COALMEM: EVIDENCE COALITION SELECTION FOR MULTI-PARTY CONVERSATIONAL MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Answering queries over long-term multi-party conversational memory often requires combining evidence distributed across participants and discussion phases. Existing memory retrieval approaches largely optimize individual message relevance, yet the resulting context may still omit critical evidence needed to answer the query. Meanwhile, generic diversity does not guarantee that different messages contribute complementary query-relevant information. We therefore formulate multi-party conversational memory retrieval as *evidence coalition selection* and propose COALMEM, a training-free framework for selecting a compact set of messages with high joint evidence value. It constructs an evidence coalition graph over seed retrieved messages, where query–evidence edges capture relevance and evidence–evidence edges characterize both non-redundancy and query-conditioned complementarity. It then performs query-anchored coalition optimization to jointly select a compact set of messages that are individually relevant and collectively supportive of the query, without additional LLM calls during selection. Experiments on GroupMemBench across six tasks show that our method consistently outperforms retrieval-based, memory-based, and alternative evidence-selection methods.

1 INTRODUCTION

Large language models increasingly rely on external memory to support long-horizon interactions that extend beyond their active context windows. Recent memory systems have explored structured storage, dynamic organization, memory consolidation, and long-term retrieval for persistent agents and multi-session dialogue (Gutiérrez et al., 2024; Xu et al., 2026b; Chhikara et al., 2025; Latimer et al., 2025). This challenge becomes more pronounced in multi-party conversations (MPC), where information is produced by different participants across interleaved topics, channels, and discussion stages. Recent benchmarks such as GroupMemBench and EverMemBench further show that multi-party memory introduces difficulties beyond conventional dyadic dialogue, including cross-participant reasoning, evolving information, user-specific grounding, and temporal dependencies (Yang et al., 2026; Hu et al., 2026). Answering a query in such settings often requires recovering not a single relevant message, but a collection of messages that together provide the evidence needed for an accurate response.

Multi-party conversational memory therefore calls for a set-level view of retrieval. First, useful evidence is often distributed across participants and discussion stages, such that individual messages reveal only part of the information required by a query. Second, individual relevance does not imply collective sufficiency: several individually relevant messages may still fail to provide all the information needed for answering the query, while their simple accumulation can consume substantial context budget without providing proportional evidence value. Third, the usefulness of a message can depend on what other messages are selected with it, as different messages may provide complementary conditions, explanations, decisions, or contextual details. Recent studies of long-context memory and retrieval similarly suggest that retrieval quality is increasingly limited by evidence composition rather than by simply increasing the amount of retrieved context (Wang et al., 2025; Yang et al., 2026; Hu et al., 2026).

Existing conversational memory methods have primarily focused on how memories are extracted, organized, consolidated, updated, and retrieved (Gutiérrez et al., 2024; Xu et al., 2026b; Chhikara

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

et al., 2025; Latimer et al., 2025). At retrieval time, however, many approaches still fundamentally rely on the relevance of individual memory items to the query. Several individually relevant messages do not necessarily form a collectively sufficient evidence set; instead, they may introduce locally relevant but globally limited information, resulting in unnecessarily long and inefficient contexts. Diversity-aware retrieval provides a partial remedy by discouraging similar candidates, and recent RAG studies have shown the benefits of incorporating diversity into evidence selection (Wang et al., 2025). Yet generic diversity only captures whether evidence items differ, not whether their differences contribute complementary information required by the current query. Thus, there remains a gap between *individual relevance*, *generic diversity*, and the *collective evidence value* needed for multi-party conversational memory.

We formulate multi-party conversational memory retrieval as *evidence coalition selection*, where the goal is to identify a compact set of messages that collectively provides sufficient evidence for the query under a limited context budget. Based on this view, we propose CoalMem, a training-free and efficient framework for post-retrieval evidence selection in multi-party conversational memory. CoalMem first retrieves a candidate message pool using message content together with conversational metadata. It then constructs an evidence coalition graph, where query–evidence edges capture the relevance of individual messages to the query, while evidence–evidence edges model evidence non-redundancy and query-conditioned complementarity. These signals are combined to characterize the pairwise utility of selecting messages within the same evidence coalition. Inspired by constrained dominant-set optimization for graph-based joint selection, CoalMem jointly estimates coalition membership weights for all candidate messages and selects the highest-weighted ones as the final evidence set. The selected coalition is then provided to the downstream answer model, without requiring additional LLM calls during evidence selection.

We evaluate CoalMem on GroupMemBench (Yang et al., 2026), a recent benchmark designed specifically for long-term multi-party conversational memory. CoalMem consistently improves downstream answer accuracy over retrieval-based and memory-based baselines, and also outperforms alternative evidence-selection methods under matched candidate pools and evidence budgets. Our contributions are threefold:

- We propose CoalMem, a training-free and efficient framework that formulates evidence selection in MPC as a query-anchored coalition optimization problem under a fixed evidence budget, without additional LLM calls during selection.
- We develop an evidence coalition graph that integrates query–evidence relevance with pairwise coalition utility. By combining evidence non-redundancy with query-conditioned evidence complementarity, it accounts for both repeated content and the additional query-term coverage provided by message pairs.
- We demonstrate CoalMem's effectiveness on GroupMemBench against retrieval-based and memory-based baselines, as well as alternative selection methods under matched candidate pools and evidence budgets.

2 Related Work

Long-Term Conversational Memory. Long-term conversational memory enables LLM agents to retain and access information beyond their active context windows (Zhang et al., 2025a). Existing systems support this capability through hierarchical context management (Packer et al., 2023) and incremental fact extraction and updating (Chhikara et al., 2025). A-Mem organizes memories as dynamically linked notes (Xu et al., 2026b), while Hindsight integrates structured memory with retrieval and reflection (Latimer et al., 2025). To support access to relevant information, SeCom combines topical segmentation with compression (Pan et al., 2025), MemGAS supports query-adaptive retrieval across multiple granularities (Xu et al., 2026a), and RF-Mem switches between one-shot retrieval and iterative memory expansion according to retrieval uncertainty (Zhang et al., 2026). These studies advance how memories are organized, maintained, and accessed. In multi-party conversations, GroupMemBench further highlights additional challenges involving speaker attribution, role-dependent language, and interleaved discussions (Yang et al., 2026).

Structured Retrieval-Augmented Generation. Structured RAG organizes external knowledge through hierarchical or relational structures to support retrieval across text segments (Gao et al.,

2023). RAPTOR recursively clusters and summarizes text into a retrieval tree (Sarthi et al., 2024), while GraphRAG constructs hierarchical entity communities and their summaries (Edge et al., 2024). Other approaches exploit graph connectivity to retrieve associated evidence. HippoRAG and HippoRAG 2 use Personalized PageRank for cross-passage retrieval (Gutiérrez et al., 2024; 2025). BrowseNet links text chunks through named entities for query-guided exploration (SANTHANA KRISHNAN et al., 2026). LinearRAG constructs an entity–sentence–passage graph for two-stage retrieval (Zhuang et al., 2026), while EHRAG builds a hybrid hypergraph that captures structural co-occurrence and semantic relationships for diffusion-based retrieval (Song et al., 2026). COALMEM instead constructs an evidence coalition graph over retrieved messages, encoding query–evidence relevance and pairwise coalition utility to guide budget-constrained evidence selection rather than navigation through the source corpus.

Budget-Constrained Evidence Selection. Selecting evidence under a limited budget requires considering both candidate relevance and interactions among the retained items. Maximal Marginal Relevance (MMR) sequentially balances query relevance against similarity to previously selected documents (Carbonell & Goldstein, 1998; Zhang et al., 2025b; Wang et al., 2025; Khan et al., 2026). Determinantal point processes (DPPs) jointly model item quality and subset diversity through a kernel that discourages the co-selection of similar items (Kulesza & Taskar, 2012; Li et al., 2024; Sun et al., 2025; 2026). Recent work formulates evidence packing for multi-hop RAG as budgeted submodular maximization(Bala, 2026). Within this line of work, COALMEM combines evidence non-redundancy with an explicit pairwise measure of evidence complementarity. It integrates these pairwise utilities with query–evidence relevance through query-anchored coalition optimization.

3 PROBLEM FORMULATION

Let $\mathcal{H} = \{m_i\}_{i=1}^{L}$ denote a multi-party conversation history containing L messages. Each message $m_i = (c_i, z_i)$ consists of its textual content c_i and conversational metadata z_i, including speaker identity, role, channel, topic, and discussion phase. Given a query q, a retriever $\mathcal{R}$ returns a candidate set $\mathcal{D}_q = \mathcal{R}(q, \mathcal{H}) \subseteq \mathcal{H}$ with $|\mathcal{D}_q| = N$.

We formulate post-retrieval evidence selection as selecting an evidence coalition: a subset of candidate messages that collectively supports an accurate answer to q under a fixed message budget. Specifically, for a prescribed selection size $K \leq N$, an evidence selector f produces

$$\mathcal{S}_q = f(q, \mathcal{D}_q, K), \qquad \mathcal{S}_q \subseteq \mathcal{D}_q, \qquad |\mathcal{S}_q| = K. \tag{1}$$

The selected messages serve as the evidence context for a downstream answer model g, which generates $\hat{y}_q = g(q, \mathcal{S}_q)$. The objective is to retain the information needed for answering the query within the prescribed budget, with evidence quality assessed by downstream answer correctness.

4 METHOD

4.1 OVERVIEW

We propose COALMEM, a training-free framework for post-retrieval evidence selection in multi-party conversational memory. As illustrated in Figure 1, COALMEM retrieves seed evidence and constructs an evidence coalition graph over the query and candidate messages. Query–evidence edges capture relevance based on message content and conversational metadata, while evidence–evidence edges encode pairwise coalition utility through non-redundancy and query-conditioned complementarity. COALMEM performs query-anchored coalition optimization to jointly estimate coalition membership weights and selects the K highest-weighted messages for answer generation, without additional LLM calls during selection.

4.2 EVIDENCE COALITION GRAPH CONSTRUCTION

Given a query q, an initial retriever (e.g., BM25[1]) selects the top-N messages from the conversation history $\mathcal{H}$ as the initial candidate pool $\mathcal{D}_q$, which serves as the basis for constructing the evidence

[1]We also experiment with dense retrieval using Qwen-based embeddings. Empirically, the lightweight BM25 retriever achieves better downstream performance; we provide a detailed comparison in Section 5.3.

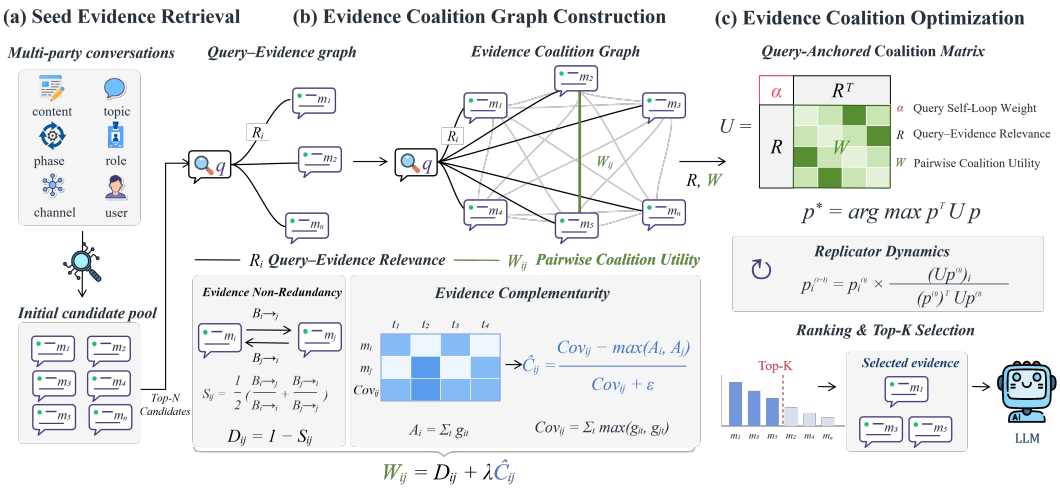

Figure 1: Overview of CoalMem.

coalition graph. Since message relevance in multi-party conversations depends not only on textual content but also on who produced the message and the discussion context in which it occurred, we augment each message with conversational metadata, including the speaker identity, participant role, channel, topic, and discussion phase. These metadata provide complementary cues about participant functions and the surrounding conversational structure, enabling the retriever to better distinguish messages with similar content but different roles or contextual significance.

To represent both individual relevance and pairwise coalition utility, CoalMem constructs a complete weighted graph $\mathcal{G}_q$ over the node set $\mathcal{V}_q = \{q\} \cup \mathcal{D}_q$. The query node connects to each candidate message through a relevance weight R_i, while each pair of distinct message nodes is connected by a utility weight W_{ij}. These two edge types encode how individual messages relate to the query and how they contribute jointly as evidence.

Query–Evidence Relevance. We derive query–evidence relevance from the retrieval scores assigned to the metadata-augmented message representations during candidate retrieval. For each $m_i \in \mathcal{D}_q$, let $\widetilde{R}_i$ denote its retrieval score with respect to query q. We normalize the scores within the candidate pool as $R_i = \widetilde{R}_i / \max_{m_j \in \mathcal{D}_q} \widetilde{R}_j$. The resulting score R_i is used as the weight of edge (q, m_i), preserving relative relevance among candidates while incorporating signals from both message content and conversational context.

Evidence Non-Redundancy. Query–evidence relevance alone does not account for content redundancy among candidate messages. We quantify this overlap through a symmetric measure of pairwise lexical similarity. Let $B_{i\to j}$ denote a directional content-matching score from c_i to c_j, with $B_{i\to i}$ denoting the corresponding self-match score. We normalize each directional score by its self-match and average the two directions to obtain a symmetric overlap score:

$$S_{ij} = \mathrm{clip}_{[0,1]}\left[\frac{1}{2}\left(\frac{B_{i\to j}}{B_{i\to i}} + \frac{B_{j\to i}}{B_{j\to j}}\right)\right]. \tag{2}$$

The self-normalization places the two directional scores on comparable scales, while their average yields a symmetric content-overlap score, $S_{ij} = S_{ji}$. We define $D_{ij} = 1 - S_{ij}$ as the evidence non-redundancy term, assigning higher values to message pairs with lower estimated overlap. In our implementation, the directional matching scores are computed using content-only BM25.

Evidence Complementarity. Evidence non-redundancy may assign high utility to lexically distinct messages even when their differences are unrelated to the query. We therefore model query-conditioned evidence complementarity through the additional query-term coverage provided by a message pair. Let $\mathcal{T}_q = (t_1, \ldots, t_T)$ denote the query terms obtained after text preprocessing, and let $g_{i\ell}$ quantify the contribution of term t_ℓ to the relevance of message content c_i to query q. The total query-term coverage of an individual message is $A_i = \sum_{\ell=1}^{T} g_{i\ell}$. For a candidate pair, we

retain the stronger message-level contribution for each query term:

$$\mathrm{Cov}_{ij} = \sum_{\ell=1}^{T} \max(g_{i\ell}, g_{j\ell})\,. \tag{3}$$

The term-wise maximum prevents a query term covered by both messages from receiving additive credit, while preserving contributions from terms covered more strongly by different messages. We define evidence complementarity as the difference between the pair's joint coverage and the larger individual coverage score, normalized by the joint coverage:

$$\widehat{C}_{ij} = \frac{\mathrm{Cov}_{ij} - \max(A_i, A_j)}{\mathrm{Cov}_{ij} + \epsilon}, \tag{4}$$

where $\epsilon > 0$ ensures numerical stability. Consequently, $\widehat{C}_{ij}$ is positive when the pair achieves greater query-term coverage than either message alone. In our implementation, $g_{i\ell}$ is the contribution of t_ℓ to the content-only BM25 score.

Pairwise Coalition Utility. We combine evidence non-redundancy and query-conditioned evidence complementarity to define the utility of jointly retaining messages m_i and m_j:

$$W_{ij} = D_{ij} + \lambda \widehat{C}_{ij}, \qquad i \neq j, \tag{5}$$

where $\lambda \geq 0$ controls the contribution of query-conditioned evidence complementarity, and we set $W_{ii} = 0$. The utility W_{ij} increases as the two messages exhibit less lexical overlap and receives an additional contribution when they cover complementary query terms. Since both components are symmetric, $W_{ij} = W_{ji}$.

4.3 Evidence Coalition Optimization

Given the evidence coalition graph $\mathcal{G}_q$, CoalMem formulates set-level evidence selection as a query-anchored coalition optimization. The query node serves as an explicit anchor, allowing query–evidence relevance and pairwise coalition utility to jointly determine candidate node weights.

Query-Anchored Coalition Formulation. We collect the query–evidence relevance weights in $\mathbf{R} = [R_1, \ldots, R_N]^\top$ and the pairwise coalition utility weights in $\mathbf{W} = [W_{ij}]_{i,j=1}^{N}$. We represent the evidence coalition graph using the query-anchored coalition matrix:

$$\mathbf{U} = \begin{bmatrix} \alpha & \mathbf{R}^\top \\ \mathbf{R} & \mathbf{W} \end{bmatrix}, \qquad \alpha = \lambda_{\max}(\mathbf{W}) + \epsilon, \tag{6}$$

where α is the query node's self-loop weight, $\lambda_{\max}(\mathbf{W})$ denotes the largest eigenvalue of $\mathbf{W}$, and $\epsilon = 10^{-3}$. This self-loop is intended to encourage the query node to remain active during optimization, allowing query–evidence relevance to guide selection alongside pairwise coalition utility. Appendix A explains the role of the query node in the optimization objective and the rationale for setting its self-loop weight.

Query-Anchored Coalition Optimization. Let $\mathbf{p} = (p_0, p_1, \ldots, p_N)^\top \in \mathbb{R}^{N+1}$ denote the node-weight vector, where p_0 corresponds to the query node and p_i to candidate message m_i for $i = 1, \ldots, N$. Given the query-anchored coalition matrix $\mathbf{U}$, we optimize the following quadratic objective over the probability simplex:

$$\mathbf{p}^\star = \arg\max_{\mathbf{p} \in \Delta} \mathbf{p}^\top \mathbf{U} \mathbf{p}, \qquad \Delta = \left\{ \mathbf{p} \in \mathbb{R}^{N+1} \;\middle|\; p_i \geq 0\ \forall i,\ \sum_{i=0}^{N} p_i = 1 \right\}. \tag{7}$$

Each candidate is evaluated in relation to the weighted coalition, so its weight reflects both its connection to the query and its pairwise coalition utility with the other candidates. A message with a lower retrieval score can therefore receive additional support through its utility with other highly weighted messages, rather than being evaluated solely by its individual relevance.

We use replicator dynamics to seek a stationary solution of the quadratic objective. Starting from the uniform initialization $\mathbf{p}^{(0)} = \frac{1}{N+1}\mathbf{1}$, the node weights are updated according to

$$p_i^{(t+1)} = p_i^{(t)} \frac{\left(\mathbf{U}\mathbf{p}^{(t)}\right)_i}{\left(\mathbf{p}^{(t)}\right)^\top \mathbf{U}\mathbf{p}^{(t)}}, \qquad i = 0, \ldots, N, \tag{8}$$

At each iteration, the dynamics shift weight toward nodes that exhibit stronger affinity with the current weighted coalition. Let $\widehat{\mathbf{p}}$ denote the converged node-weight vector. We then remove the query coordinate and select the K candidate messages with the largest weights as the final evidence set, $\mathcal{S}_q = \mathrm{TopK}_{m_i \in \mathcal{D}_q}(\widehat{p}_i)$.

Evidence-Grounded Answer Generation. The selected messages in $\mathcal{S}_q$ are arranged in descending order of their final node weights, with their original content and conversational metadata preserved. Together with the query q, this evidence context is provided to the answer model g, which is instructed to answer based on the supplied evidence and to abstain when it is insufficient.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Dataset and Evaluation. We evaluate COALMEM on GroupMemBench (Yang et al., 2026) and EverMemBench (Hu et al., 2026). GroupMemBench evaluates long-term memory in multi-party conversations with user-specific queries spanning multi-hop reasoning, knowledge update, term ambiguity, user-implicit reasoning, temporal reasoning, and abstention. EverMemBench evaluates long-term conversational memory under evolving, temporally extended interactions. We report category-wise answer accuracy and overall accuracy computed across all evaluated queries. More dataset statistics and the evaluation protocol are provided in Appendix B.1 and Appendix B.2.

Baselines. We compare COALMEM with a broad and competitive set of baselines spanning retrieval, memory, and diversity-aware evidence selection. Retrieval-based baselines include BM25 and dense retrieval, as well as graph-based methods LinearRAG (Zhuang et al., 2026) and EHRAG (Song et al., 2026). Memory-based baselines include HippoRAG (Gutiérrez et al., 2024), Hindsight (Latimer et al., 2025), and A-Mem (Xu et al., 2026b), which construct structured memory representations for subsequent retrieval and reasoning. We further compare against diversity-aware subset selection methods, including MMR (Carbonell & Goldstein, 1998; Khan et al., 2026), DPP (Kulesza & Taskar, 2012; Sun et al., 2026), and SEP (Bala, 2026), under the same candidate pool and evidence budget. This controlled setting isolates the contribution of COALMEM's selection strategy from differences in initial retrieval.

Implementation Details. We evaluate GPT-5 and DeepSeek-V4-Flash as answer models, each paired with a judge of the same model. GPT-5 uses `reasoning_effort=medium` for both answer generation and judging. For DeepSeek-V4-Flash, thinking is enabled for answer generation and disabled for judging. Both answer models use a generation limit of 4,096 tokens. We use BM25 as the initial retriever. Unless otherwise specified, we set the candidate pool size to $N = 48$, the evidence budget to $K = 10$, and the complementarity coefficient to $\lambda = 0.5$. Additional implementation details and baseline configurations are provided in Appendix B.3, while the prompt templates are included in Appendix B.4.

5.2 MAIN RESULTS

As shown in Table 1, COALMEM achieves the highest overall accuracy under both answer models, surpassing the strongest baseline in each setting by 11.71 percentage points with GPT-5 and 7.92 points with DeepSeek-V4-Flash. These results demonstrate the effectiveness of coalition-based evidence selection for multi-party conversational memory. COALMEM achieves these improvements without parameter training, memory rewriting, or additional LLM calls during evidence selection.

The category-level results show that COALMEM's improvements span key challenges in multi-party conversational memory. Under both answer models, COALMEM outperforms the baselines in multi-hop reasoning, knowledge update, user-implicit reasoning, and temporal reasoning. These tasks require combining information across messages or identifying evidence associated with the relevant participant or conversational state, making the gains consistent with COALMEM's query-aware, set-level selection design. COALMEM also achieves strong abstention accuracy, showing that these improvements are not accompanied by poorer performance on unsupported queries.

We further evaluate COALMEM on EverMemBench to assess its effectiveness beyond GroupMemBench. To limit evaluation costs, we use only DeepSeek-V4-Flash as the answer model for these ex-

Table 1: Accuracy (%) of memory-based and retrieval-based methods on GroupMemBench across six query categories. The Embedding baseline uses Qwen3-Embedding-8B. Within each model setting, the best available result in each column is shown in **bold**, and the second-best is underlined.

Method	Multi-Hop	Update	Ambiguity	Implicit	Temporal	Abstention	Overall
GPT-5							
BM25	40.11	25.23	14.15	40.82	54.94	77.98	43.22
Embedding	38.46	29.91	24.53	42.86	38.89	76.98	42.82
LinearRAG	28.02	23.36	16.98	40.82	22.84	**88.49**	36.78
EHRAG	28.02	21.50	16.98	42.86	25.31	83.45	36.24
HippoRAG	39.56	27.10	30.19	42.86	29.63	75.23	39.72
Hindsight	42.31	17.76	**37.74**	40.82	54.94	77.06	46.01
A-Mem	35.16	22.43	26.42	46.94	23.46	67.89	35.10
CoALMem	**52.75**	**42.99**	34.91	**67.35**	**61.73**	84.89	**57.72**
DeepSeek-V4-Flash							
BM25	43.96	27.10	14.15	46.94	57.41	87.77	48.59
Embedding	40.66	29.91	25.47	46.94	37.04	77.70	43.49
LinearRAG	34.62	27.10	17.92	42.86	34.57	**89.21**	41.88
EHRAG	35.71	22.43	16.98	44.90	37.65	**89.21**	42.15
HippoRAG	57.14	36.45	34.91	65.31	46.91	76.26	52.89
Hindsight	45.60	18.69	34.91	42.86	54.32	76.98	47.79
A-Mem	39.56	30.84	29.25	65.31	38.89	56.83	41.61
CoALMem	**63.19**	**37.38**	**37.74**	**67.35**	**63.58**	87.77	**60.81**

periments. As reported in Appendix C.1, CoALMem achieves the highest overall accuracy among the evaluated methods, further supporting the effectiveness of coalition-based evidence selection across different conversation structures and memory evaluation tasks.

We also conduct an efficiency analysis, with results reported in Appendix C.2. Compared with BM25, CoALMem delivers substantial accuracy gains with limited additional online latency. It also achieves higher accuracy and considerably lower online latency than dense retrieval and memory-based baselines, without the LLM calls required for memory construction. These results highlight CoALMem's favorable accuracy–efficiency trade-off.

5.3 Discussions

Table 2: Comparison of CoALMem with alternative evidence set selection methods on GroupMemBench. All methods use the same candidate pool and evidence budget.

Method	Multi-Hop	Update	Ambiguity	Implicit	Temporal	Abstention	Overall
MMR	56.04	33.64	32.08	63.27	58.64	84.17	55.70
DPP	57.69	31.78	34.91	61.22	59.88	82.73	56.11
SEP	37.91	11.21	25.47	38.78	32.10	84.89	39.86
Reranker	60.44	**40.19**	36.79	**69.39**	63.58	69.78	57.18
CoALMem	**63.19**	37.38	**37.74**	67.35	**63.58**	**87.77**	**60.81**

Comparison with Alternative Selection Methods. We compare CoALMem with MMR, DPP, SEP, and Qwen3-Reranker-4B using the same candidate pool and evidence budget. As shown in Table 2, CoALMem achieves the highest overall accuracy of 60.81%, outperforming both the alternative selection methods and the LLM-based reranker. With candidate availability and evidence budget held fixed, these results support the effectiveness of CoALMem's coalition-based formulation for evidence selection. CoALMem outperforms MMR, DPP, and SEP across all six query categories. Compared with the LLM-based reranker, it achieves higher accuracy in multi-hop reasoning and term ambiguity while matching its temporal reasoning performance. Its advantage is particularly pronounced in abstention (87.77% vs. 69.78%). Unlike the reranker, CoALMem requires no additional neural model inference during evidence selection. These findings support CoALMem's joint modeling of query–evidence relevance and pairwise coalition utility, suggesting that effective con-

Table 3: Ablation study of CoALMem on GroupMemBench. Comp and NR denote evidence complementarity and evidence non-redundancy, respectively. CoALMem (Greedy) replaces query-anchored coalition optimization with greedy selection.

Method	Multi-Hop	Update	Ambiguity	Implicit	Temporal	Abstention	Overall
CoALMem w/o Comp	57.69	30.84	35.85	63.27	59.26	85.61	56.64
CoALMem w/o NR	57.14	36.45	34.91	65.31	59.26	83.45	56.91
CoALMem (Greedy)	54.95	35.51	**39.62**	59.18	54.94	85.61	55.97
CoALMem	**63.19**	**37.38**	37.74	**67.35**	**63.58**	**87.77**	**60.81**

versational memory benefits not only from identifying relevant messages, but also from considering their contributions to the evidence set as a whole.

Ablation Study. Table 3 examines the contributions of pairwise coalition utility components and the selection procedure. CoALMem w/o Comp removes evidence complementarity, while CoALMem w/o NR removes evidence non-redundancy. CoALMem (Greedy) replaces query-anchored coalition optimization with greedy selection while retaining the same query–evidence relevance and pairwise coalition utility scores. Removing evidence complementarity reduces overall accuracy from 60.81% to 56.64%, with the largest decline in knowledge update. Removing non-redundancy lowers overall accuracy to 56.91%, with the largest decline in multi-hop reasoning. Both ablations reduce accuracy across all six query categories, supporting the benefit of combining evidence non-redundancy with complementarity rather than relying on either component alone. Replacing query-anchored coalition optimization with greedy selection further lowers overall accuracy to 55.97%, highlighting the benefit of evaluating candidates jointly within the evidence coalition.

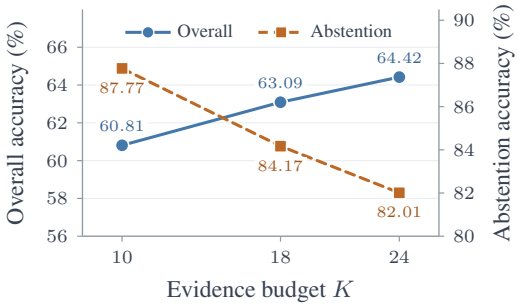

Figure 2: Effect of evidence budget K on CoALMem's performance on GroupMemBench. Overall accuracy (left axis) and abstention accuracy (right axis) are reported.

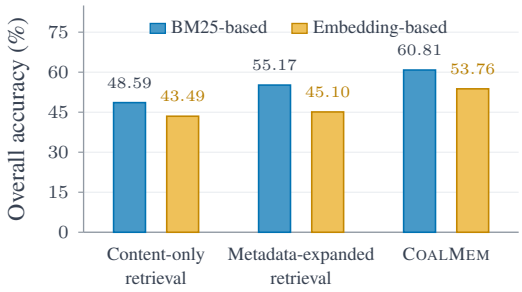

Figure 3: Comparison of BM25 and embedding-based variants on GroupMemBench across content-only retrieval, metadata-expanded retrieval, and CoALMem.

Effect of Evidence Budget. We examine the effect of the evidence budget K on CoALMem's performance. As shown in Figure 2, overall accuracy improves as K increases within the evaluated range, suggesting that CoALMem can benefit from additional evidence. Abstention accuracy (right axis) exhibits a different trend: it decreases from 87.77% at $K = 10$ to 82.01% at $K = 24$. A plausible explanation is that a larger evidence set introduces more context that appears relevant but does not establish an answer, potentially encouraging the answer model to respond rather than abstain. These results suggest that increasing evidence quantity does not improve performance uniformly across query types. The observed trade-off highlights the importance of the selected evidence coalition, rather than treating a larger evidence budget as sufficient for better performance.

Effect of Graph Sparsification. We examine two sparsification schemes while keeping the candidate pool and evidence budget fixed. Relevance-guided sparsification retains the h highest-weighted query–evidence edges ($h \in \{36, 24, 12\}$), with the full graph ($h = 48$) serving as the reference. Structure-guided sparsification retains evidence–evidence edges only between messages sharing the same scope, linked by a direct reply, or temporally adjacent across scopes. The two schemes are evaluated separately, leaving the other edge type unchanged.

Table 4: Effect of graph sparsification on CoalMem's accuracy (%) on GroupMemBench. Relevance-guided sparsification retains the h highest-weighted query–evidence edges, while structure-guided sparsification restricts evidence–evidence edges to those supported by shared scope membership, direct replies, or cross-scope temporal adjacency. Each scheme modifies only one edge type, leaving the other unchanged, with $N = 48$ candidate messages and evidence budget $K = 10$.

Variant	Multi-Hop	Update	Ambiguity	Implicit	Temporal	Abstention	Overall
CoalMem (Full graph)	63.19	37.38	37.74	67.35	63.58	87.77	60.81
Relevance-guided ($h = 36$)	60.99	36.45	33.96	67.35	59.26	87.05	58.52
Relevance-guided ($h = 24$)	59.89	37.38	35.85	65.31	58.02	82.73	57.45
Relevance-guided ($h = 12$)	58.79	37.38	33.96	65.31	59.26	82.73	57.18
Structure-guided	57.14	39.25	32.08	67.35	48.15	82.73	54.50

As shown in Table 4, relevance-guided sparsification reduces overall accuracy relative to the full graph, but decreasing the number of retained query edges from 36 to 12 incurs only limited additional degradation. One possible explanation is that the strongest query connections continue to guide selection, while messages without direct query edges can still contribute through their pairwise coalition utility with query-connected messages. Moreover, graph optimization operates on the top-N messages retrieved for the query, rather than the entire conversation history. Query relevance is therefore incorporated into candidate selection before graph optimization, which evaluates evidence combinations within an already filtered pool. This relevance filtering may help limit the additional performance loss when fewer query connections are retained.

Structure-guided sparsification causes a larger decline, reducing overall accuracy from 60.81% to 54.50%. In multi-party conversations, evidence relevant to the same query may be distributed across participants, channels, and discussion phases. Shared scopes, direct replies, and temporal adjacency describe conversational organization, but do not necessarily capture which messages jointly support an answer. Using these relations as hard connectivity constraints can therefore exclude useful interactions between structurally distant messages. These results support modeling pairwise coalition utility across all candidate pairs rather than restricting message interactions to these relations.

Effect of Seed Retrieval Methods. We compare BM25-based and embedding-based variants for content-only retrieval, metadata-expanded retrieval, and CoalMem. For the embedding-based CoalMem variant, we replace candidate retrieval, query–evidence relevance scoring, and evidence non-redundancy estimation with embedding-based counterparts, while retaining BM25-based evidence complementarity. As shown in Figure 3, BM25-based variants consistently outperform their embedding-based counterparts across all three evaluated settings. In this comparison, embedding-based matching may capture broader passage semantics while overlooking salient entities and fine-grained details. This potential limitation is relevant to multi-party conversations, where messages sharing a topic may concern different participants, entities, or discussion phases. For such queries, broad semantic similarity alone may not adequately distinguish relevant evidence. CoalMem's lexical formulation uses explicit term matching to assess query relevance and content overlap to estimate non-redundancy. These results support the use of lexical matching for retrieval and evidence scoring on GroupMemBench.

6 Conclusion

We presented CoalMem, a training-free framework for post-retrieval evidence selection in multi-party conversations. CoalMem constructs an evidence coalition graph that integrates query–evidence relevance with pairwise coalition utility, distinguishing evidence non-redundancy from query-conditioned evidence complementarity. Through query-anchored coalition optimization, it jointly estimates message weights and selects the highest-weighted messages under a fixed evidence budget, without additional LLM calls during selection. Experiments on GroupMemBench demonstrate improved answer accuracy over retrieval-based and memory-based baselines, as well as alternative selection methods under matched candidate pools and evidence budgets.

AI USE STATEMENT

We used generative AI tools to assist with coding during this work, supporting the development and refinement of research software. The authors retain full responsibility for the implementation, its correctness, and the results reported in this paper.

REPRODUCIBILITY STATEMENT

Sections 4.2 and 4.3 describe the evidence coalition graph and query-anchored coalition optimization. Appendix B.3 provides implementation details, hyperparameters, and baseline configurations. Appendices B.1, B.2, and B.4 document the datasets, evaluation protocol, and prompts for answer generation and evaluation. We will publicly release our code, including preprocessing, evidence selection, and evaluation scripts, after the review process.

REFERENCES

Ananto Nayan Bala. What survives into context: A diagnostic for budget-constrained multi-hop rag and when submodular evidence packing improves it. *arXiv preprint arXiv:2607.00725*, 2026.

Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 335–336. Association for Computing Machinery, 1998. doi: 10.1145/290941.291025.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

Bernal J Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. Hipporag: Neurobiologically inspired long-term memory for large language models. *Advances in neural information processing systems*, 37:59532–59569, 2024.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. From rag to memory: Non-parametric continual learning for large language models. *arXiv preprint arXiv:2502.14802*, 2025.

Chuanrui Hu, Tong Li, Xingze Gao, Hongda Chen, Yi Bai, Dannong Xu, Tianwei Lin, Xiaohong Li, Yunyun Han, Jian Pei, et al. Evaluating long-horizon memory for multi-party collaborative dialogues. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 9082–9090, 2026.

Saadat Hasan Khan, Spencer Hong, Jingyu Wu, Kevin Lybarger, Youbing Yin, Erin Babinsky, and Daben Liu. Df-rag: Query-aware diversity for retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 2873–2894, 2026.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2–3):123–286, 2012. doi: 10.1561/2200000044.

Chris Latimer, Nicoló Boschi, Andrew Neeser, Chris Bartholomew, Gaurav Srivastava, Xuan Wang, and Naren Ramakrishnan. Hindsight is 20/20: Building agent memory that retains, recalls, and reflects. *arXiv preprint arXiv:2512.12818*, 2025.

Jiatao Li, Xinyu Hu, and Xiaojun Wan. Smart-rag: Selection using determinantal matrices for augmented retrieval. *arXiv preprint arXiv:2409.13992*, 2024.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G Patil, Ion Stoica, and Joseph E Gonzalez. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

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

PAVAN KUMAR SANTHANA KRISHNAN, Kiran Kumar Nakka, C Reddy, Divyateja Pasupuleti, Prakhar Agarwal, Harpinder Singh, Anshu Avinash, and Nirav Bhatt. Browsenet: Graph-based associative memory for contextual information retrieval. In *International Conference on Learning Representations*, volume 2026, pp. 70597–70619, 2026.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher Manning. Raptor: Recursive abstractive processing for tree-organized retrieval. In *International Conference on Learning Representations*, volume 2024, pp. 32628–32649, 2024.

Yifan Song, Xingjian Tao, Zhicheng Yang, Yihong Luo, and Jing Tang. Ehrag: Bridging semantic gaps in lightweight graphrag via hybrid hypergraph construction and retrieval. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 24638–24652, 2026.

Hui Sun, Shiyin Lu, Huanyu Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Ming Li. Mdp 3: a training-free approach for list-wise frame selection in video-llms. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24090–24101. IEEE, 2025.

Xun Sun, Baiheng Xie, Li Huang, and Qiang Gao. Scaling dpps for rag: Density meets diversity. *arXiv preprint arXiv:2604.03240*, 2026.

Zhichao Wang, Bin Bi, Yanqi Luo, Sitaram Asur, and Claire Na Cheng. Diversity enhances an llm's performance in rag and long-context task. *arXiv preprint arXiv:2502.09017*, 2025.

Derong Xu, Yi Wen, Pengyue Jia, Yingyi Zhang, Wenlin Zhang, Yichao Wang, Huifeng Guo, Ruiming Tang, Xiangyu Zhao, Enhong Chen, et al. From single to multi-granularity: Toward long-term memory association and selection of conversational agents. In *International Conference on Learning Representations*, volume 2026, pp. 91581–91617, 2026a.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. *Advances in Neural Information Processing Systems*, 38:17577–17604, 2026b.

Jingbo Yang, Kwei-Herng Lai, Xiaowen Wang, Shiyu Chang, Yaar Harari, and Evgeniy Gabrilovich. Groupmembench: Benchmarking llm agent memory in multi-party conversations. *arXiv preprint arXiv:2605.14498*, 2026.

Yingyi Zhang, Junyi Li, Wenlin Zhang, Pengyue Jia, Xianneng Li, Yichao Wang, Derong Xu, Yi Wen, Huifeng Guo, Yong Liu, et al. Evoking user memory: Personalizing llm via recollection-familiarity adaptive retrieval. In *International Conference on Learning Representations*, volume 2026, pp. 54178–54213, 2026.

Zeyu Zhang, Quanyu Dai, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. A survey on the memory mechanism of large language model-based agents. *ACM Transactions on Information Systems*, 43(6):1–47, 2025a.

Zichen Zhang, Kunlong Zhang, Hongwei Ruan, and Yiming Luo. A comprehensive evaluation of transformer-based question answering models and rag-enhanced design. *arXiv preprint arXiv:2509.21845*, 2025b.

Luyao Zhuang, Shengyuan Chen, Yilin Xiao, Huachi Zhou, Yujing Zhang, Hao Chen, Qinggang Zhang, and Xiao Huang. Linearrag: Linear graph retrieval augmented generation on large-scale corpora. In *International Conference on Learning Representations*, volume 2026, pp. 147053–147075, 2026.

A Query Anchoring in Coalition Optimization

The query node allows query–evidence relevance and pairwise coalition utility to jointly influence evidence selection. Let $\mathbf{p} = [p_0; \mathbf{x}]$, where p_0 denotes the query node's weight and $\mathbf{x}$ contains the candidate message weights. The objective expands as

$$\mathbf{p}^\top \mathbf{U}\mathbf{p} = \alpha p_0^2 + 2p_0 \mathbf{R}^\top \mathbf{x} + \mathbf{x}^\top \mathbf{W}\mathbf{x}. \tag{9}$$

The cross term incorporates direct query relevance, while the final term captures pairwise coalition utility. Specifically, the replicator update for candidate m_i depends on

$$(\mathbf{U}\mathbf{p})_i = p_0 R_i + \sum_{j=1}^{N} W_{ij} x_j, \qquad i = 1, \ldots, N. \tag{10}$$

When $p_0 > 0$, each candidate receives a direct relevance contribution from the query node alongside its utility with the current weighted evidence set. The query node therefore serves as a shared relevance anchor within coalition optimization.

The self-loop adds αp_0^2 to the objective, rewarding weight assigned to the query node. It is introduced to encourage query participation during optimization, while the query–evidence edges convey the relevance signal to individual candidates.

We set $\alpha = \lambda_{\max}(\mathbf{W}) + \epsilon$ to determine the self-loop weight from the spectral scale of the candidate utility matrix. Since $\mathbf{W}$ is symmetric, $\mathbf{x}^\top \mathbf{W}\mathbf{x} \leq \lambda_{\max}(\mathbf{W})\|\mathbf{x}\|_2^2$. The largest eigenvalue thus provides an upper bound on pairwise utility per unit squared weight, making it a graph-dependent reference scale for the self-loop.

B Additional Experimental Details

B.1 Dataset Statistics

GroupMemBench (Yang et al., 2026) comprises synthetic multi-party conversations across four domains: Finance, Technology, Healthcare, and Manufacturing. Each domain contains 30,000 messages, yielding 120,000 messages in total. The conversations are organized into project channels, topics, and discussion phases, with participants exhibiting distinct roles and communication styles. Each evaluation question is associated with a specific asking user, making speaker attribution and conversational context relevant to interpreting the requested information. The evaluation set contains 745 questions spanning multi-hop reasoning, knowledge update, term ambiguity, user-implicit reasoning, temporal reasoning, and abstention. Table 5 summarizes the corpus size and question distribution across domains.

Table 5: Question distribution across the four domains of GroupMemBench.

Question type	**Finance**	**Technology**	**Healthcare**	**Manufacturing**	**Total**
Multi-hop	48	41	48	45	182
Knowledge update	32	36	17	22	107
Term ambiguity	45	43	7	11	106
User-implicit	15	28	2	4	49
Temporal	45	37	37	43	162
Abstention	29	28	34	48	139
Total	214	213	145	173	745

EverMemBench (Hu et al., 2026) contains 51,023 synthetic dialogue turns totaling approximately 4.2 million tokens, with 170 participants across five domains: technology, operations, marketing, financial services, and governance. The conversations span one year and are organized into interconnected group chats, featuring interleaved topics, evolving decisions, and participants with distinct roles and communication styles. The evaluation set comprises 2,400 questions across three dimensions: fine-grained recall (762), memory awareness (1,097), and profile understanding (541). These

dimensions assess the retrieval of specific facts, the use of past information and constraints, and the understanding of participant characteristics, respectively. Fine-grained recall uses open-ended questions, whereas memory awareness and profile understanding use multiple-choice questions. Table 6 summarizes the evaluation dimensions, constituent tasks, and question counts.

Table 6: Evaluation dimensions, tasks, and question distribution of EverMemBench.

Evaluation dimension	**Tasks**	**Questions**
Fine-grained recall	Single-hop retrieval, multi-hop trajectory, temporal duration	762
Memory awareness	Constraint, proactivity, update	1,097
Profile understanding	Style, skill, role	541
Total		2,400

B.2 Evaluation Protocol

We use GPT-5 and DeepSeek-V4-Flash as answer models, with each generated answer evaluated against the reference answer by the same model in a separate judging call. GPT-5 uses `reasoning_effort=medium` for both generation and judging. DeepSeek-V4-Flash uses thinking enabled for generation and disabled for judging. Answer generation is limited to 4,096 output tokens. The answer generation and evaluation prompts are provided in Appendix B.4. We report answer accuracy as the percentage of evaluated queries whose responses are judged correct. Category-wise accuracy is computed separately for each query category, while overall accuracy is computed across all evaluated queries rather than averaged over category-level scores.

B.3 Implementation and Baseline Configurations

CoalMem Implementation. For seed evidence retrieval, each message is represented by its content together with its channel, discussion phase, topic, speaker, and role. The question text is used as the retrieval query. Unless otherwise specified, CoalMem uses BM25 to retrieve $N = 48$ candidate messages and selects $K = 10$ messages as evidence. Query–evidence relevance is computed over the metadata-expanded representations, whereas pairwise coalition utility is computed from message content only. We use a complementarity coefficient of $\lambda = 0.5$ and set the query node's self-loop weight to $\alpha = \lambda_{\max}(\mathbf{W}) + 10^{-3}$. Replicator dynamics starts from uniform node weights and terminates when the ℓ_1 distance between consecutive weight vectors falls below 10^{-8}, or when 1,000 iterations have been reached. The query node is then excluded, and the K highest-weighted messages are supplied to the answer model in descending order of their final node weights. Our code is available at the anonymous repository.

Retrieval and Memory Baselines. BM25 and Embedding retrieve messages without LLM-based memory construction. The Embedding baseline uses Qwen3-Embedding-8B, and both methods return the top-10 messages for each query. LinearRAG (Zhuang et al., 2026) and EHRAG (Song et al., 2026) both use all-MiniLM-L6-v2 for dense representations and `en_core_web_sm` for entity extraction. Each message is treated as a passage, and a separate graph index is constructed for each domain. Both methods return the top-10 messages and use a damping factor of 0.5. LinearRAG uses three expansion iterations, an activation threshold of 0.5, and a passage ratio of 1.5. EHRAG uses five expansion iterations, an activation threshold of 0.1, one initial entity seed, at most two semantic seeds with a decay factor of 0.2, and a concept-clustering threshold of 0.9. For A-Mem (Xu et al., 2026b), HippoRAG (Gutiérrez et al., 2024), and Hindsight (Latimer et al., 2025), memory construction uses GPT-4o-mini, separately from the answer and judge models. A-Mem consults five neighboring memories when establishing links and updating notes and refreshes its retrieval index after every 100 successful memory-evolution updates. Hindsight uses its native recall interface with `budget=mid`. HippoRAG performs OpenIE with temperature 0 and a 4,096-token extraction output limit, using the top-5 candidates for entity linking and returning the top-10 passages.

Alternative Selection Methods. All selection baselines operate on the same $N = 48$ candidates retrieved by metadata-expanded BM25, with at most $K = 10$ messages selected. Let $\widehat{r}_i$

denote max-normalized BM25 relevance and S the content-only TF–IDF cosine-similarity matrix. MMR (Carbonell & Goldstein, 1998; Khan et al., 2026) greedily balances query relevance against the maximum similarity to previously selected messages, with $\lambda = 0.5$. DPP (Kulesza & Taskar, 2012; Sun et al., 2026) uses the quality–similarity kernel $L = \mathrm{diag}(\mathbf{q})S\,\mathrm{diag}(\mathbf{q}) + 10^{-8}I$, where $q_i = \max(\widehat{r}_i, 10^{-6})$. We select ten messages using greedy MAP inference. For SEP (Bala, 2026), we set the weights for relevance, query-term coverage, saturated content coverage, and source diversity to $(1.0, 0.5, 0.4, 0.3)$, respectively, with a content-coverage saturation coefficient of 0.3. Source groups are defined by `_channel`, `phase_name`, and `topic`. SEP greedily selects messages by marginal gain per token, subject to the message-count limit and a token budget equal to the combined token length of the top-10 candidates.

B.4 Prompt Templates

All methods use the same prompts for answer generation and evaluation. The answer model generates an answer from the question and the evidence returned by the corresponding method, while the judge model assesses its correctness against the gold answer. Variables enclosed in braces are instantiated at runtime.

Figure 4 presents the answer generation prompts and an illustrative evidence format. Each retrieved message is assigned a one-based evidence index and includes its content together with available metadata: speaker, role, channel, conversation phase, topic, timestamp, and reply-to relation.

Answer Generation Prompt

System prompt

```
You are a careful QA agent. Use only the retrieved conversation
passages to answer the question.
First provide a chain-of-thought reasoning paragraph. Then
provide the final answer on a new line using the format:
Final: <answer>
If the answer is not supported by the retrieved passages, say so
in the Final line.
```

User prompt template

```
Asking user: {asking_user_id}
Question:
{question}
Retrieved evidence:
{retrieved_evidence}
Answer the question using the retrieved passages.
```

Evidence format (illustration)

```
[1] [user={author} / speaker_role={role} / channel={channel} /
phase_name={phase} / topic={topic} / timestamp={timestamp} /
reply_to={parent}]
{message content}
...
[K] {K-th retrieved message}
```

Figure 4: Answer generation prompts and an illustrative evidence format.

Figure 5 presents the evaluation prompts. The judge model receives the question, the gold answer, and the agent's final answer, but not the retrieved evidence. Answer correctness is determined by the judge model's final verdict: `Final: Correct` or `Final: Incorrect`.

Answer Evaluation Prompt

System prompt

```
You are a strict judge evaluating whether an agent's answer
matches the gold answer for a question.
Consider paraphrases correct if they have the same meaning as the
gold answer.
First provide a brief reasoning paragraph.  Then provide the
final judgment on a new line using the format:
Final:  Correct
or
Final:  Incorrect
```

User prompt template

```
Question:
{question}
Gold Answer:
{gold_answer}
Agent Answer:
{agent_answer}
```

Figure 5: Answer evaluation prompts used to judge the agent's final answer against the gold answer.

C Additional Experimental Results

C.1 Evaluation on EverMemBench

We additionally evaluate CoalMem on EverMemBench (Hu et al., 2026) to assess its effectiveness beyond GroupMemBench. EverMemBench contains 51,023 synthetic dialogue turns totaling approximately 4.2 million tokens, with 170 participants across five project domains: technology, operations, marketing, financial services, and governance. The conversations span one year and are organized into interconnected group chats, featuring interleaved topics, evolving decisions, and participants with distinct roles and communication styles. The evaluation set comprises 2,400 questions across three dimensions: fine-grained recall (762), memory awareness (1,097), and profile understanding (541). Fine-grained recall assesses factual retrieval, multi-hop reasoning, and temporal reasoning. Memory awareness evaluates the application of past constraints and updated information to new situations, including proactive recognition of relevant memories. Profile understanding assesses the ability to infer participants' communication styles, expertise, and professional roles from their interactions. The benchmark uses open-ended questions for fine-grained recall and multiple-choice questions for the other two dimensions.

Table 7: Answer accuracy (%) on EverMemBench across three evaluation dimensions. Overall accuracy is computed across all 2,400 evaluated questions.

Method	Fine-Grained Recall	Memory Awareness	Profile Understanding	Overall
BM25	32.81	71.29	49.91	54.25
Embedding	31.76	73.11	45.66	53.79
LinearRAG	15.35	67.18	48.06	46.42
EHRAG	18.50	69.46	48.98	48.67
Hindsight	35.17	80.13	46.03	58.17
HippoRAG	33.20	77.39	49.72	57.13
A-Mem	32.02	74.38	44.36	54.17
CoalMem	50.00	78.85	58.23	65.04

The results on EverMemBench provide further evidence of CoalMem's effectiveness beyond GroupMemBench. Despite differences in conversation structure and evaluation tasks, CoalMem improves answer accuracy through set-level evidence selection. These findings suggest that its joint modeling of query–evidence relevance and pairwise coalition utility is also useful in a different multi-party conversational memory setting.

C.2 Efficiency of Evidence Retrieval and Selection

Table 8: Efficiency comparison on GroupMemBench. Construction LLM calls are reported as totals for the full conversation corpus, while online latency is averaged per query. Call counts exclude embedding model calls, answer generation, and judging.

Method	Overall Accuracy (%) ↑	Construction LLM Calls ↓	Online Latency (ms / Query) ↓
BM25	48.59	0	1562
Embedding	43.49	0	3682
LinearRAG	41.88	0	1717
EHRAG	42.15	0	2368
HippoRAG	52.89	240,000	8514
Hindsight	47.79	$\geq$ 120,000	5443
A-Mem	41.61	239,996	3720
CoalMem	60.81	0	2162

We compare CoALMem with retrieval-based and memory-based baselines in terms of answer accuracy, construction LLM calls, and online latency. We report the total number of LLM calls for index and memory construction, excluding answer generation, judging, and embedding model calls. Online latency measures the time required to retrieve and prepare the final evidence context, excluding offline construction, answer generation, and judging.

As shown in Table 8, CoALMem achieves the highest overall accuracy of 60.81%. Compared with BM25, it improves accuracy by 12.22 percentage points at the cost of an additional 600 ms of online latency per query. It also achieves higher accuracy and lower online latency than the embedding-based and memory-based baselines. Overall, CoALMem offers a favorable balance between answer accuracy and online efficiency without requiring LLM-based memory construction.

D Case Studies

We present qualitative comparisons of CoALMem with MMR, DPP, and SEP under the same candidate pool and evidence budget. By examining the selected messages alongside the generated answers, we illustrate how CoALMem's query-anchored coalition optimization brings together messages that jointly support an answer, whereas alternative selections can contain relevant information without providing the evidence needed to resolve the query.

CoALMem Retains Direct Evidence for a Multi-Hop Question

Instance. Manufacturing / `multi_hop_10`

Question.
What factor determines the sequence during the Boiler Retrofit Implementation phase?

Alternative selections

The alternative evidence sets include sequence-related statements such as the following:

> **`Msg_28873` — Procedural cue**
> "On Boiler Retrofit Implementation, field measure first—proof over paperwork."
> **`Msg_7514` — Sequence-related statement**
> "Then burner lead sets sequence."

The answer model returns "the burner lead" with DPP-selected evidence, "field measurement" with MMR-selected evidence, and "lead time" with SEP-selected evidence. **All three methods omit `Msg_1583`**, which explicitly identifies the shutdown window as the factor driving the retrofit sequence.

Incorrect

CoALMem selection

CoALMem retains **`Msg_1583`**, which explicitly states that the shutdown window for boiler tie-ins drives the retrofit sequence, directly supporting the answer.

> **`Msg_1583` — Answer-critical evidence**
> "One item in the way is confirming the **shutdown window for boiler tie-ins**, since that **drives the retrofit sequence**. . . ."

Gold answer

The shutdown window.

CoALMem answer

"The **shutdown window for boiler tie-ins**, including the approved access dates and timing."

Correct

Figure 6: Evidence selection for a multi-hop question under the same candidate pool and evidence budget. CoALMem retains the direct evidence omitted by DPP, MMR, and SEP. Blue text highlights the key evidence and the corresponding content in CoALMem's answer.

CoALMem Retains Evidence Confirming the Revised Policy

Instance. Technology / `knowledge_update_24`

Question.
What is the current approach for handling partial aggregation in the scheduling requirements?

Alternative selections

DPP, MMR and SEP all select `Msg_4182`, which states that partial aggregation remains confirm-only unless Platform certifies freshness and completeness. DPP and MMR also select the preview-based proposal in `Msg_4210`, but **all three omit the subsequent confirmation, `Msg_4797`**. The answer model identifies the earlier confirm-only policy as current under all three selections. When given the evidence selected by DPP or MMR, it describes the revision as a proposal whose adoption has not been confirmed.

Incorrect

CoALMem selection

CoALMem retains the earlier policy, the proposed revision, and **`Msg_4797`**. The three messages establish the change from an earlier rule to an adopted revision.

> **`Msg_4182` — Earlier policy**
> "Partial aggregation stays confirm-only unless Platform certifies freshness and completeness."
> **`Msg_4210` — Proposed revision**
> "New proposal: keep partial aggregation visible as a non-bookable preview with a distinct incomplete-data label, then require confirm-time revalidation before any booking action."
> **`Msg_4797` — Confirmation of the revised policy**
> "**Keep partial aggregation as non-bookable preview.** … `partial_aggregation` = **visible, labeled incomplete-data, not selectable**; `stale_source` = selectable, loud warning, confirm-time revalidation."

Gold answer

Update the specification so that partial aggregation is shown as a non-bookable preview with a distinct incomplete-data label, while preserving confirm-time revalidation before any booking action.

CoALMem answer

"The current approach is to keep partial aggregation as a **non-bookable preview with an incomplete-data label—visible but not selectable**—rather than treating it as actionable availability; stale sources, by contrast, remain selectable with a loud warning and confirm-time revalidation."

Correct

Figure 7: Evidence selection for a knowledge-update question under the same candidate pool and evidence budget. CoALMem retains the confirmation omitted by DPP, MMR and SEP. Blue text highlights the key evidence and the corresponding content in CoALMem's answer.

CoALMem Retains Complementary Evidence to Resolve the Approval Deadline

Question.
What is the deadline for finance workflows to approve exception changes during the Implement Report Generation Engine phase?

Alternative selections

DPP, MMR and SEP all omit `Msg_28303`, which supplies the calendar-date anchor for the Friday approval deadline. Given their respective selected evidence sets, the answer model reports that the deadline is not specified.

Incorrect

CoALMem selection

CoALMem retains the approval-workflow evidence, the Friday deadline, and `Msg_28303`. Together, these messages connect the exception-approval process to a specific calendar date.

Selected evidence (summarized)

`Msg_5942` — Exception approval
Exception changes enter the exception lane and require approval from Finance Ops and Controls.

`Msg_5087` — Approval responsibilities
Finance Ops owns the cutoff, while Controls co-approves exception releases after the freeze.

`Msg_6168` — Relative deadline
The approval window closes on **Friday**; the message also discusses the consequences of Finance missing the window.

`Msg_28303` — Calendar-date anchor
The corresponding Friday deadline is identified as **July 18**.

Gold answer
2025-07-18.

CoALMem answer
"**July 18** (the referenced Friday deadline)."

Correct

Figure 8: Evidence selection for a temporal question under the same candidate pool and evidence budget. CoALMem retains the date-anchoring evidence omitted by DPP, MMR and SEP. Blue text highlights the key evidence and the corresponding content in CoALMem's answer.