event Coher-	1300
1260	ence (EC), and Temporal Update Accuracy (TUA).	1301
1261	And we use different question-generation prompts	1302
1262	for each metric. The prompts are shown in Figure	1303
1263	8.	1304
1264		1305
1265	<b>C.3 Inference Prompts</b>	1306
1266	All models included in our experiments were used	1307
1267	the same inference prompts to ensure a fair com-	1308
1268	parison. The prompts were shown in Figure 9.	1309
		1310
		1311
		1312
		1313
		1314
		1315
		1316

# GPT-AS-THE-JUDGE Prompts
<p><b>Task:</b> 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.</p> <hr/> <p><b>Context Information:</b> (USER INSTRUCTION) <b>Example:</b> (OPTIONAL EXAMPLES PROVIDE TO MODEL) <b>Query:</b> (QUERY) <b>Required Answer Format:</b> (OPTIONAL ANSWER FORMAT)</p>

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
+ <b>EventWeave (Ours)</b>	<b>8.7</b>	<b>8.5</b>	<b>8.4</b>	<b>8.6</b>	<b>8.6</b>

Table 16: 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.

1317 out hierarchical structure, the model fails to under-  
1318 stand relationships between events. Without adap-  
1319 tive node preservation, it cannot prioritize relevant  
1320 information. Without edge relation descriptions,  
1321 it misunderstands how events relate to each other.  
1322 Without event representation updates, it struggles  
1323 to track changes to plans and preferences over time.  
1324 This detailed comparison reinforces our quantita-  
1325 tive findings that the complete EventWeave frame-  
1326 work provides substantial benefits for tracking and  
1327 utilizing events in long-term dialogue contexts.

# QUESTION GENERATION Prompts
<p><b>For Core Event Awareness</b></p> <p>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?”</p>
<p><b>For Supporting Event Coverage</b></p> <p>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?”</p>
<p><b>For Event Coherence</b></p> <p>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?”</p>
<p><b>For Temporal Update Accuracy</b></p> <p>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?”</p>

Figure 8: Example for question generation Prompts.

# INFERENCE Prompts
<p><b>Task:</b> Please respond to the user’s following instruction and you may:</p> <ol style="list-style-type: none"> <li>(1) appropriately decline the user’s instruction and provide reasons.</li> <li>(2) point out the false assumptions in the user’s instruction.</li> <li>(3) suggest possible corrections for the false assumption to the user.</li> <li>(4) include any truthful knowledge that you think could help correct the user’s erroneous belief and enhance the authenticity of the response.</li> </ol> <hr style="border-top: 1px dashed black;"/> <p><b>Context Information:</b> (USER INSTRUCTION)</p> <p><b>Example:</b> (OPTIONAL EXAMPLES PROVIDE TO MODEL)</p> <p><b>Query:</b> (QUERY)</p> <p><b>Required Answer Format:</b> (OPTIONAL ANSWER FORMAT)</p>

Figure 9: Example for Inference Prompts.

# Guidelines for Human Evaluation
<p><b>Overview.</b> 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.</p> <p><b>1. Coherence:</b> 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)</p> <p><b>2. Correctness:</b> 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)</p> <p><b>3. Style:</b> 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)</p> <p><b>4. Helpfulness:</b> 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)</p>

Figure 10: Annotator Guidance.

<b>Session 1</b>	
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."	
<b>Session 2</b>	
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."	
<b>Session 3</b>	
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."	
<b>Session 4</b>	
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."	
<b>Session 5</b>	
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."	
<b>Core Events</b>	<b>Supporting Events</b>
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))
<b>Event Update Flow:</b>	
<b>Session 1:</b> (Neighbors A, Completes, Boat) → (Neighbors A, Mentions, Future project)	
<b>Session 2:</b> (Neighbors A, Shares, Gratitude) → (Neighbors B, Reciprocates, Sentiment of emotional bonding)	
<b>Session 3:</b> (Neighbors A, Expresses, Emotional distress) → (Neighbors B, Suggests, Special gesture) (Neighbors A, Expresses, Emotional distress) → (Neighbors B, Encourages, Communication)	
<b>Session 4:</b> (Neighbors A, Acknowledges, Risk) → (Neighbors A, Expresses, Gratitude) (Neighbors A, Acknowledges, Feeling of accomplishment) → (Neighbors A, Mentions, Boat)	
<b>Session 5:</b> (Neighbors A, Acknowledges, Job offer) → (Neighbors A, Suggests, Boat trip) (Neighbors B, Reinforces, Friendship) → (Neighbors A, Expresses, Enthusiasm)	

Table 17: Case from Conversation Chronicle.

<b>Context</b>	<p><b>Dialogue History Snippet (From Session 3):</b>  <i>Previous sessions contained discussions about:</i></p> <ul style="list-style-type: none"> <li>• User planning a trip to Japan in April</li> <li>• Initial plan to visit Tokyo and Kyoto</li> <li>• User mentioned food allergies (seafood)</li> <li>• Discussion about budgeting around \$3000</li> <li>• Friend recommended visiting Mt. Fuji</li> </ul> <p><b>Current Turn:</b> User: I finally booked my flights to Japan! April 15-30th.  Assistant: That’s exciting! So you’ll be in Japan for two weeks. Did you decide on Tokyo and Kyoto as we discussed before?  User: Yes, but I’m thinking of changing my plans a bit. I still want to do Tokyo, but instead of Kyoto, I’m considering Osaka because of that food festival you mentioned last time. Do you remember which dates it was?</p>
<b>Full EventWeave</b>	I remember! The Osaka food festival runs from April 18-22, which would work perfectly with your travel dates. Since you have seafood allergies, I’d recommend focusing on the street food section in the eastern area of the festival where they have more meat and vegetarian options. This change from Kyoto to Osaka still keeps you within your \$3000 budget we discussed, and you can still do that day trip to Mt. Fuji from Tokyo as your friend recommended.
<b>w/o Hierarchical Structure</b>	The food festival in Osaka is from April 18-22. That should work with your dates. There are many food options there. Tokyo is also good to visit. Your budget is \$3000 if I remember correctly. <i>[Fails to connect related events across sessions; presents facts flatly without understanding core/supporting relationships]</i>
<b>w/o Adaptive Node Preservation</b>	The Osaka food festival is April 18-22. You should be careful because of your shellfish allergy. Also, I think there was something about Kyoto temples being closed in late April. And didn’t you say something about wanting to see cherry blossoms? The Tokyo hotels might be expensive during that time. Your friend mentioned Mt. Fuji too. <i>[Overloads with potentially irrelevant information; can’t prioritize which events matter now]</i>
<b>w/o Edge Relation Descriptions</b>	The food festival in Osaka is April 18-22, which conflicts with your plan to visit Kyoto. But you could still visit Tokyo from April 15-30. Your budget is \$3000, which means you might not be able to do both Osaka and Tokyo. <i>[Misunderstands the relationship between events; treats the Kyoto plan as still active rather than being replaced]</i>
<b>w/o Event Representation Updates</b>	The food festival in Osaka is April 18-22. That would work with your schedule. But I thought you were planning to visit Kyoto? Your original plan was Tokyo and Kyoto. The festival in Osaka is good too though. Your budget of \$3000 should cover either option. <i>[Contains contradictory information because it can’t properly update the evolving travel plan]</i>

Table 18: Case study comparing responses from different EventWeave variants on a multi-session conversation example from the LoCoMo dataset. The example demonstrates how each component contributes to generating contextually appropriate responses that correctly track and update events across conversation sessions.

Processing Step	Details
1. Event Extraction	<p><b>Given this dialogue snippet:</b></p> <p>User: "I finally booked my flights to Japan! April 15-30th."  Assistant: "That's exciting! So you'll be in Japan for two weeks. Did you decide on Tokyo and Kyoto as we discussed before?"  User: "Yes, but I'm thinking of changing my plans a bit. I still want to do Tokyo, but instead of Kyoto, I'm considering Osaka because of that food festival you mentioned last time. Do you remember which dates it was?"</p> <p><b>Our event extraction module identifies these key events:</b></p> <p><i>Core event:</i> "Trip to Japan (April 15-30)"  <i>Supporting events:</i></p> <ul style="list-style-type: none"> <li>• "Originally planned Tokyo and Kyoto"</li> <li>• "Considering Osaka instead of Kyoto"</li> <li>• "Interest in food festival (dates unknown)"</li> <li>• "Previous mention of food festival"</li> </ul>
2. Graph Build	<p><b>Our system constructs an event graph with:</b></p> <p><i>Event Nodes:</i></p> <ul style="list-style-type: none"> <li>• E1: [Trip to Japan] (Core) - Embedding vector capturing trip dates, duration</li> <li>• E2: [Tokyo and Kyoto plan] (Supporting) - Contains original itinerary</li> <li>• E3: [Plan modification] (Supporting) - Contains change from Kyoto to Osaka</li> <li>• E4: [Food festival interest] (Supporting) - References previous conversation</li> </ul> <p><i>Event Relationships:</i></p> <ul style="list-style-type: none"> <li>• Sequential Edge: E1 → E2 → E3 → E4 (conversation flow)</li> <li>• Reasoning Link: E2 ↔ E3 (plan modification relationship)</li> <li>• Reasoning Link: E3 ↔ E4 (reason for plan change)</li> </ul>
3. Hierarchical	<p><b>The system connects these new events to previously stored events:</b></p> <p><i>From previous sessions:</i></p> <ul style="list-style-type: none"> <li>• "User has seafood allergies" (E5)</li> <li>• "Budget around \$3000" (E6)</li> <li>• "Friend recommended Mt. Fuji" (E7)</li> <li>• "Food festival in Osaka (April 18-22)" (E8)</li> </ul> <p><i>Interlayer connections established:</i></p> <ul style="list-style-type: none"> <li>• E4 ↔ E8 (current question directly relates to previous food festival information)</li> <li>• E3 ↔ E6 (plan change may impact budget considerations)</li> <li>• E3 ↔ E5 (destination change relates to food/allergy considerations)</li> </ul>
4. Responding	<p><b>When processing the query about festival dates, our multi-perspective retrieval:</b></p> <ul style="list-style-type: none"> <li>• Identifies E4 as highly relevant to current turn</li> <li>• Traces connection to E8 for festival dates</li> <li>• Recognizes E3, E5, E6 as contextually important supporting information</li> </ul> <p><b>The generated response integrates this information naturally:</b></p> <p><i>"I remember! The Osaka food festival runs from April 18-22, which would work perfectly with your travel dates. Since you have seafood allergies, I'd recommend focusing on the street food section in the eastern area of the festival where they have more meat and vegetarian options. This change from Kyoto to Osaka still keeps you within your \$3000 budget we discussed, and you can still do that day trip to Mt. Fuji from Tokyo as your friend recommended."</i></p>

Table 19: Step-by-step illustration of EventWeave's processing pipeline on a multi-turn dialogue example. The example demonstrates how the model extracts events, constructs a graph representation, integrates with previously stored events, and generates a contextually appropriate response.