SAME MEMORY, DIFFERENT INTERPRETATIONS: MULTI-FORM MEMORY RETRIEVAL WITH COARSE-TO-FINE RERANKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Personalized memory agents retrieve key information from long conversational histories to generate query-aligned responses. However, lengthy dialog archives contain extensive redundancy and exceed the limited context budgets of large language models (LLMs). Existing methods organize dialogue histories into static memory representations, lacking the capacity to unlock diverse semantic interpretations embedded within individual memory records. Inspired by the core insight that identical memories support multi-view understanding, this paper proposes **SMDI** (**S**ame **M**emory, **D**ifferent **I**nterpretations), a multi-form memory retrieval framework with coarse-to-fine reranking. We derive three complementary memory units for factual facets, contextual expansions, and thematic summaries to enrich memory semantic modeling. All units are retrieved independently and projected back to source memories to construct unified candidate pools. We further propose a coarse-to-fine reranking strategy to produce global initial rankings and adaptively refine high-quality top candidates with multi-view ranking signals. Experimental results on three long-term memory benchmarks demonstrate that SMDI consistently outperforms state-of-the-art methods.

1 INTRODUCTION

Large language models (LLMs) have enabled conversational agents to provide personalized responses by leveraging information from previous interactions (Li et al., 2016; Zhang et al., 2018; Mazaré et al., 2018). However, as conversations accumulate over time, dialogue histories grow increasingly long and contain large amounts of redundant information, while relevant user facts, preferences, and events may be scattered across distant sessions. Constrained by limited input context, LLMs may struggle to capture key information distributed throughout such long interaction histories. Besides, different queries require different aspects of historical interactions (e.g., specific information in a single conversation, evidence distributed across multiple sessions or information that evolves over time). These characteristics make it challenging to identify the historical evidence most relevant to each query.

Existing studies on long-term conversational memory follow Retrieval-Augmented Generation (RAG) pipelines, where historical interactions are organized into external memory representations and retrieved to support personalized response generation (Xu et al., 2022a; Packer et al., 2023; Wu et al., 2025). For memory organization, existing methods aim to organize long dialogue histories into compact and structured memory representations. SeCom (Pan et al., 2025) segments long conversations into topically coherent units and further compresses each segment into a compact representation for subsequent retrieval. A-MEM (Xu et al., 2025) constructs structured memory notes that combine the original content with contextual descriptions, keywords, tags, and explicit links to related memories. MemGAS (Xu et al., 2026) organizes historical interactions at multiple granularities and associates semantically related memories through clustering, forming memory units at different abstraction levels. For memory reranking, recent approaches introduce additional relevance signals or ranking models to improve the ordering of retrieved memory candidates (Sun et al., 2023; Gangi Reddy et al., 2024). MemReranker (Li et al., 2026) trains a reranker model over memory queries involving temporal, causal, and contextual information to score retrieved candidates.

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.

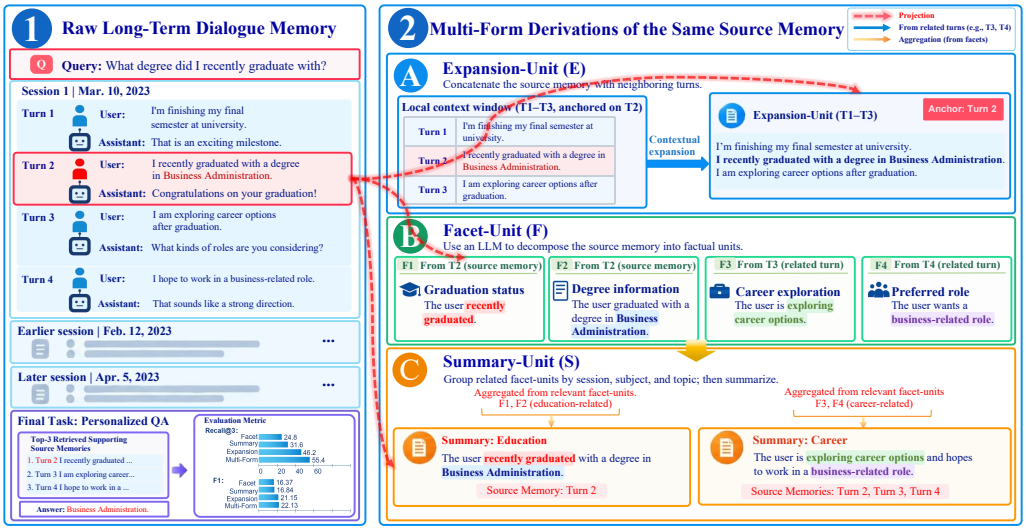


Figure 1: Illustration of multi-form memory organization from the same source memory.

EARM (Feng et al., 2026) further incorporates relevance scores obtained from previous queries as additional signals when reranking memory candidates for later queries.

However, two key limitations remain in these methods: (1) **Single-form source memory interpretations.** Existing methods typically represent and retrieve historical information through a particular memory form, such as compressed segments, structured notes, or hierarchical graphs. Such single-form interpretations may overlook complementary information contained in the same source memory, making it difficult to identify relevant memory for queries that require different aspects of the original interactions; (2) **Insufficient integration of ranking signals.** current reranking methods often rely on a fixed strategy to rerank retrieved memory candidates, which may limit their ability to distinguish among highly relevant candidates, as different ranking signals can provide varying levels of useful information across queries.

To address these limitations, we propose **SMDI** (**S**ame **M**emory, **D**ifferent **I**nterpretations), a multi-form memory retrieval framework with coarse-to-fine reranking. Specifically, SMDI first derives three complementary memory forms from the same source memory: Facet-Units for factual interpretation, Summary-Units for thematic interpretation, and Expansion-Units for contextual interpretation. Each derived unit preserves a provenance link to its supporting source memory, allowing retrieved results from different forms to be projected back and fused into a unified source memory candidate set. SMDI then adopts a coarse-to-fine reranking strategy: the coarse stage integrates complementary signals to obtain an initial ranking, while the fine stage adaptively combines multiple ranking signals to further refine the ordering of these candidates for each query. Experiments on the three long-memory datasets demonstrate the effectiveness of our framework that SMDI achieves state-of-the-art performance on both retrieval and generation tasks. Ablation studies further validate the contributions of each module in our framework.

2 Methodology

2.1 Problem Formulation

This work focuses on long-term conversational memory retrieval, which aims to answer user queries by leveraging valuable evidence from extensive historical interaction records. The entire dialogue history is hierarchically organized as $\mathcal{H} = \{S_1, \ldots, S_N\}$, where each *session* $S_i = \{t_{i,1}, \ldots, t_{i,L_i}\}$ denotes an ordered sequence of dialogue *turns*. Given a user query q, our goal is to retrieve and rerank a compact subset of relevant historical turns to form the final memory set $\mathcal{M}_q = \mathcal{R}(q, \mathcal{H})$. The LLM generates grounded responses by taking the query and retrieved memories as contextual inputs, i.e.,

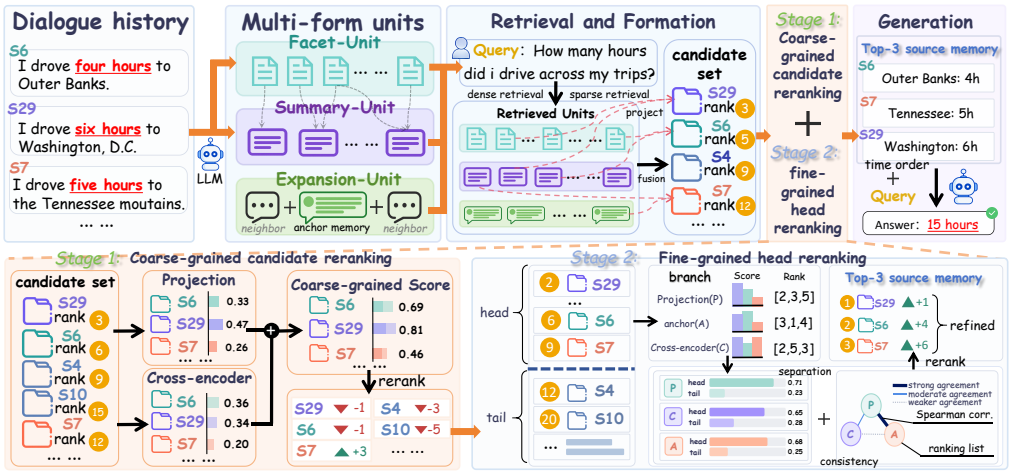


Figure 2: Overview of our proposed framework SMDI.

$y = \mathrm{LLM}(q, \mathcal{M}_q)$. The core challenge lies in capturing diverse semantic perspectives of historical memories adaptively and distilling valid evidence from redundant or irrelevant dialogue contents.

2.2 Multi-Form Derivations of a Native Memory

Existing memory construction methods typically compress historical dialogues into a single fixed representation. However, a native memory (i.e., a single dialogue turn) contains multi-dimensional semantic information that can be interpreted from distinct perspectives. To fully exploit diverse contextual and factual clues, we construct three complementary memory units to enrich the semantics of each source memory, covering factual, thematic, and contextual perspectives.

Intra-Memory Facet Decomposition (Facet-Unit). We adopt an LLM to decompose each source memory m into multiple Facet-Units u_f, where each unit corresponds to one independent, fine-grained factual statement. The decomposed facets cover four typical semantic types: temporal descriptions, participant information, behavioral actions, and entity attributes. Each Facet-Unit preserves categorical and temporal metadata, providing a precise factual interpretation of the original memory.

Cross-Memory Topic Summary (Summary-Unit). We aggregate Facet-Units with consistent topics within the same session and employ an LLM to generate a topic-level Summary-Unit u_s. The summarized topics mainly include long-term user preferences, ongoing personal plans, and multi-turn event narratives. This unit captures shared themes, key events, and chronological characteristics across scattered source memories, yielding a global thematic interpretation.

Neighboring Memory Contextual Expansion (Expansion-Unit). For each source memory m, we construct an Expansion-Unit u_e by concatenating the source memory as an anchor with its neighboring memories within a local session window, while completely retaining the raw dialogue content. Unlike abstract summary units or isolated factual facets, Expansion-Units focus on contextual supplementation, capturing implicit logical connections, turn-level dialogue coherence, and situational background knowledge that are otherwise lost during decomposition and abstraction. This unit provides a context-aware interpretation and complements the semantic incompleteness of decomposed memory units.

All derived units maintain explicit provenance links to their original source memories. Such traceable metadata enables all three types of memory units to be accurately projected back to corresponding dialogue turns, supporting subsequent multi-form retrieval and source-level candidate fusion.

2.3 Multi-Form Retrieval and Candidate Formation

Memory units derived from the same source memory provide complementary retrieval evidence from distinct semantic perspectives. Since multiple heterogeneous units may originate from identical source memories, their individual relevance signals should be aggregated to evaluate the comprehensive

relevance of each source memory. Accordingly, we project unit-level retrieval results back to the source-memory level and fuse multi-perspective signals to construct high-quality candidate sets.

Multi-Form Retrieval. We define $\mathcal{F} = \{\text{Facet}, \text{Summary}, \text{Expansion}\}$ as the set of three memory forms. For each memory form $f \in \mathcal{F}$, we perform an independent hybrid retrieval pipeline for the input query q. Specifically, sparse and dense retrievers separately output top-K_1 results, and a deduplicated union is adopted to form the preliminary unit-level candidate list $L_f^{(1)}$. A pretrained cross-encoder reranker further scores each candidate unit, and we retain the top-K_2 high-relevance units to obtain the refined list $L_f^{(2)}$ for each memory form.

Source-Memory Projection and Fusion. Retrieved units from different forms may correspond to identical source memories. Let $\mathcal{M}$ denote the entire source memory set, and $\mathcal{U}(m)$ represent all derived units originating from source memory m. Guided by provenance links, each retrieved unit in $L_f^{(2)}$ is projected to its parent source memory. We aggregate reciprocal ranks of all associated units to compute the unified projection score for each source memory m:

$$S(m) = \sum_{f \in \mathcal{F}} \sum_{u \in L_f^{(2)} \cap \mathcal{U}(m)} \frac{1}{r_f(u)}, \tag{1}$$

where $r_f(u)$ denotes the ranking position of unit u under memory form f. Finally, we rank all source memories by $S(m)$ and select the top-K_3 items as the source candidate set L_{src}.

2.4 Coarse-to-Fine Memory Candidate Reranking

Multi-form memory signals provide complementary relevance cues, while their discriminative capacities differ across scenarios. Direct fusion of heterogeneous rankings fails to resolve fine-grained ambiguity among highly similar candidates. To address this issue, we design a coarse-to-fine reranking paradigm, which first produces globally robust rankings via multi-view fusion and then adaptively refines ambiguous head candidates with dynamic branch weights.

Coarse-Grained Candidate Reranking. Based on the source memory candidates L_{src} obtained in the previous stage, we conduct coarse-grained reranking to eliminate obviously irrelevant items. For each candidate memory m, we derive two independent ranking signals: the projection fusion rank $r_p(m)$ computed from $S(m)$, and the cross-encoder rank $r_c(m)$ obtained by directly scoring the raw source memory against the query. We fuse the two rankings through a weighted reciprocal strategy to obtain the coarse-grained score $S^{(1)}(m)$:

$$S^{(1)}(m) = \lambda \frac{1}{r_p(m) + \alpha} + (1 - \lambda) \frac{1}{r_c(m) + \alpha}, \tag{2}$$

where λ balances the two ranking modalities, and α serves as a constant offset to avoid numerical division-by-zero issues and stabilize ranking distributions. After coarse reranking, we preserve the full candidate set and select the top-H items as the high-confidence head candidate set $\mathcal{K}_H$ for subsequent fine-grained refinement.

Fine-Grained Head Reranking. To resolve ambiguous rankings among semantically similar head candidates, we establish three complementary ranking branches $\mathcal{B} = \{p, c, a\}$, corresponding to projection fusion ranking, cross-encoder ranking, and native contextual anchor ranking, respectively. For each branch $i \in \mathcal{B}$, we compute two core metrics: score separation degree Δ_i and pairwise ranking consistency $C_{(i,j)}$:

$$\Delta_i = \frac{\bar{S}_i^{\text{head}} - \bar{S}_i^{\text{tail}}}{\max(S_i) - \min(S_i)}, \qquad C_{(i,j)} = \frac{1 + \rho_S(r_i, r_j)}{2}, \tag{3}$$

where $\bar{S}_i^{\text{head}}$ and $\bar{S}_i^{\text{tail}}$ denote the average scores of head and tail candidates under branch i, and ρ_S denotes the Spearman rank correlation coefficient between two ranking lists.

We first compute the average pairwise consistency to obtain the overall stability $\bar{C}_i$ for each branch. The final branch confidence S_i comprehensively considers both discriminative separation ability and ranking stability:

$$\bar{C}_i = \frac{1}{|\mathcal{B}| - 1} \sum_{\substack{j \in \mathcal{B} \\ j \neq i}} C_{i,j}, \qquad S_i = \frac{1}{2} \max(\Delta_i, 0) + \frac{1}{2} \bar{C}_i, \tag{4}$$

The adaptive fusion weight of each branch is normalized from its confidence score:

$$w_i = \frac{S_i}{\sum_{j \in \mathcal{B}} S_j}. \tag{5}$$

With dynamically optimized branch weights, we fuse multi-branch rankings to yield the fine-grained relevance score for each head candidate memory:

$$S^{(2)}(m) = \sum_{i \in \mathcal{B}} \frac{w_i}{r_i(m)}. \tag{6}$$

Ultimately, head candidates are reordered by $S^{(2)}(m)$. Following standard dialogue generation settings with retrieved memories, we select the top-3 refined source memories and concatenate them with the original query as grounded contextual evidence for LLM response generation.

3 Experiments

3.1 Experimental Settings

Datasets and Metrics. We evaluate our method on three long-term conversational memory datasets: LoCoMo (Maharana et al., 2024), LongMemEval-S, and LongMemEval-M (Wu et al., 2025). For retrieval evaluation, we report Recall@K and NDCG@K at $K \in \{3, 5, 10\}$ to measure evidence coverage and ranking quality, respectively. For generation evaluation, we report the GPT-4o-based judge score (4o-J), token-level F1, BLEU-4, ROUGE-1/2/L, and BERTScore. More details of the datasets and metrics are provided in Appendix C.

Baselines. We compare our method with representative long-term memory and retrieval approaches covering different memory construction paradigms. The baselines include MPNet (Song et al., 2020), Contriever (Izacard et al., 2022), MPC (Lee et al., 2023), RecurSum (Wang et al., 2025), SeCom (Pan et al., 2025), RAPTOR (Sarthi et al., 2024), HippoRAG2 (Gutiérrez et al., 2025), and A-MEM (Xu et al., 2025). We additionally include the full conversation history as a reference setting for generation. The details of these baselines can be found in Appendix D.1.

Table 1: Retrieval performance on LongMemEval-s, LongMemEval-m, and LoCoMo. The best results are in bold, and the sub-optimal results are underlined.

Model	Recall@3	NDCG@3	Recall@5	NDCG@5	Recall@10	NDCG@10
LongMemEval-s						
RecurSum (Wang et al., 2025)	67.23	78.33	79.79	81.76	87.66	83.28
MPC (Lee et al., 2023)	60.00	70.90	68.09	73.27	80.00	76.59
SeCom (Pan et al., 2025)	71.06	80.88	80.43	83.08	89.15	85.11
HippoRAG2 (Gutiérrez et al., 2025)	75.53	85.44	84.68	87.32	91.28	88.73
MemGAS (Xu et al., 2026)	<u>78.51</u>	<u>86.83</u>	<u>88.94</u>	<u>88.77</u>	<u>94.47</u>	<u>89.96</u>
SMDI (Ours)	**82.20**	**90.92**	**90.20**	**92.10**	**97.00**	**93.38**
LongMemEval-m						
RecurSum (Wang et al., 2025)	23.19	32.10	31.70	36.99	43.62	41.40
MPC (Lee et al., 2023)	35.96	47.51	42.55	50.36	54.26	53.88
SeCom (Pan et al., 2025)	44.26	56.61	55.32	60.76	66.60	63.78
MemGAS (Xu et al., 2026)	<u>51.06</u>	<u>61.36</u>	<u>63.62</u>	<u>66.07</u>	<u>77.02</u>	<u>69.46</u>
SMDI (Ours)	**64.20**	**73.74**	**76.00**	**78.05**	**81.20**	**79.64**
LoCoMo						
RecurSum (Wang et al., 2025)	47.23	48.99	59.01	54.58	74.97	60.07
MPC (Lee et al., 2023)	49.50	51.47	57.45	55.53	71.85	60.47
SeCom (Pan et al., 2025)	55.24	57.90	64.80	62.36	78.30	66.97
HippoRAG2 (Gutiérrez et al., 2025)	56.60	58.37	65.06	62.50	78.05	66.79
MemGAS (Xu et al., 2026)	<u>57.30</u>	<u>58.76</u>	<u>67.32</u>	<u>63.62</u>	<u>81.82</u>	<u>68.42</u>
SMDI (Ours)	**66.40**	**64.99**	**73.51**	**68.44**	**82.09**	**70.87**

Implementation Details. We use DeepSeek-V4-Flash[1] to construct the multi-form memory units and apply the same evidence filtering procedure to all methods before answer generation. For fair

[1] https://api-docs.deepseek.com/

comparison, we follow the generation and evaluation protocol of MemGAS (Xu et al., 2026), using the same answer generator, prompt template, and evaluation settings across all methods. Additional implementation details, including memory construction, evidence filtering, prompts, and generation configurations, are provided in Appendix 3.

3.2 Overall Results

The results of retrieval and generation are illustrated in table 1 and 2, respectively. The analysis of these results is as follows.

Retrieval Results Table 1 compares the retrieval performance between our SMDI framework and the competitors on the LongMemEval-S, LongMemEval-M, and LoCoMo datasets. It is observed that our SMDI achieves competitive performance across multiple metrics on all datasets, demonstrating notable improvements over baseline models in both memory recall and ranking. Our SMDI framework improves Recall@3 and NDCG@3 by 3.66 and 4.09 points over the second-best results on LongMemEval-S, with Recall@3 and NDCG@3 gains of 13.14 and 12.38 points on LongMemEval-M, and 9.10 and 6.23 points on LoCoMo, respectively. These improvements validate our design motivations and objectives. We argue that multi-form derivation captures complementary interpretations of a native memory from different perspectives, source memory projection integrates complementary information while preserving coherent source context, and coarse-to-fine reranking further improves the query-specific ordering of candidate memories. These results indicate that the effectiveness of SMDI comes not only from organizing memories into multiple complementary forms, but also from source memory projection and coarse-to-fine reranking, which better align with the task of long-term conversational retrieval.

Table 2: Generation performance on LongMemEval-s, LongMemEval-m, and LoCoMo. The best results are in bold, and the sub-optimal results are underlined.

Model	4o-J	F1	B-4	R-1	R-2	R-L	BS	Avg. Tokens	Avg. Latency
LongMemEval-s									
Full History	50.60	11.48	1.40	12.10	5.47	10.85	83.07	103,137	9.39
MPC (Lee et al., 2023)	53.80	13.60	1.74	14.27	6.49	12.95	83.49	8,457	2.66
RecurSum (Wang et al., 2025)	35.40	12.29	2.09	13.01	5.55	11.52	83.60	8,853	1.94
SeCom (Pan et al., 2025)	56.00	12.95	2.25	13.80	6.09	11.93	83.51	2,741	1.67
HippoRAG2 (Gutiérrez et al., 2025)	57.60	14.73	2.15	15.30	7.36	13.83	83.86	8,530	4.51
RAPTOR (Sarthi et al., 2024)	32.20	12.08	1.90	12.73	5.82	11.25	83.50	6,254	2.25
A-Mem (Xu et al., 2025)	55.60	13.73	2.11	14.82	6.81	12.98	83.88	9,018	2.59
MemGAS (Xu et al., 2026)	60.20	20.38	4.22	21.05	10.47	19.47	85.21	8,829	2.55
SMDI (Ours)	**64.80**	**22.13**	**5.24**	**22.90**	**11.84**	**21.51**	**86.67**	8,802	1.73
LongMemEval-m									
Full History	12.20	5.70	0.78	6.27	2.08	5.28	81.62	128,000	12.88
MPC (Lee et al., 2023)	37.80	11.28	1.37	11.93	5.12	10.57	82.98	8,428	2.75
RecurSum (Wang et al., 2025)	23.80	10.04	1.70	10.89	4.26	9.21	83.12	8,927	1.99
SeCom (Pan et al., 2025)	42.80	11.33	1.79	12.03	5.07	10.49	83.36	2,821	1.63
MemGAS (Xu et al., 2026)	45.40	16.85	3.39	17.60	8.25	16.14	84.69	8,852	2.45
SMDI (Ours)	**57.80**	**19.83**	**4.20**	**20.56**	**9.87**	**18.96**	**86.33**	8,810	1.96
LoCoMo									
Full History	33.43	12.23	1.84	12.70	5.66	11.73	84.07	20,078	4.92
MPC (Lee et al., 2023)	40.38	14.81	1.99	15.10	6.83	14.13	84.43	2,683	1.95
RecurSum (Wang et al., 2025)	22.56	9.14	0.99	9.82	3.38	8.98	83.45	3,074	1.58
SeCom (Pan et al., 2025)	44.21	13.79	2.30	14.28	6.17	13.30	84.04	1,021	1.02
HippoRAG2 (Gutiérrez et al., 2025)	45.62	16.66	2.91	17.01	8.27	15.93	84.88	2,991	3.56
RAPTOR (Sarthi et al., 2024)	31.72	14.55	2.88	15.09	7.49	14.18	84.48	1,931	1.73
A-Mem (Xu et al., 2025)	40.81	14.72	2.83	16.22	7.71	14.89	84.72	3,042	1.98
MemGAS (Xu et al., 2026)	41.07	17.66	3.61	18.00	8.93	16.99	85.13	2,825	1.88
SMDI (Ours)	**46.50**	**23.58**	**5.69**	**25.53**	**12.99**	**24.05**	**88.04**	1,437	1.76

Generation Results The generation results of our SMDI framework and the baselines on three datasets are reported in Table 2. Our SMDI achieves consistently strong generation performance across all three datasets. Specifically, SMDI improves the F1 scores by 1.75, 2.98, and 5.92 points

over the second-best baselines on LongMemEval-S, LongMemEval-M, and LoCoMo, respectively, with consistent improvements across the other generation metrics. We attribute these improvements to the multi-form interpretations, source memory projection, and coarse-to-fine reranking strategy, which respectively enrich memory representations, preserve coherent source context, and refine the ranking of candidate memories, thereby providing more relevant evidence for response generation.

Besides, we note that SMDI maintains balanced token consumption and generation latency, which demonstrates its computational efficiency. It is interesting to observe that using the full conversation history leads to larger input contexts and higher inference latency. While other baselines reduce these costs by compressing or selecting historical information, they may overlook useful details or complementary evidence, which makes it difficult to provide sufficient evidence for response generation. In contrast, our SMDI framework combines multi-form retrieval with source memory projection and reranking to preserve useful evidence in a compact context, which supports stronger generation without substantially increasing inference costs.

Table 3: Ablation of memory forms on LongMemEval-S. u_f, u_s, and u_e denote facet-unit, summary-unit, and expansion-unit, respectively.

	Recall-all			NDCG			Generation						
Memory forms	@3	@5	@10	@3	@5	@10	4o-J	F1	BLEU-4	R-1	R-2	R-L	BS
u_f only	50.40	58.40	63.60	67.03	68.62	69.97	46.50	16.37	3.42	17.17	7.87	15.63	85.70
u_s only	59.40	65.80	67.40	74.49	75.15	75.47	48.60	16.84	3.49	17.68	8.18	15.86	85.84
u_e only	78.80	86.40	93.20	86.48	87.80	89.30	56.40	21.15	4.30	21.67	10.70	20.01	86.47
w/o u_f	80.00	88.20	96.20	88.19	89.68	91.25	61.80	20.06	4.01	20.80	10.05	19.12	86.36
w/o u_s	81.40	89.80	96.00	88.90	90.30	91.57	59.20	21.00	4.29	21.61	10.52	19.91	86.52
w/o u_e	60.00	67.20	69.00	74.57	75.71	76.03	51.40	17.64	3.71	18.40	8.69	16.67	85.97
Full model	**82.20**	**90.20**	**97.00**	**90.92**	**92.10**	**93.38**	**64.80**	**22.13**	**5.24**	**22.90**	**11.84**	**21.51**	**86.67**

3.3 Ablation Study

Ablation of Memory Forms. To evaluate the contribution of different memory forms, we consider three settings: (1) each form individually, (2) different combinations of two forms, and (3) all three forms together. The results on LongMemEval-S are presented in Table 3. Among the single-form settings, the expansion-unit achieves the strongest performance, indicating the importance of contextual information. While the combinations of two forms, removing the expansion unit causes the largest performance drop, while excluding the facet or summary unit also leads to consistent degradation, showing that the three forms provide complementary information. Combining all three forms achieves the best overall retrieval and generation performance, validating the benefit of jointly modeling factual, thematic, and contextual interpretations of the same memory.

Table 4: Main ablation study on LongMemEval-S. The cumulative rows evaluate the contribution of each reranking stage. The branch rows isolate the ranking facets used by the coarse and fine stages.

	Recall-all			NDCG			Generation						
Configuration	@3	@5	@10	@3	@5	@10	4o-J	F1	BLEU-4	R-1	R-2	R-L	BS
Coarse-stage branch ablations													
w/o coarse BGE branch	82.20	89.40	95.20	90.57	91.63	92.78	59.60	18.17	4.05	18.75	9.31	17.56	84.86
w/o coarse projection branch	76.00	82.80	87.40	87.23	88.14	88.91	54.40	17.33	3.89	17.77	8.96	16.77	84.70
Fine-stage branch ablations													
w/o fine projection branch	80.40	88.80	97.00	88.11	89.62	91.22	57.20	18.21	4.03	18.65	9.43	17.64	84.82
w/o fine BGE branch	77.20	87.00	97.00	86.09	88.24	90.31	58.60	17.56	4.05	18.06	9.28	16.91	84.71
w/o fine anchor branch	81.80	89.20	97.00	88.76	90.21	91.47	57.20	18.06	4.02	18.68	9.74	17.63	84.85
Cumulative stage ablations													
Initial ranking (projection)	80.00	89.20	95.20	88.83	90.49	91.73	57.60	17.43	4.08	17.91	9.05	16.97	84.75
+ Coarse reranking	76.80	88.60	97.00	85.94	88.55	90.33	55.00	17.07	4.05	17.57	8.91	16.67	84.67
+ Fine reranking (full model)	**82.20**	**90.20**	**97.00**	**90.92**	**92.10**	**93.38**	**64.80**	**22.13**	**5.24**	**22.90**	**11.84**	**21.51**	**86.67**

Ablation Study of Retrieval and Reranking Stages. To examine the contribution of different reranking stages, we conduct both branch-level ablations and ranking stage ablations on LongMemEval-S. As shown in Table 4, removing the projection branch from the coarse-grained candidate reranking

stage causes a clear performance drop, demonstrating the importance of source-memory projection in the coarse reranking stage. In the fine-grained head reranking stage, removing the BGE branch leads to the largest drop in top-ranked retrieval performance, while removing the projection or anchor branch also degrades the results, showing that the three ranking signals provide complementary information. The ranking stage ablations further show that adding coarse reranking increases Recall@10 from 95.20 to 97.00 but decreases Recall@3 and NDCG@3, suggesting that this stage mainly improves the coverage of relevant candidates rather than refining their ranking positions. After further introducing fine head reranking, Recall@3 and NDCG@3 increase to 82.20 and 90.92, and the GPT-4o judge score improves to 64.80, demonstrating that the fine stage effectively refines the ordering of high-quality candidates and improves the final retrieval and generation performance.

3.4 Further Comparison Analysis

Retrieval budget analysis. Figure 3 illustrates the effect of the retrieval budget on performance and computational cost. Increasing the budget initially improves both Recall and NDCG, suggesting that a broader candidate set helps recover more useful evidence. However, the improvement becomes limited once the budget reaches a moderate scale, while the computational cost continues to grow. We also observe that the metrics of recall@10 benefit slightly from a larger reranking budget, whereas NDCG shows little further improvement, indicating that retrieving more relevant source turns does not necessarily improve their ranking quality. Therefore, a moderate retrieval budget provides a better balance between effectiveness and efficiency, and we use $K_1/K_2/K_3 = 20/40/20$ as the default setting.

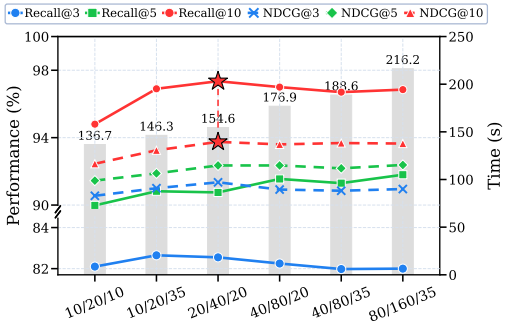


Figure 3: Performance–efficiency trade-off under different retrieval budgets on LongMemEval-S.

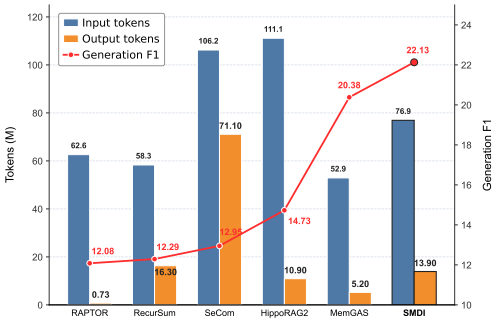


Figure 4: Comparison of token consumption and generation F1 on LongMemEval-S.

Generation budget analysis. As shown in Figure 4, we report the total LLM token consumption, including memory organization and response generation. Notably, although **SMDI** requires higher computational costs than previous methods, it achieves the highest generation F1 score. We argue that this additional cost indicates a trade-off for constructing more comprehensive memory representations. Specifically, the increased token consumption mainly stems from organizing multiple complementary representations derived from the same source memory. Such multi-form derivations preserve different interpretations of historical information, thereby providing precise evidence for response generation.

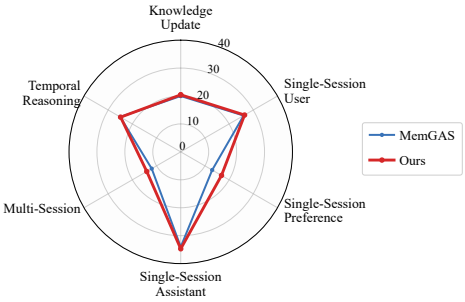


Figure 5: Comparison of F1 across different query types in LongMemEval-S.

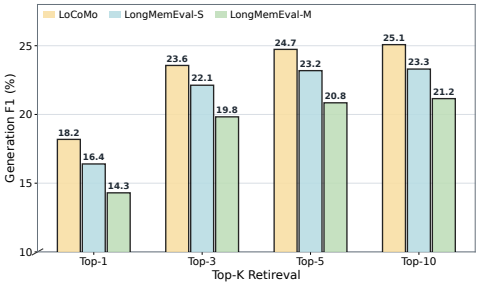


Figure 6: Comparison on different Top-K retrieval.

Comparison of different query types. Figure 5 compares the performance of different methods across query types. SMDI achieves stronger results on single-session user, multi-session, and

temporal-reasoning queries. Besides, the improvement on knowledge-update queries further suggests that our multi-view retrieval and reranking framework remains effective when relevant information changes over time. It is interesting to note that SMDI shows a slight performance drop on single-session assistant queries. We argue that such queries typically rely on explicit information from a single session, where the answer can often be obtained from the conversation context. Consequently, the benefit of integrating multiple memory forms is less significant compared with queries requiring cross-session evidence aggregation or temporal reasoning. Overall, these results indicate that SMDI is effective for queries that require integrating relevant information across multiple sessions or reasoning over changes in historical information over time.

Comparison on different Top-K retrieval. Figure 6 presents SMDI under varying Top-K retrieval sizes. SMDI consistently achieves better performance on both LoCoMo and LongMemEval-S across different values of K. The largest improvement occurs from Top-1 to Top-3, suggesting that a single retrieved memory often provides incomplete evidence, whereas a small set of highly ranked memories can better cover the complementary information required for answering the query. It is interesting to note that the performance changes only slightly from Top-5 to Top-10. We argue that increasing K brings limited additional evidence while introducing more redundant memories into the generation context, which suggests that retrieving a compact set of high-quality memories is more effective than simply increasing the amount of retrieved context.

4 Related Work

Long-Term Memory Organization. Existing methods use different memory units and structure to organize long-term conversational memory (Zhong et al., 2024; Packer et al., 2023; Liu et al., 2026; Shen et al., 2026; Chen et al., 2026). Mem0 (Chhikara et al., 2025) extracts important facts as memories and can further organize entities mentioned in these facts and their relations in a graph. MemoryOS (Kang et al., 2025) organizes memory into short, intermediate, and long-term storage, and progressively consolidates earlier interactions into longer term memory. MemoryBank (Zhong et al., 2024) maintains long-term user memories through incremental memory updates and retrieval over historical interactions. However, these methods often organize memories following a single construction scheme, making it difficult to capture different interpretations of the same memory. In addition, summarizing or compressing conversations into memory units may overlook details contained in the original context.

Memory Retrieval and Reranking. Memory retrieval aims to select relevant information from stored memories based on the current query (Wu et al., 2025; Cao et al., 2026; Banerjee et al., 2026). LongMemEval (Wu et al., 2025) combines semantic retrieval with fact-augmented indexing and time-aware query expansion. HiGMem (Cao et al., 2026) first identifies relevant events through their summaries and then selects detailed conversation turns within these events. APEX-MEM (Banerjee et al., 2026) uses multiple retrieval tools to search structured memories according to different query requirements. Beyond initial retrieval, reranking methods further refine the candidate ranking using additional relevance signals (Sun et al., 2023; Gangi Reddy et al., 2024). ERMAR (Alselwi et al., 2026) applies a pointwise reranker to score retrieved memory candidates. However, these retrieval methods often search different memory forms separately, with limited use of their complementary information. In addition, existing reranking methods often follow a fixed ranking strategy, which may overlook the varying importance of different ranking signals across queries.

5 Conclusion

In this work, we ask a simple question: *instead of representing each memory in a single form, can multiple complementary interpretations provide better access to the same historical information?* To this end, we propose **SMDI** for long-term conversational retrieval, which captures factual, thematic, and contextual information through different memory units, projects retrieved units back to their source memories, and refines the head candidates through coarse and fine-grained reranking. Extensive experiments on LongMemEval-S, LongMemEval-M, and LoCoMo demonstrate consistent improvements in both memory retrieval and downstream response generation. In future work, we plan to investigate more efficient memory construction and updating mechanisms, and extend SMDI to larger-scale and continuously evolving conversational memories.

AI USE STATEMENT

In this work, we used generative AI tools for assisting with manuscript preparation, including improving language clarity, grammar, and readability, as well as refining the organization and presentation of technical descriptions. We have not used generative AI tools to independently generate scientific claims, experimental conclusions, or unverified research results. The remaining required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for recommended disclosure tasks, including assisting with literature organization and improving manuscript presentation. We have reviewed all AI-assisted work, including manually verifying the correctness of technical content and reference citations. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Ghadir Alselwi, Hao Xue, Shoaib Jameel, Basem Suleiman, Flora D. Salim, and Imran Razzak. Long context modeling with ranked memory-augmented retrieval. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3576–3590, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.162.

Pratyay Banerjee, Masud Moshtaghi, Shivashankar Subramanian, Amita Misra, and Ankit Chadha. APEX-MEM: Agentic semi-structured memory with temporal reasoning for long-term conversational AI. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16470–16489, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.749.

Shuqi Cao, Jingyi He, and Fei Tan. HiGMem: A hierarchical and LLM-guided memory system for long-term conversational agents. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 33853–33862, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.1690.

Xingyu Chen, Rui Wang, Zhaopeng Tu, and Liefeng Bo. Adamem: Learning what to remember with adaptive memory policies for personalized agents. *arXiv preprint arXiv:2606.21144*, 2026.

Yi-Pei Chen, Noriki Nishida, Hideki Nakayama, and Yuji Matsumoto. Recent trends in personalized dialogue generation: A review of datasets, methodologies, and evaluations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pp. 13650–13665, 2024.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory. In Inês Lynce, Nello Murano, Mauro Vallati, Serena Villata, Federico Chesani, Michela Milano, Andrea Omicini, and Mehdi Dastani (eds.), *ECAI 2025 - 28th European Conference on Artificial Intelligence, 25-30 October 2025, Bologna, Italy - Including 14th Conference on Prestigious Applications of Intelligent Systems (PAIS 2025)*, volume 413 of *Frontiers in Artificial Intelligence and Applications*, pp. 2993–3000. IOS Press, 2025. doi: 10.3233/FAIA251160. URL https://doi.org/10.3233/FAIA251160.

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

Qi Feng, Chris Ding, and Jicong Fan. The retriever should remember: Experience-amortized reranking for long-term agent memory. *arXiv preprint arXiv:2608.22767*, 2026.

Revanth Gangi Reddy, JaeHyeok Doo, Yifei Xu, Md Arafat Sultan, Deevya Swain, Avirup Sil, and Heng Ji. FIRST: Faster improved listwise reranking with single token decoding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8642–8652, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.491. URL https://aclanthology.org/2024.emnlp-main.491/.

Jingsheng Gao, Yixin Lian, Ziyi Zhou, Yuzhuo Fu, and Baoyuan Wang. Livechat: A large-scale personalized dialogue dataset automatically constructed from live streaming. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 15387–15405, 2023.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. From RAG to memory: Non-parametric continual learning for large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 21497–21515. PMLR, 2025. URL https://proceedings.mlr.press/v267/gutierrez25a.html.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=jKN1pXi7b0.

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

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. Memory OS of AI agent. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 25961–25970, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1318.

Gibbeum Lee, Volker Hartmann, Jongho Park, Dimitris Papailiopoulos, and Kangwook Lee. Prompted llms as chatbot modules for long open-domain conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 4536–4554. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.277.

Chunyu Li, Mengyuan Zhang, Jingyi Kang, Ding Chen, Jiajun Shen, Bo Tang, Xuanhe Zhou, Feiyu Xiong, and Zhiyu Li. MemReranker: Reasoning-aware reranking for agent memory retrieval. *arXiv preprint arXiv:2605.06132*, 2026.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 994–1003, 2016.

Xintong Li, Jalend Bantupalli, Ria Dharmani, Yuwei Zhang, and Jingbo Shang. Toward multi-session personalized conversation: A large-scale dataset and hierarchical tree framework for implicit reasoning. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 11493–11506. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.EMNLP-MAIN.580. URL https://doi.org/10.18653/v1/2025.emnlp-main.580.

Jiaqi Liu, Yaofeng Su, Peng Xia, Siwei Han, Zeyu Zheng, Cihang Xie, Mingyu Ding, and Huaxiu Yao. Simplemem: Efficient lifelong memory for LLM agents. *CoRR*, abs/2601.02553, 2026. doi: 10.48550/ARXIV.2601.02553. URL https://doi.org/10.48550/arXiv.2601.02553.

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

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2775–2779, 2018.

Charles Packer, Vivian Fang, Shishir G. Patil, Kevin Lin, Sarah Wooders, and Joseph E. Gonzalez. Memgpt: Towards llms as operating systems. *CoRR*, abs/2310.08560, 2023. doi: 10.48550/ARXIV.2310.08560. URL https://doi.org/10.48550/arXiv.2310.08560.

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Xufang Luo, Hao Cheng, Dongsheng Li, Yuqing Yang, Chin-Yew Lin, H. Vicky Zhao, Lili Qiu, and Jianfeng Gao. Secom: On memory construction and retrieval for personalized conversational agents. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=xKDZAW0He3.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. RAPTOR: recursive abstractive processing for tree-organized retrieval. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=GN921JHCRw.

Zhiyu Shen, Ziming Wu, Fuming Lai, Shaobing Lian, and Yanghui Rao. Membuilder: Reinforcing llms for long-term memory construction via attributed dense rewards. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27868–27887, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1284.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MPNet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems*, volume 33, pp. 16857–16867, 2020.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. Is ChatGPT good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14918–14937, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.923.

Qingyue Wang, Yanhe Fu, Yanan Cao, Shuai Wang, Zhiliang Tian, and Liang Ding. Recursively summarizing enables long-term dialogue memory in large language models. *Neurocomputing*, 639: 130193, 2025. doi: 10.1016/j.neucom.2025.130193.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

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

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

Jing Xu, Arthur Szlam, and Jason Weston. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 5180–5197, 2022a.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-Mem: Agentic memory for LLM agents. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0593.

Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. Long time no see! open-domain conversation with long-term persona memory. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2639–2650, 2022b.

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

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 19724–19731, 2024.

Appendix Contents

A Notation and Pseudocode **14**

B Related Work **15**

C Datasets Statistics **15**

D Details of experiments **16**

D.1 Baselines 16

D.2 Metrics 16

D.3 Experiments Setting 18

E Comparison of Single-Perspectives and Multi-Perspectives **18**

E.1 Retrieval Performance 18

E.2 Generation Performance 18

F Further Experiments **19**

G Generalization Analysis **20**

G.1 Comparision on different retriever 20

G.2 Comparision on different reranker 20

G.3 Comparision on different generator 21

G.4 Comparison on different query types 21

H Hyperparameter sensitivity analysis **22**

I Retrieval and Generation Error Analysis **24**

J LLM Prompts **25**

K Case Study **27**

A Notation and Pseudocode

Table 5 summarizes the notation and pseudocode of our SMDI framework. The left part lists the main symbols used in different components, including memory construction, retrieval and projection, and coarse and fine reranking, while the right part provides a compact pseudocode overview of the complete framework.

Table 5: Notation (left) and pseudocode (right) of the SMDI framework.

Task & inputs	
$\mathcal{H} = \{S_i\}$	history of N sessions
$S_i = \{t_{i,j}\}$	i-th session of turns
q	user query
$\mathcal{M}_q$	final ranked set, $y = \mathrm{LLM}(q, \mathcal{M}_q)$
Memory construction	
$\mathcal{M},\ m$	source-memory set / a memory
$\mathcal{U}(m)$	units derived from m
$\mathcal{F}$	forms {Facet, Summary, Expansion}
u_f, u_s, u_e	facet / summary / expansion unit
w	neighboring window size
Retrieval & projection	
K_1, K_2, K_3	retrieval budgets
$L_f^{(1)}, L_f^{(2)}$	initial / refined lists under f
$s_{\mathrm{ce}},\ r_f(u)$	CE score / rank of u under f
$S(m)$	projection score (Eq. 1)
L_{src}	top-K_3 source candidates
Coarse reranking	
$r_p(m), r_c(m)$	projection / CE rank of m
$S^{(1)}(m)$	coarse score (Eq. 2)
λ, α	fusion weight / rank offset
$H,\ \mathcal{K}_H$	head count / head set
Fine reranking	
$\mathcal{B} = \{p, c, a\}$	ranking branches
$r_i(m)$	rank of m under branch i
$\bar{S}_i^{\mathrm{h}}, \bar{S}_i^{\mathrm{t}}$	mean head / tail scores
Δ_i	head–tail separation (Eq. 3)
$C_{(i,j)}$	Spearman consistency (Eq. 3)
$\bar{C}_i,\ S_i$	mean consistency / confidence (Eq. 4)
w_i	fusion weight (Eq. 5)
$S^{(2)}(m)$	fine score (Eq. 6)

Require: $\mathcal{H}, q; K_1, K_2, K_3, H; \lambda, \alpha; w$

Ensure: $\mathcal{M}_q$ and y

Stage 1: multi-form derivation

for all $m \in \mathcal{M}$ **do**
  decompose m into u_f
end for
for all S_i **do**
  group same-subject facets $\rightarrow u_s$
end for
for all $m \in \mathcal{M}$ **do**
  concat. w-neighbors $\rightarrow u_e$
end for

Stage 2: per-form retrieval

for all $f \in \mathcal{F}$ **do**
  $L_f^{(1)} \leftarrow \mathrm{dedup}(\text{sparse top-}K_1 \cup \text{dense top-}K_1)$
  $L_f^{(2)} \leftarrow \text{top-}K_2 \text{ by } s_{\mathrm{ce}}(q, u)$
end for

Stage 3: projection & fusion

project units via $\mathcal{U}(\cdot)$; compute $S(m)$
$L_{\mathrm{src}} \leftarrow \text{top-}K_3 \text{ by } S(m)$

Stage 4: coarse reranking

$r_p, r_c \leftarrow$ ranks by $S(m), s_{\mathrm{ce}}$
$S^{(1)} \leftarrow \lambda \frac{1}{r_p+\alpha} + (1-\lambda)\frac{1}{r_c+\alpha}$
$\mathcal{K}_H \leftarrow \text{top-}H \text{ by } S^{(1)}$

Stage 5: adaptive fine reranking

branches p, c, a: proj., CE, anchor

for all $i \in \mathcal{B}$ **do**
  $\Delta_i \leftarrow (\bar{S}_i^{\mathrm{h}} - \bar{S}_i^{\mathrm{t}})/(\max S_i - \min S_i)$
  $\bar{C}_i \leftarrow \frac{1}{|\mathcal{B}|-1} \sum_{j \neq i} C_{(i,j)}$
  $S_i \leftarrow \frac{1}{2} \max(\Delta_i, 0) + \frac{1}{2}\bar{C}_i$
end for
$w_i \leftarrow S_i / \sum_{j \in \mathcal{B}} S_j$
$S^{(2)}(m) \leftarrow \sum_{i \in \mathcal{B}} \frac{w_i}{r_i(m)}$
rerank $\mathcal{K}_H$ by $S^{(2)}(m)$; append tail
$\mathcal{M}_q \leftarrow \text{top-3}; y \leftarrow \mathrm{LLM}(q, \mathcal{M}_q)$
return $\mathcal{M}_q, y$

B Related Work

Personalized Memory Agents. Personalized conversational agents aim to adapt their responses to individual users based on user specific information, such as personal facts, preferences, and interaction history (Li et al., 2016; Zhang et al., 2018; Mazaré et al., 2018; Gao et al., 2023; Chen et al., 2024). As user interactions extend over longer periods, recent studies further consider settings where agents need to maintain and use information across multiple conversation sessions (Xu et al., 2022a;b; Jang et al., 2023; Maharana et al., 2024; Du et al., 2024; Wu et al., 2025; Li et al., 2025). In these settings, an agent is expected to recall information from earlier conversations, connect information distributed across different sessions, track changes in user information over time, and use relevant memories when responding to later queries.

C Datasets Statistics

We evaluate our method on LoCoMo (Maharana et al., 2024) and two settings of LongMemEval (Wu et al., 2025), namely LongMemEval-S and LongMemEval-M. All experiments are conducted using the official releases of LoCoMo[2] and LongMemEval[3]. The dataset statistics and question-type distributions are represented in Tables 6 and 7, respectively.

Table 6: Dataset statistics of LoCoMo and LongMemEval.

Statistic	LoCoMo	LongMemEval-S	LongMemEval-M
# Conversations / histories	10	500	500
# Evaluation questions	1,986	500	500
Avg. sessions / history	27.2	47.7	475.3
Avg. turns / history	588.2	493.5	4,894.0
Avg. turns / session	21.6	10.3	10.3
Avg. whitespace tokens / history	13.4K	78.4K	775.3K
Avg. whitespace tokens / turn	22.7	158.9	158.4
Max. sessions / history	32	62	490

Table 7: Distribution of question types in LoCoMo and LongMemEval.

Dataset	Question Type	# Questions	Percentage
LoCoMo	Single-hop Retrieval	841	42.3%
	Multi-hop Retrieval	282	14.2%
	Temporal Reasoning	321	16.2%
	Open-domain Knowledge	96	4.8%
	Adversarial	446	22.5%
	Total	**1,986**	**100%**
LongMemEval-S/M	Single-session User	70	14.0%
	Single-session Assistant	56	11.2%
	Single-session Preference	30	6.0%
	Multi-session	133	26.6%
	Temporal Reasoning	133	26.6%
	Knowledge Update	78	15.6%
	Total	**500**	**100%**

LoCoMo is a benchmark designed to evaluate long-term memory in multi-session conversations. It consists of extended dialogues spanning multiple interaction sessions, where questions may require retrieving and integrating information from distant parts of the conversation. This setting provides a comprehensive evaluation of memory retention and reasoning over long-term conversational histories.

[2] https://github.com/snap-research/locomo
[3] https://github.com/xiaowu0162/LongMemEval

LongMemEval is designed to evaluate the long-term interactive memory of conversational assistants. **LongMemEval-S** provides lengthy conversational histories containing both relevant evidence and substantial distracting information, requiring the model to accurately identify and utilize information accumulated across previous interactions. **LongMemEval-M** follows the same evaluation setting and question set as LongMemEval-S, but substantially increases the length of the associated interaction history. It therefore provides a more challenging setting for evaluating the effectiveness and scalability of memory systems when handling large-scale conversational histories.

D Details of Experiments

D.1 Baselines

We compare our framework with the baselines covering different memory construction and retrieval paradigms, including memory pools, recursive summarization, semantic segmentation, hierarchical retrieval, graph-based memory, structured memory notes, and multi-granularity memory systems. The details of these baselines are as follows:

MPNet (Song et al., 2020) is a pretrained sentence representation model that combines masked language modeling and permuted language modeling objectives. It learns contextualized text embeddings by considering token dependencies and positional information, and serves as a dense retrieval encoder for matching queries with memory contents.

RecurSum (Wang et al., 2025) adopts a recursive summarization mechanism for long-term dialogue memory. It continuously updates historical summaries by integrating newly observed conversations with previous summaries, forming compact representations of long-term dialogue history.

SeCom (Pan et al., 2025) constructs long-term memory by segmenting conversations into semantically coherent units and applying memory compression. It represents historical interactions as compressed semantic segments and organizes them as retrievable memory units.

RAPTOR (Sarthi et al., 2024) builds a hierarchical retrieval structure by recursively clustering and summarizing text chunks. The resulting tree representation organizes information at multiple abstraction levels, where each node stores summarized information from its corresponding subtree.

HippoRAG2 (Gutiérrez et al., 2025) integrates knowledge graph structures with retrieval-augmented generation. It organizes stored information into interconnected entities and relations, forming a graph-based memory structure for associative retrieval.

A-MEM (Xu et al., 2025) adopts a Zettelkasten-inspired memory mechanism that converts conversational experiences into structured memory notes. Each note contains contextual descriptions, keywords, tags, and links, allowing memories to be organized through explicit relationships.

MemGPT (Packer et al., 2023) introduces a virtual memory management framework for LLM agents. It separates the limited context window into working memory and external archival memory, and manages information transfer between these memory spaces.

Mem0 (Chhikara et al., 2025) provides a scalable memory framework for LLM agents by extracting, consolidating, and storing information from previous interactions. It organizes memories at multiple levels, including user-level, session-level, and interaction-level memories.

MemGAS (Xu et al., 2026) introduces a multi-granularity memory framework that organizes historical information at different abstraction levels. It models memory units with Gaussian Mixture Models and performs memory association across different granularities.

D.2 Metrics

Retrieval Metrics. Following the evaluation setting of MemGAS Xu et al. (2026), we use Recall@k and NDCG@k with $k \in \{3, 5, 10\}$ to evaluate the retrieval results. For query q, let $\mathcal{G}$ be the set of gold memories and $\mathcal{R}^k$ the top-k retrieved memories. Recall@k measures the proportion of queries

for which all gold memories are retrieved:

$$\text{Recall@}k = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left[\mathcal{G}_i \subseteq \mathcal{R}_i^k\right]. \tag{7}$$

NDCG@k further accounts for the ranking positions of relevant memories. Let $r_{i,j} \in \{0,1\}$ indicate whether the memory ranked at position j is relevant to query q_i. We compute

$$\text{NDCG@}k = \frac{1}{N}\sum_{i=1}^{N}\frac{\text{DCG}_i@k}{\text{IDCG}_i@k}, \qquad \text{DCG}_i@k = r_{i,1} + \sum_{j=2}^{k}\frac{r_{i,j}}{\log_2 j}, \tag{8}$$

where $\text{IDCG}_i@k$ denotes the maximum possible DCG@k, obtained by ranking all relevant memories before irrelevant ones.

Generation Metrics. For generation, we follow MemGAS and report GPT4o-J, F1, BLEU4, ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore. GPT4o-J uses GPT-4o as an LLM-based evaluator to determine whether a generated response correctly answers the question with respect to the reference answer. Its score is the percentage of responses judged as correct:

$$\text{GPT4o-J} = \frac{1}{N}\sum_{i=1}^{N} J_i \times 100, \tag{9}$$

where $J_i = 1$ if the response is judged correct and $J_i = 0$ otherwise. F1 measures token-level overlap between the generated and reference answers:

$$P = \frac{n_{\text{match}}}{n_{\text{pred}}}, \qquad R = \frac{n_{\text{match}}}{n_{\text{gold}}}, \qquad F_1 = \frac{2PR}{P+R}. \tag{10}$$

BLEU4 evaluates n-gram precision up to four-grams with a brevity penalty:

$$\text{BLEU4} = \text{BP}\,\exp\left(\frac{1}{4}\sum_{n=1}^{4}\log p_n\right), \tag{11}$$

where p_n is the modified n-gram precision. ROUGE-1 and ROUGE-2 measure unigram and bigram overlap, while ROUGE-L measures sequence overlap based on the longest common subsequence. BERTScore evaluates semantic similarity by matching contextual token representations between the generated and reference answers.

Table 8: Experimental configurations and hyperparameter settings on LongMemEval-S.

Component	**Configuration**	**Component**	**Configuration**
Models		*Model Setting*	
Memory construction model	DeepSeek-V4-Flash	Generation temperature	0
Evidence filtering model	DeepSeek-V4-Flash	Maximum output length	4000
Answer generation model	GPT-4o-mini	Thinking mode	Disabled
Judge model	GPT-4o	Reranker max length	512
Embedding model	Contriever	Batch size	256
Reranking model	BGE-reranker-large		
Retrieval and Reranking		*Memory Construction and Filtering*	
Head candidate size (H)	10	Expansion window size	3
Coarse fusion weight (λ)	0.5	Expansion stride	1
RRF offset (α)	10	Memory ordering	Chronological
Per-form retrieval budget (K_1)	20		
Per-form cross-encoder budget (K_2)	40		
Source-memory projection budget (K_3)	20		

D.3 Experiments Setting

All experiments are conducted on a single machine equipped with an NVIDIA RTX 4090 GPU. Table 8 provides an overview of the models, configurations, and hyperparameters adopted throughout the experiments.

Memory Construction and Evidence Filtering. We use DeepSeek-V4-Flash[4]as the memory construction model to generate the multi-form memory units used in SMDI. Specifically, DeepSeek-V4-Flash is used to construct the different memory units described, including the factual, thematic, and contextual interpretations of historical interactions. Since concatenating all retrieved source memories may exceed the context window of the answer generator, we introduce an additional extractive evidence filtering step before generation. Given the retrieved memories and the query, DeepSeek-V4-Flash is used to identify query-relevant evidence while preserving the original wording of the selected content. This filtering procedure is applied uniformly to all compared methods to ensure a fair comparison.

Answer Generation and Evaluation Protocol. For answer generation, we follow the generation and evaluation protocol of MemGAS (Xu et al., 2026). All methods use the same GPT-4o-mini answer generator and the same prompt template. After retrieval and reranking, the selected memories at top-K positions, where $K \in \{3, 5, 10\}$, are first reordered chronologically and then provided to the filtering–generation pipeline together with the query and its associated temporal information. The decoding temperature of GPT-4o-mini is set to 0, while all other generation configurations remain identical across different methods.

E Comparison of Single-Perspectives and Multi-Perspectives

E.1 Retrieval Performance

As shown in Table 9, the three retrieval pathways retrieve substantially different sets of source memory, with relatively low pairwise Jaccard similarities that further decrease as k increases, indicating clear complementarity among the different memory forms. Among the individual pathways, u_e achieves the strongest gold-evidence retrieval performance, while u_f and u_s provide additional evidence that is not fully covered by u_e. In particular, combining u_e with either u_f or u_s improves R-All and evidence coverage over using u_e alone, whereas removing u_e leads to a clear performance drop. These results show that u_e achieves the strongest performance among individual retrieval forms, while u_f and u_s provide additional relevant evidence, supporting the necessity of multi-form retrieval.

Table 9: Complementarity analysis of the retrieval pathways on LongMemEval-S. Jacc.@k denotes the Jaccard similarity between the top-k retrieved source-memory sets. R-All and R-Any measure whether all or at least one gold evidence is retrieved, respectively, while Cov. denotes the average proportion of retrieved gold evidence.

Retrieval-Set Similarity				**Gold-Evidence Retrieval**			
Compared Pair	**Jacc.@3**	**Jacc.@5**	**Jacc.@10**	**Memory Form**	**R-All**	**R-Any**	**Cov.**
u_e vs. u_f	0.387	0.323	0.280	u_e	78.80	96.80	88.60
u_e vs. u_s	0.425	0.351	0.316	u_f	50.40	77.80	65.24
u_f vs. u_s	0.505	0.438	0.397	u_s	59.40	85.40	73.56
u_e vs. $w/o\ u_e$	0.386	0.324	0.314	$w/o\ u_s$	81.40	98.40	93.85
u_f vs. $w/o\ u_f$	0.429	0.371	0.338	$w/o\ u_f$	80.00	98.60	94.53
u_s vs. $w/o\ u_s$	0.452	0.389	0.381	$w/o\ u_e$	60.00	86.00	75.91

E.2 Generation Performance

Table 10 compares the generation performance of different memory units and the full SMDI framework on LongMemEval-M and LoCoMo. We evaluate three individual memory representations, including Expansion-Unit, Facet-Unit, and Summary-Unit, and compare them with the complete

[4] https://api-docs.deepseek.com/

Table 10: Generation results of single-form and full model on LongMemEval-M and LoCoMo.

Memory Unit	GPT4o-J	F1	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
LongMemEval-M							
Expansion-Unit (u_e)	53.80	18.42	4.02	19.05	9.07	17.54	85.91
Facet-Unit (u_f)	42.60	16.76	3.32	18.39	8.35	16.60	84.30
Summary-Unit (u_s)	44.20	16.52	3.86	18.09	8.74	16.46	84.97
Full model	57.80	19.83	4.20	20.56	9.87	18.96	86.33
LoCoMo							
Expansion-Unit (u_e)	36.02	17.06	3.58	17.44	9.01	16.71	86.70
Facet-Unit (u_f)	41.78	19.40	4.09	19.76	9.95	18.62	87.18
Summary-Unit (u_s)	37.94	18.04	3.72	18.49	9.03	17.43	86.99
Full model	46.50	23.58	5.69	25.53	12.99	24.05	88.04

model that integrates all memory units. The results show that each memory unit provides complementary information for generation, while combining multiple memory views consistently achieves better performance across different datasets.

F Further Experiments

Study of head reranking weights. Table 11 compares the proposed adaptive head fusion with fixed uniform fusion across three datasets. Adaptive fusion consistently matches or outperforms uniform fusion on all retrieval metrics, with more evident improvements at smaller retrieval depths. On LongMemEval-S, it improves Recall@3 from 81.20 to 82.20 and NDCG@3 from 90.80 to 90.92. On LongMemEval-M, the gains reach 1.38 points in NDCG@3 and 1.20 points in Recall@5. Similar improvements are observed on LoCoMo, including gains of 0.61 and 0.65 points in Recall@3 and Recall@5, respectively. These results show that adaptive fusion provides more reliable head-candidate ranking than assigning uniform weights to different ranking signals.

Table 11: Comparison of fixed uniform and adaptive head fusion across three datasets.

Dataset	**Fusion strategy**	**@3**		**@5**		**@10**	
		Recall	**NDCG**	**Recall**	**NDCG**	**Recall**	**NDCG**
LongMemEval-S	Fixed uniform fusion	81.20	90.80	**90.20**	91.36	**97.00**	93.11
	SMDI (Ours)	**82.20**	**90.92**	**90.20**	**92.10**	**97.00**	**93.38**
LongMemEval-M	Fixed uniform fusion	63.60	72.36	74.80	77.92	**81.20**	79.56
	SMDI (Ours)	**64.20**	**73.74**	**76.00**	**78.05**	**81.20**	**79.64**
LoCoMo	Fixed uniform fusion	65.79	64.05	72.86	67.41	**82.09**	70.74
	SMDI (Ours)	**66.40**	**64.99**	**73.51**	**68.44**	**82.09**	**70.87**

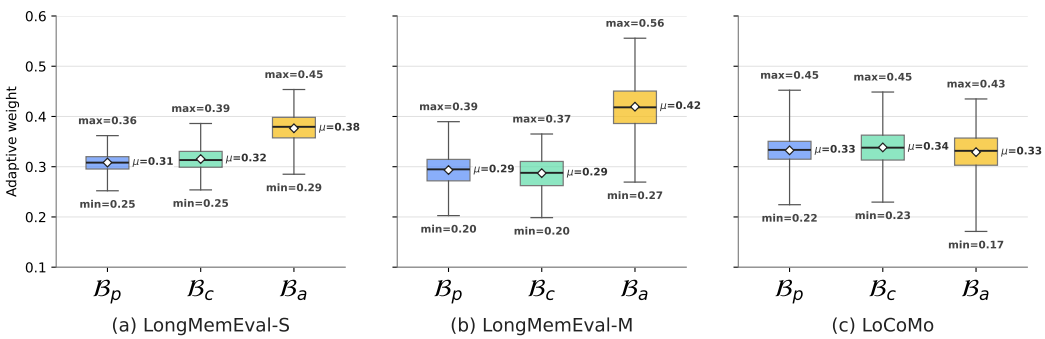


Figure 7: Adaptive weight allocation among B_p, B_c, and B_a across different datasets.

Head Weight Analysis. Figure 7 shows the head weight distributions of the projection branch B_p, the cross-encoder branch B_c, and the anchor branch B_a across the three datasets. On LongMemEval-S and LongMemEval-M, B_a receives the largest average weight, which further confirms the importance of the anchor ranking derived from the Expansion-Unit in fine-grained reranking. Meanwhile, the weight distributions of all three branches vary across queries and datasets, rather than remaining fixed. In particular, the three branches receive more balanced weights on LoCoMo, while the anchor branch is assigned greater importance on LongMemEval. These results show that the proposed adaptive reranking can dynamically adjust the contributions of different ranking signals according to the query.

We further analyze the adaptive weights across different query types. As illustrated in Figure 8, the weights assigned to B_p, B_c, and B_a vary across query types, indicating that different queries rely on different ranking signals. This further shows that the proposed adaptive reranking strategy can adjust the contribution of each branch according to the characteristics of the query rather than using a fixed weighting scheme.

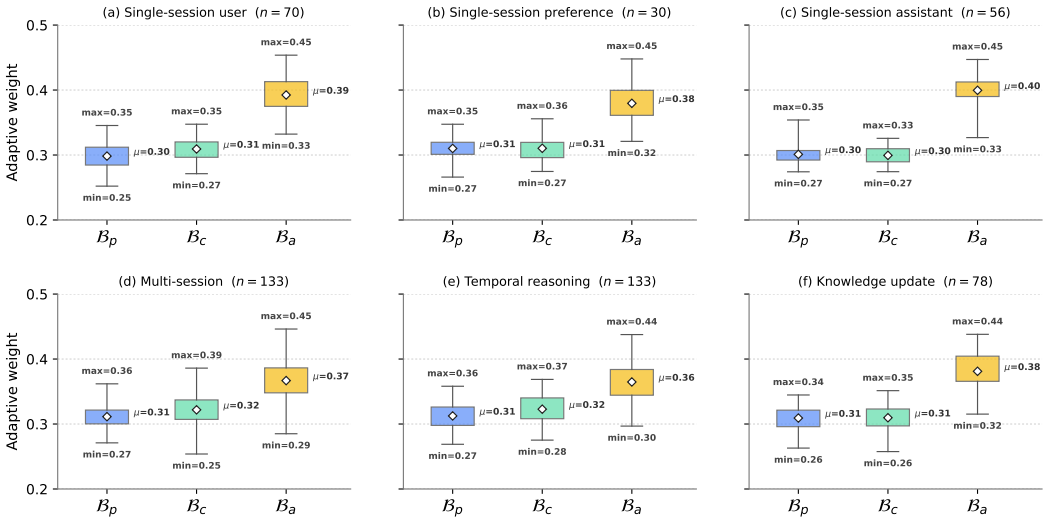


Figure 8: Adaptive weight allocation among B_p, B_c, and B_a across different query types.

G Generalization Analysis

We evaluate the robustness of our SMDI framework under different retriever, reranker, and generator configurations.

G.1 Comparision on different retriever

Table 12 shows the performance of SMDI with different base retrievers on LoCoMo. SMDI consistently achieves the best performance across MiniLM, MPNet, and Contriever, demonstrating its robustness to different retrieval backbones. The improvements are particularly notable on Recall@3 and NDCG@3, showing that SMDI can retrieve relevant memories more effectively and rank them higher among the top candidates. Moreover, its Recall@3 remains stable across all three retrievers. These results demonstrate that SMDI can provide robust and high-quality memory retrieval for downstream QA.

G.2 Comparision on different reranker

Table 13 investigates the effect of different rerankers on **SMDI**. Overall, retrieval performance improves as the reranker scales from BGE-Base (278M) to BGE-Large (560M), with consistent gains across all evaluation metrics, suggesting that larger model capacity enables more effective relevance modeling. It is interesting to note that BGE-v2-m3 (568M) further outperforms BGE-Large despite their comparable parameter sizes, achieving the best performance across all metrics. We argue that

Table 12: Retrieval Performance based on different retriever on LoCoMo.

Model	Recall@3	NDCG@3	Recall@5	NDCG@5	Recall@10	NDCG@10
Base Retriever: MiniLM						
MiniLM (Wang et al., 2020)	42.55	44.19	52.37	49.01	67.98	54.59
MPC (Lee et al., 2023)	42.30	43.59	51.21	48.08	68.03	54.03
RecurSum (Wang et al., 2025)	44.76	46.82	54.73	51.64	72.16	57.52
SeCom (Pan et al., 2025)	45.77	47.25	54.93	51.70	71.15	57.37
MemGAS (Xu et al., 2026)	47.73	49.11	56.60	53.46	71.30	58.59
SMDI (Ours)	**67.00**	**65.18**	**75.68**	**68.54**	**83.05**	**71.22**
Base Retriever: MPNet						
MPNet (Song et al., 2020)	45.92	47.71	53.98	51.79	68.58	56.88
MPC (Lee et al., 2023)	45.47	47.35	54.08	51.68	68.28	56.59
RecurSum (Wang et al., 2025)	49.50	51.15	59.47	56.16	76.64	61.99
SeCom (Pan et al., 2025)	47.53	49.03	57.05	53.57	70.90	58.57
MemGAS (Xu et al., 2026)	52.77	54.63	62.79	59.56	80.51	65.48
SMDI (Ours)	**66.90**	**65.48**	**76.44**	**69.26**	**82.85**	**71.62**
Base Retriever: Contriever						
Contriever (Izacard et al., 2022)	49.90	52.15	58.26	56.29	71.80	60.92
MPC (Lee et al., 2023)	49.50	51.47	57.45	55.53	71.85	60.47
RecurSum (Wang et al., 2025)	47.23	48.99	59.01	54.58	74.97	60.07
SeCom (Pan et al., 2025)	55.24	57.90	64.80	62.36	78.30	66.97
MemGAS (Xu et al., 2026)	57.30	58.76	67.32	63.62	81.82	68.42
SMDI (Ours)	**66.40**	**64.99**	**73.51**	**68.44**	**82.09**	**70.87**

this improvement may be attributed to the stronger representation and relevance modeling capabilities of the former. Overall, these results show that **SMDI** consistently benefits from stronger rerankers, while remaining effective across different model scales.

Table 13: Retrieval performance with different rerankers on LoCoMo.

Method	**Reranker**	**@3**		**@5**		**@10**	
		Recall	**NDCG**	**Recall**	**NDCG**	**Recall**	**NDCG**
SMDI (Ours)	BGE-Base (278M)	63.82	61.94	73.06	65.67	81.18	68.49
	BGE-Large (560M)	66.40	64.99	73.51	68.44	82.09	70.87
	BGE-v2-m3 (568M)	**67.05**	**65.44**	**75.38**	**68.76**	**82.29**	**71.11**

G.3 Comparision on different generator

To evaluate whether the effectiveness of SMDI generalizes across different generators, we conduct generation experiments on LongMemEval-S using GPT-4o-mini, Qwen3-8B, and Qwen3-1.7B while keeping the retrieval setting unchanged. As shown in Table 14, SMDI consistently outperforms the non-oracle baselines across different generators on both answer accuracy and text similarity metrics. For example, its GPT-4o judge scores reach 64.80, 52.40, and 42.20 with GPT-4o-mini, Qwen3-8B, and Qwen3-1.7B, respectively, all exceeding the corresponding MemGAS results. Although generation performance generally decreases with smaller generators, the advantage of SMDI remains consistent, indicating that the improvements brought by our memory retrieval and reranking framework are not tied to a specific generator and can provide higher-quality evidence for models with different generation capabilities.

G.4 Comparison on different query types

To examine how different memory forms contribute to different query types, we compare Recall-all@3 on LongMemEval-S using individual retrieval forms, their pairwise combinations, and all three forms together. As shown in Table 15, the contextual form u_e provides the strongest individual retrieval performance across most query types, while u_f and u_s provide complementary evidence

Table 14: Generation robustness under different base generators on LongMemEval-S.

Model	GPT-4o-J	F1	B-4	R-1	R-2	R-L	BERTScore
Base Generator: GPT-4o-mini							
Contriever (Izacard et al., 2022)	55.40	13.78	2.21	14.46	6.93	12.89	83.70
SeCom (Pan et al., 2025)	56.00	12.95	2.25	13.80	6.09	11.93	83.51
MemGAS (Xu et al., 2026)	60.20	20.38	4.22	21.05	10.47	19.47	85.21
Oracle	67.20	23.53	7.48	23.96	12.80	22.91	88.72
SMDI(Ours)	**64.80**	**22.13**	**5.24**	**22.90**	**11.84**	**21.51**	**86.67**
Base Generator: Qwen3-8B							
Contriever (Izacard et al., 2022)	50.60	14.76	1.85	15.37	7.44	13.81	83.73
SeCom (Pan et al., 2025)	50.60	14.83	1.89	17.66	8.45	14.08	85.15
MemGAS (Xu et al., 2026)	51.40	20.57	3.50	21.17	10.15	19.09	85.54
Oracle	60.20	31.06	4.39	31.07	17.56	30.20	86.98
SMDI(Ours)	**52.40**	**21.21**	**4.25**	**21.89**	**11.06**	**20.34**	**86.24**
Base Generator: Qwen3-1.7B							
Contriever (Izacard et al., 2022)	36.40	9.39	1.30	9.92	4.37	8.80	82.62
SeCom (Pan et al., 2025)	36.00	8.50	1.07	9.26	3.66	7.91	82.26
MemGAS (Xu et al., 2026)	38.00	16.79	3.60	17.58	7.78	16.19	84.84
Oracle	57.60	27.28	9.32	27.68	13.93	27.22	87.84
SMDI(Ours)	**42.20**	**18.74**	**4.16**	**19.49**	**9.75**	**18.12**	**86.01**

for more complex queries. Combining all three forms achieves the best overall Recall-all@3 of 82.20, with particularly clear improvements on multi-session, single-session preference, and temporal reasoning queries. In contrast, for single-session assistant and knowledge-update queries, u_e alone already retrieves most required evidence, leaving limited room for additional gains. These results indicate that the simple queries can often be handled effectively by a single contextual form, while benefit of multi-form retrieval is more pronounced when relevant evidence is distributed across multiple memories or requires different semantic perspectives.

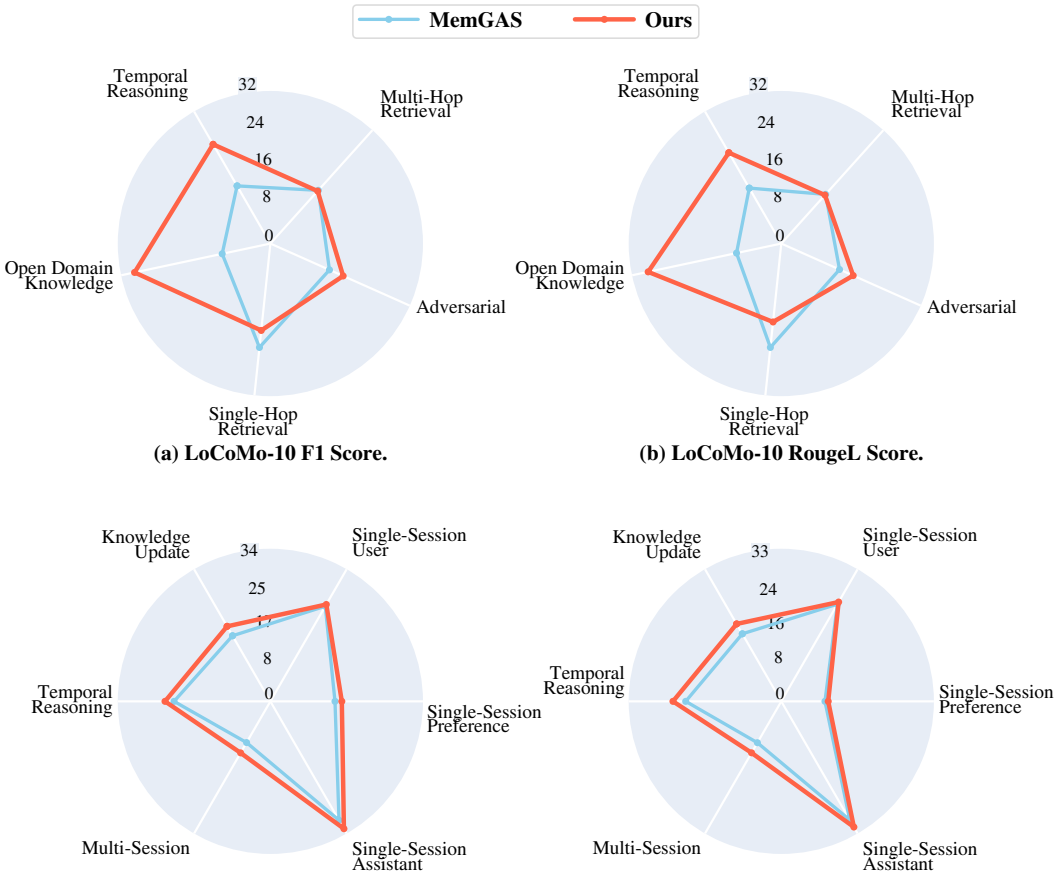
Table 15: Comparison of retrieval performance (Recall-all@3) across query types and retrieval views on LongMemEval-S. u_e, u_f, and u_s denote local context, atomic memory, and topic summary, respectively.

Query Type	u_e	u_f	u_s	$w/o\ u_f$	$w/o\ u_s$	$w/o\ u_e$	**full forms**
knowledge-update	92.31	38.46	53.85	92.31	92.31	52.56	**92.31**
multi-session	60.15	47.37	46.62	63.91	66.17	52.63	**69.17**
single-session-assistant	100.00	48.21	48.21	100.00	100.00	48.21	**100.00**
single-session-preference	80.00	66.67	80.00	83.33	83.33	80.00	**90.00**
single-session-user	95.71	45.71	71.43	95.71	95.71	67.14	**94.29**
temporal-reasoning	71.43	60.15	69.17	71.43	74.44	68.42	**74.44**
Overall	78.80	50.40	59.40	80.00	81.40	60.00	**82.20**

We further examine whether the retrieval improvements translate into better generation performance across different query types. As shown in Figure 9, SMDI generally achieves higher F1 and ROUGE-L scores than MemGAS on both LoCoMo and LongMemEval-S. The improvements are more evident on query types involving temporal reasoning, multi-hop or multi-session information, and open-domain knowledge, where relevant evidence is often distributed across different memories. These results are consistent with the retrieval analysis above and further show that multi-form retrieval provides more useful evidence for downstream response generation.

H Hyperparameter Sensitivity Analysis

Analysis of λ Figure 10 presents the sensitivity of the coarse-stage fusion weight λ on LongMemEval-S. When λ increases from 0 to 0.3, both Recall-all and NDCG consistently improve across different retrieval depths, indicating that introducing the projection ranking is beneficial for the coarse reranking stage. After $\lambda = 0.3$, changes in λ have limited impact on performance across



(a) LoCoMo-10 F1 Score. **(b) LoCoMo-10 RougeL Score.**

(c) LongMemEval-s F1 Score. **(d) LongMemEval-s RougeL Score.**

Figure 9: Generation performance across different query types on LoCoMo and LongMemEval-S.

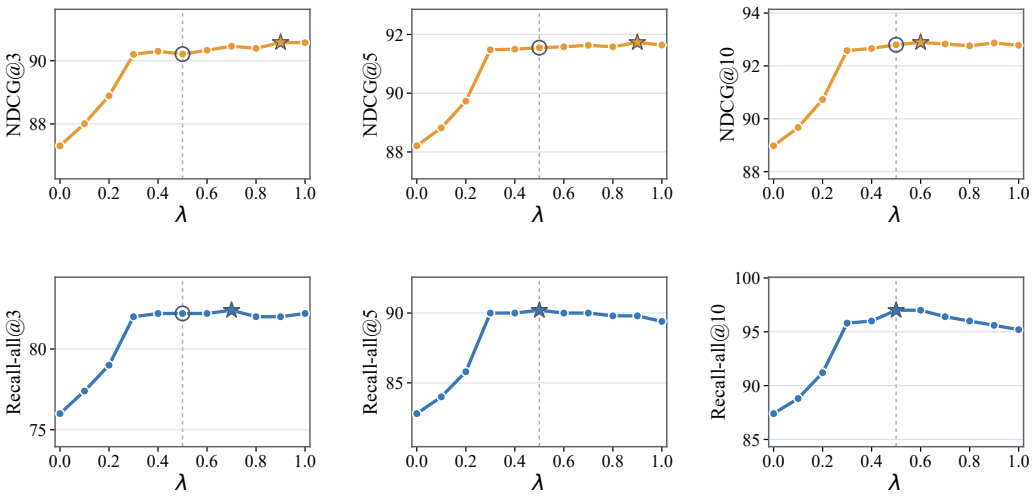


Figure 10: Sensitivity analysis of the coarse-stage fusion weight λ on LongMemEval-S.

all metrics. We can observe that the setting of $\lambda = 0.5$ achieves consistently strong performance and stays close to the best result on each metric. Therefore, we use $\lambda = 0.5$ as the default setting in our experiments.

Analysis of α Table 16 presents the sensitivity of the coarse-stage offset α on LongMemEval-S. The performance remains stable across a broad range of α values. While $\alpha = 10$ shows consistently strong performance across different metrics. Notably, increasing α beyond 10 leads to a gradual decrease in NDCG. In contrast, the recall metrics remain stable and only decline at $\alpha = 100$. Therefore, we use $\alpha = 10$ as the default setting in our experiments.

Table 16: Sensitivity analysis of the coarse-stage RRF offset α on LongMemEval-S.

α	Recall-all			NDCG		
	@3	@5	@10	@3	@5	@10
0.1	82.20	89.80	97.00	90.24	91.49	92.81
0.5	82.20	89.80	97.00	90.24	91.49	92.81
1	82.20	89.80	97.00	90.24	91.49	92.81
2	82.20	89.80	97.00	90.21	91.47	92.79
5	82.20	89.80	97.00	90.21	91.47	92.79
10	82.20	90.20	97.00	90.92	92.10	93.38
20	82.20	90.20	97.00	90.19	91.46	92.68
50	82.20	90.20	97.00	90.10	91.29	92.34
100	82.00	90.00	96.60	90.00	91.09	92.11

Analysis of head size H Figure 11 presents the sensitivity of the head candidate size H on LongMemEval-S. As H increases, the performance consistently improves across NDCG@3, NDCG@5, and NDCG@10, followed by a clear saturation trend. In particular, the gains become marginal when H reaches over 8, indicating that a moderate value is sufficient to capture most of the useful information. Further increasing H yields only negligible improvements or slight fluctuations. Overall, these results demonstrate that **SMDI** is relatively robust to the choice of H within a reasonable range. We therefore set $H = 10$ as the default in our experiments.

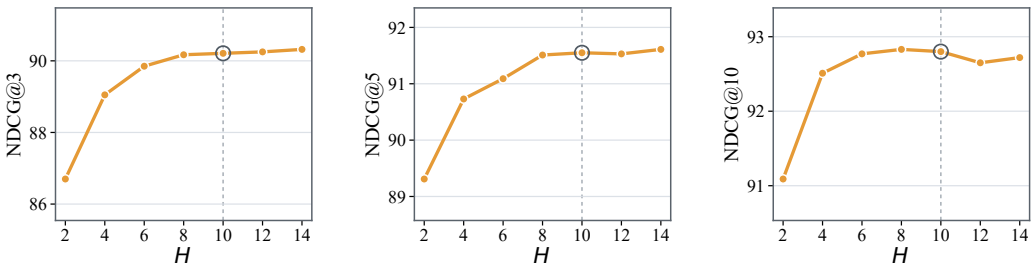


Figure 11: Sensitivity analysis of the head candidate size H on LongMemEval-S.

Analysis of retrieval budget To further investigate the impact of retrieval budget allocation, we provide a more comprehensive evaluation under different $K_1/K_2/K_3$ configurations in Figure 12. These budgets control the number of candidates retained at different retrieval stages of SMDI, and their variations reflect different trade-offs between retrieval effectiveness and inference cost. The results show that increasing retrieval budgets generally improves performance by preserving more potentially relevant memories, while introducing additional computational overhead. Among the evaluated configurations, $K_1/K_2/K_3 = 20/40/20$ achieves a favorable balance between retrieval performance and efficiency, which is adopted as the default setting in our experiments.

I Retrieval and Generation Error Analysis

To analyze the sources of errors, we divide the answers into four groups according to the correctness of retrieval and generation. As shown in Figure 13, correct retrieval with correct generation accounts for the largest proportion across all three datasets, indicating that successful generation is closely related to accurate retrieval. Meanwhile, a considerable number of cases have correct retrieval but incorrect

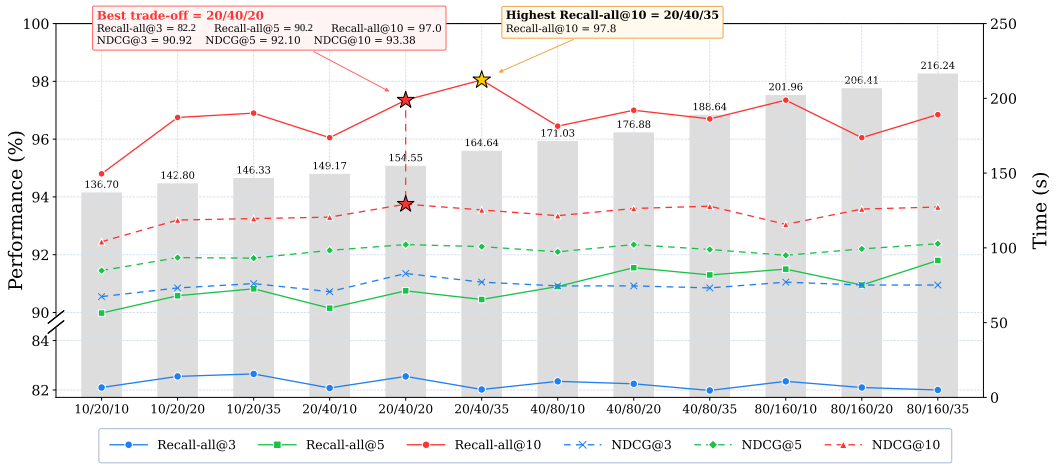


Figure 12: Performance–efficiency trade-off under different retrieval budgets. Stars highlight the selected trade-off and best-performing configurations.

generation, indicating that relevant evidence is not always used effectively by the generator. In contrast, incorrect retrieval with correct generation occurs only rarely, further showing the importance of retrieving the correct evidence. We also observe that LongMemEval-M and LoCoMo contain more cases where both retrieval and generation fail than LongMemEval-S. Overall, the results suggest that end-to-end performance depends on both accurate memory retrieval and effective use of the retrieved evidence.

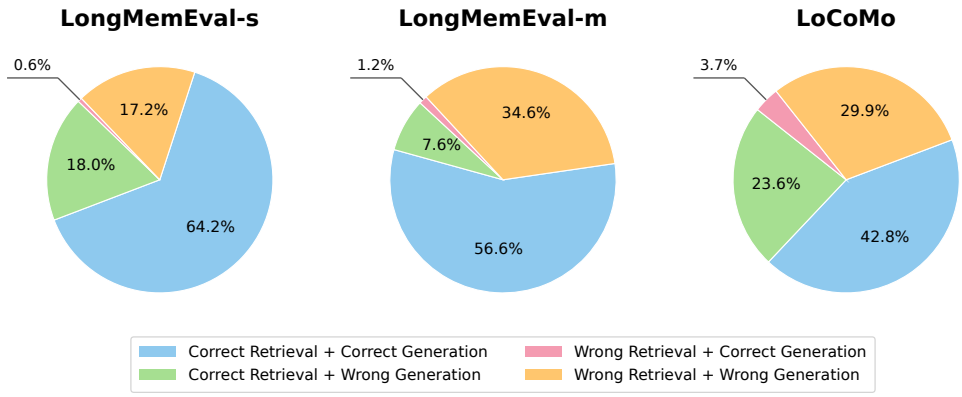


Figure 13: Retrieval and Generation Error Analysis.

J LLM Prompts

Figures 14-18 present the LLM prompts used in different stages of our framework. Figure 14 shows the prompt for intra-memory facet decomposition, which extracts standalone Facet-Units from a dialogue window together with their semantic and temporal information. Figure 15 presents the prompt for cross-memory topic summarization, which aggregates related Facet-Units from the same session and subject into a concise Summary-Unit. Figure 16 shows the evidence filtering prompt, which removes irrelevant or redundant retrieved information while preserving evidence relevant to the current query. Figure 17 presents the response generation prompt, where the LLM generates the final answer based on the filtered evidence and temporal information. Finally, Figure 18 shows the automatic judging prompt used to determine whether the generated response correctly answers the question with respect to the reference answer.

Demonstration of Intra-Memory Facet Decomposition

Description: This prompt extracts durable and question-answerable atomic memories from a dialogue window. Each memory should be independently understandable and traceable to its supporting dialogue turns.

Input: **Session Date :** {session_datetime}

Dialogue Window : {dialogue_window}

Prompt Details:

You are a memory construction module. Extract only durable, question-answerable memories from the current dialogue window.
Do not copy every utterance. Ignore greetings, generic praise, empty acknowledgments, and information that is not useful for future question answering.
Each claim must be concise and understandable without the original dialogue. Use the speaker or entity actually associated with the fact. Preserve explicit or relative temporal information. Do not infer unsupported facts.

For each extracted memory, provide the following fields:

Subject : the speaker or entity the memory is about
Claim : one concise standalone memory claim
Memory Type : the semantic category that describes the nature of the extracted memory
Temporal Status : the temporal state indicating when or how the extracted memory applies
Time Text : explicit or relative time expression; leave empty if unavailable
Predicate : a short normalized predicate
Object : the main object or complement
Evidence Dialogue IDs : supporting dialogue IDs, e.g., D1:3

Output:

Memory 1

Subject : Taylor
Claim : Taylor is training for a marathon.
Memory Type : goal
Temporal Status : current
Time Text : this year
Predicate : trains_for
Object : a marathon
Evidence Dialogue IDs : D1:1

Memory 2

Subject : Taylor
Claim : Taylor runs five days a week.
Memory Type : habitual
Temporal Status : habitual
Time Text :
Predicate : runs
Object : five days a week
Evidence Dialogue IDs : D1:3

Figure 14: Prompt of Intra-memory facet decomposition.

Demonstration of Topic-level Memory Summary

Description: This prompt compresses related atomic memories from the same session and subject into one concise topic-level summary, while preserving factual faithfulness and correct subject attribution.

Input: **Subject :** {subject}

Atomic Memories : {atomic_memories}

Prompt Details:

You are a topic-level memory summarization module. Given atomic memories from the same session and concerning the same subject, write one concise, standalone subject-level summary.
Use only information supported by the provided atomic memories. Preserve the correct person or entity attribution and temporal status. Do not add new facts, causal explanations, or details absent from the atomic memories.
Provide the following fields:

Summary : one concise standalone subject-level summary
Memory Type : the semantic category that describes the nature of the extracted memory
Temporal Status : the temporal state indicating when or how the extracted memory applies
Importance : a score between 0 and 1

Output:

Summary : Taylor is currently training for a marathon and follows a regular running schedule.
Memory Type : mixed
Temporal Status : current
Importance : 0.82

Figure 15: Prompt of Cross-memory topic summary.

Demonstration of Multi-view Evidence Filtering

Description: This prompt filters retrieved dialogue turns, atomic memories, and topic-level summaries, retaining only evidence directly relevant to the current question.

Input:
Retrieved Evidence : {retrieved_texts}
Question Date : {question_date}
Question : {question}

Prompt Details:
You are an intelligent dialog bot. You will be shown History Dialogs and corresponding multi-granular information.
Filter the History Dialogs, summaries, and keywords to extract only the parts directly relevant to the Question.
Preserve original tokens and do not paraphrase the selected evidence. Remove irrelevant turns, redundant information, and non-essential details. Do not introduce information that is absent from the retrieved evidence.

Output:
Filtered Evidence:
Dialogue : D1:1 Taylor: I started training for a marathon this year.
Dialogue : D1:3 Taylor: I run five days a week.
Atomic Memory : Taylor is training for a marathon.
Topic Summary : Taylor follows a regular running schedule.

Figure 16: Prompt of evidence filtering.

Demonstration of Answer Generation

Description: This prompt generates one concise, coherent, and helpful answer using the filtered evidence retrieved for a user question.

Input:
Filtered Evidence : {filtered_history}
Question Date : {question_date}
Question : {question}

Prompt Details:
You are an intelligent dialog bot. You will be shown History Dialogs.
Please read, memorize, and understand the given Dialogs, then generate one concise, coherent, and helpful response for the Question.
Ground the answer in the provided evidence. Respect the temporal information in the question and avoid adding unsupported details.

Output:
Answer : Taylor is currently training for a marathon and runs five days a week.

Figure 17: Prompt of answer generation.

K Case Study

We further conduct case studies to qualitatively examine how SMDI retrieves and utilizes historical memories under different reasoning scenarios. Figures 19-25 illustrate representative examples involving numerical aggregation, temporal reasoning, multi-form memory retrieval, adaptive fine reranking, multi-session evidence composition, evolving user states, and retrieval–generation errors.

As shown in Figure 19, answering some queries requires combining information distributed across multiple source memories rather than relying on a single retrieved item. SMDI successfully retrieves the relevant memories and composes their evidence to produce the correct answer. The temporal-ordering example further shows that retrieving related events alone is insufficient, the model must also distinguish when each event occurred to determine the correct order.

Figure 20 illustrates how different memory forms provide complementary views of the same source memory. The Facet-Unit captures the key factual event, the Summary-Unit connects it to the broader session-level context, and the Expansion-Unit preserves nearby conversational information. Although these representations organize the memory differently, their provenance links allow the retrieved evidence to be traced back to the original source memory. This enables the retrieval process to

Demonstration of Automatic Answer Judging

Description: This prompt evaluates whether a generated model response contains the information required by the reference answer. It is used to compute the GPT-4o-based generation accuracy metric.

Input: **User Question :** {question}

Reference Answer : {reference_answer}

Model Response : {response}

Prompt Details:

I will give you a question, a reference answer, and a response from a model.

Return [[yes]] if the model response contains the reference answer. Return [[yes]] if the response is equivalent to the correct answer or contains all intermediate information required to obtain the reference answer.

Return [[no]] if the response contains only a subset of the information required by the reference answer, contradicts the reference answer, or does not answer the question.

Output [[yes]] or [[no]] only.

Output:

Judgment : [[yes]]

Figure 18: Prompt of answer judging.

Same-Query Response Comparison

Description: Each case shows the query, the ground-truth answer, the retrieved Top-3 source memories, and the final responses from SMDI and MemGAS-style.

Case 1. Multi-session numerical aggregation

Query: What is the total amount I spent on luxury items in the past few months?
Gold answer: $2,500

Retrieved Source Memories (Top-3)

May 25, 2023	"I was so excited about that designer handbag I just got from Gucci for $1,200. It's one of those classic pieces I've wanted for years, and I finally decided to treat myself. The quality is amazing and I know it will be a staple in my wardrobe for a long time."
May 28, 2023	"I ended up getting a pair of leather boots from a high-end Italian designer that I got for $500. They were definitely a splurge, but I love them and I know I'll wear them a lot. The craftsmanship is excellent and they go with so many outfits."
May 24, 2023	"I recently bought a luxury evening gown for a wedding. It was a big purchase, $800, but it fit the occasion perfectly and made the event feel extra special. I've been looking for the right dress for a while and I'm really happy with it."
Evidence composition:	$1,200 + $500 + $800 = $2,500.

Response :
- **SMDI (Ours):** $2,500 (Gucci handbag $1,200 + Italian leather boots $500 + luxury gown $800). ✓
- **MemGAS-style:** The total amount you spent on luxury items in the past few months is **$1,700**. ✗

Case 2. Temporal ordering

Query: Which device did I get first, the Samsung Galaxy S22 or the Dell XPS 13?
Gold answer: Samsung Galaxy S22

Retrieved Source Memories (Top-3)

Samsung Galaxy S22	"I recently got the Samsung Galaxy S22 on February 20th, and I've been really enjoying it. The camera is great and it feels like a nice upgrade. I've been using it a lot since then and it's been working well for me."
Dell XPS 13	"I pre-ordered the Dell XPS 13 on January 28th, and it finally arrived on February 25th. I've been setting it up and looking forward to using it for work. The setup process was smooth and it's been great so far."
Related purchase history	"I also picked up a pair of Madewell boots recently, which was another purchase around that time, though unrelated to the device ordering question. They're really comfortable and have been perfect for everyday wear."
Temporal order:	Samsung Galaxy S22 (Feb 20) → Dell XPS 13 (arrived Feb 25).

Response :
- **SMDI (Ours):** Samsung Galaxy S22. ✓
- **MemGAS-style:** You got the **Dell XPS 13** first. ✗

Figure 19: Case study of multi-session numerical aggregation and temporal ordering.

exploit different interpretations of the same historical information while ultimately returning source memories for generation.

Figures 21-23 further demonstrate the role of multi-form retrieval and reranking under different evidence scenarios. Figure 21 shows that different memory forms may assign different ranks to the same source memory, while individual ranking signals can favor different candidates. By combining these signals, SMDI preserves multiple answer-bearing memories in the final retrieval set, allowing

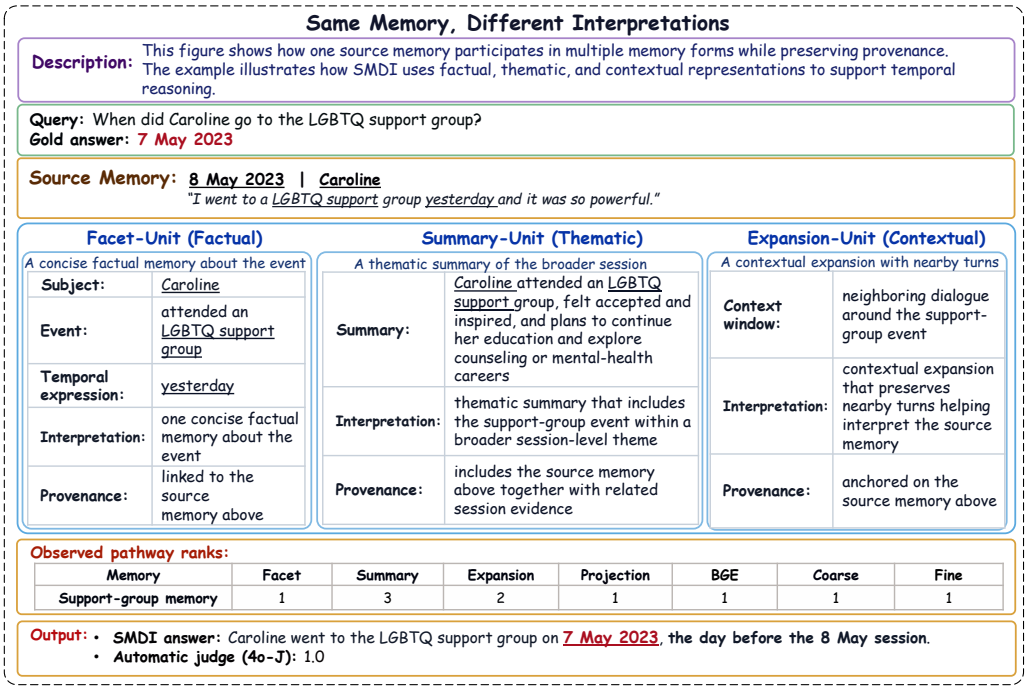


Figure 20: Case study of multi-form representations of a source memory.

the generator to jointly reason over the required evidence. Figure 22 further shows how adaptive fine reranking corrects candidate ordering after coarse fusion. Although the answer-bearing memory D15:1 is ranked eighth after coarse reranking, the fine stage promotes it to rank 2 by adaptively combining the projection, cross-encoder, and anchor ranking signals, increasing the number of gold memories in the final Top-3 from one to two. Figure 23 presents a multi-session query that requires combining complementary evidence from multiple source memories while distinguishing related but non-contributing information. Together, these cases illustrate how SMDI integrates complementary memory representations and ranking signals to preserve and refine answer-bearing memories for downstream generation.

Figure 24 considers an evolving-state query where both earlier and later observations remain in memory. SMDI retrieves both states while preserving their temporal order, allowing the generator to use the more recent observation when answering a current-state question. This example highlights the importance of retaining temporal information rather than treating historical observations as independent facts.

Finally, Figure 25 separates retrieval and generation errors through representative examples. Correct retrieval does not always guarantee a correct answer, as the generator may omit part of the retrieved evidence. Conversely, an exact annotated gold memory may be absent from the top-ranked results while semantically related evidence is still sufficient for answering the query correctly. When neither the gold memory nor sufficient alternative evidence is retrieved, generation also fails. These cases indicate that the final performance depends on both evidence coverage during retrieval and effective evidence utilization during generation.

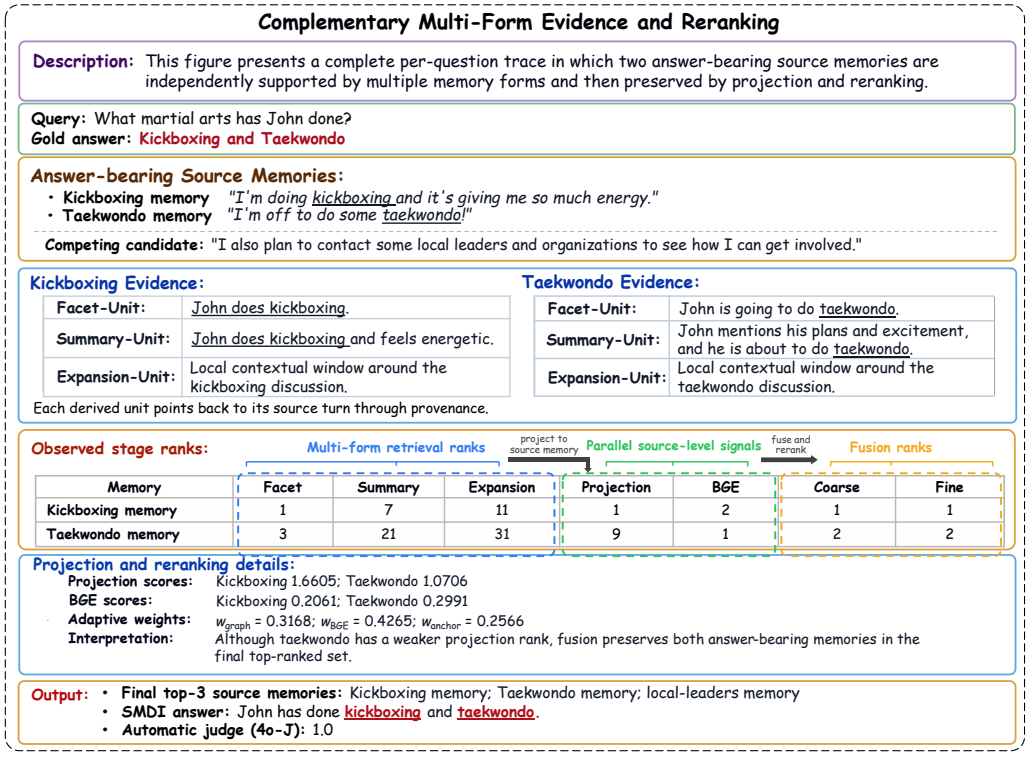


Figure 21: Case study of complementary multi-form evidence and reranking.

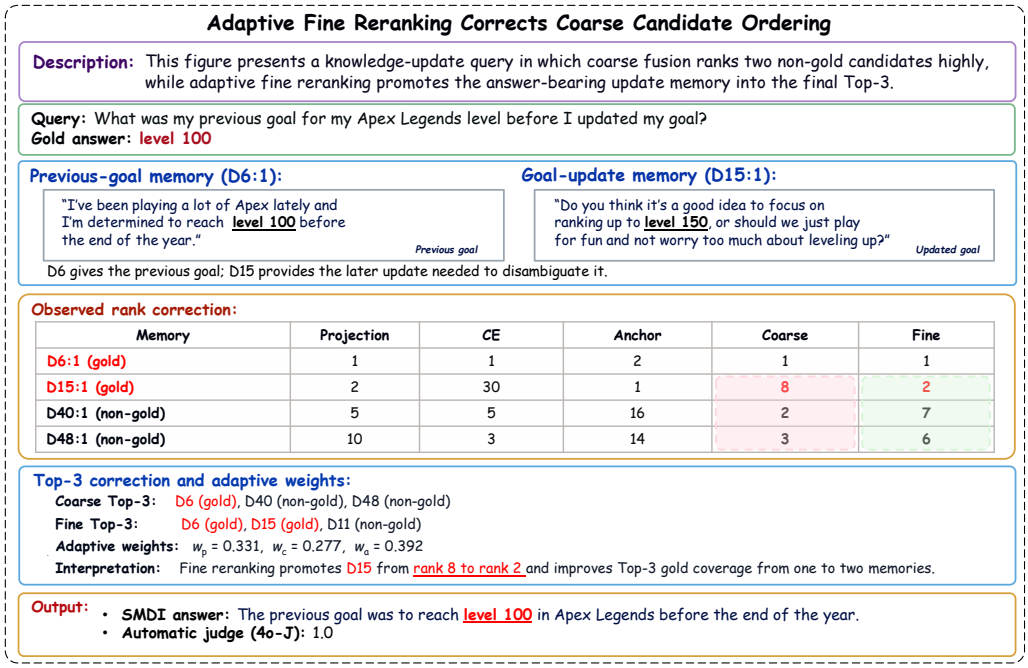


Figure 22: Case study of adaptive fine reranking for correcting coarse candidate ordering.

Multi-Session Complementary Evidence

Description: This figure illustrates a multi-session query that requires complementary evidence from multiple source memories, together with a negative qualification that should not be added into the final sum.

Query: How many days did I spend on camping trips in the United States this year?
Gold answer: 8 days

Supporting Source Memories:
- **Big Sur camping trip** *"I just got back from a 3-day solo camping trip to Big Sur in early April 2023."*
- **Yellowstone camping trip** *"I just got back from an amazing 5-day camping trip to Yellowstone National Park last month."*
- **Utah road trip** *"We had a 7-day family road trip in Utah in February. We did a lot of driving and hiking, but not camping."*

Supporting Evidence:

Big Sur camping trip	3-day solo camping trip in early April 2023.
Yellowstone camping trip	5-day camping trip last month.

Negative Qualification:

Utah road trip	7-day family road trip in Utah in February; lots of driving and hiking, but not camping.

Evidence composition:
Supporting evidence: Big Sur contributes 3 days; Yellowstone contributes 5 days.
Negative qualification: The Utah trip is related travel evidence, but explicitly states "not camping" and should not be added to the camping-duration sum.
Computation: 3 + 5 = 8 days

Observed final retrieval:
Final top-3 source memories: Big Sur camping trip; Yellowstone camping trip; Utah road trip
Final source-memory scores: 0.0871, 0.0839, 0.0801

Output:
- **SMDI answer:** 8 days.
- **Automatic judge (4o-J):** 1.0

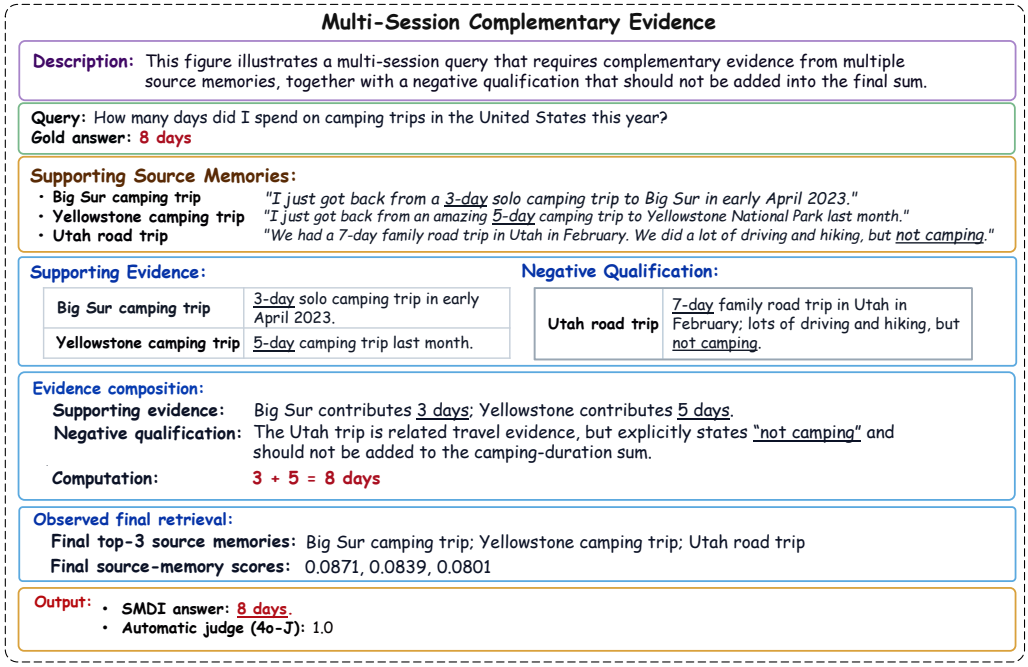


Figure 23: Case study of multi-session complementary evidence.

Current-State Retrieval under Evolving Memory

Description: This figure illustrates a current-state query in which two temporally ordered source memories coexist. The system retrieves both historical observations and uses the later observation to answer the current-state question.

Query: How many projects have I completed since starting painting classes?
Gold answer: 5

Earlier Source Memory:

"The user has completed 4 projects since starting painting classes."

Earlier state

Later Source Memory:

"I just finished my 5th project since starting painting classes, and I'm feeling pretty accomplished!"

Later observation

Time 1 (earlier) — *Time progresses* — Time 2 (later)

Evolving historical state:
- The earlier memory records a state of four completed projects.
- The later memory records the observation after completion of the fifth project.
- The available artifacts do not encode an explicit replacement or supersession operation.
- This example illustrates temporally ordered evidence for a current-state query.

Observed final retrieval:
Final top-3 source memories: earlier project count, later fifth-project observation, another related memory. The two relevant observations are both available to the generator. Their session dates preserve the earlier/later ordering.

Output:
- **SMDI answer:** 5 projects.
- **Automatic judge (40-J):** 1.0
- Interpretation: The main-run answer uses the later observation for the current-state query.

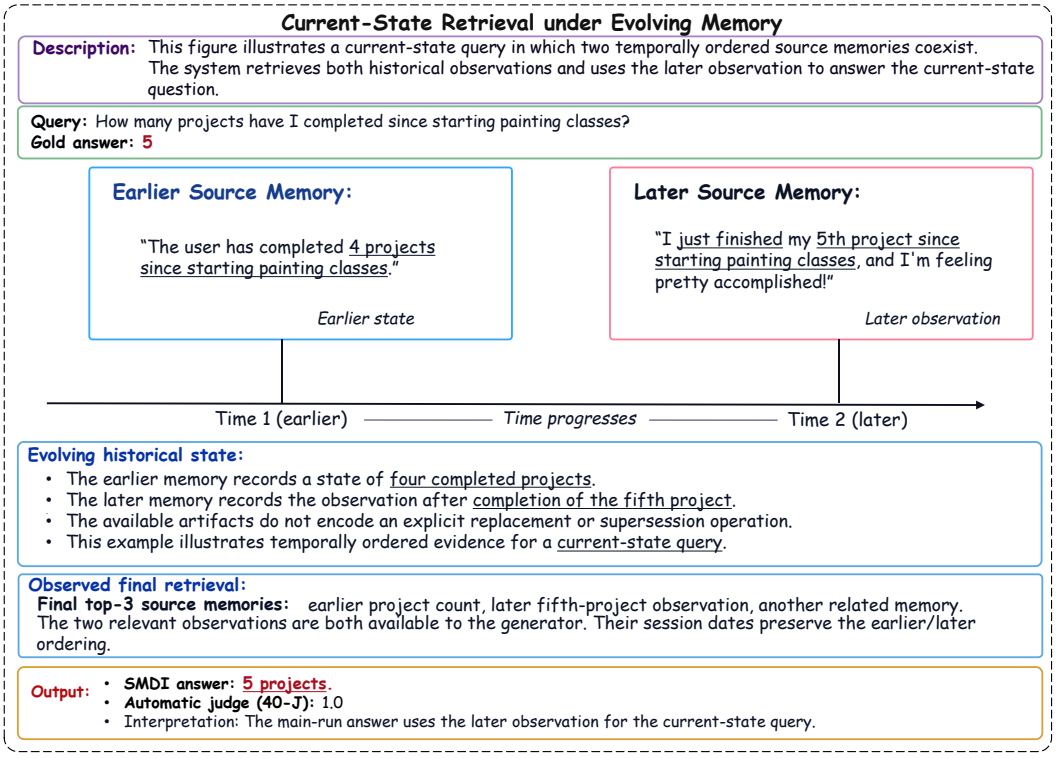


Figure 24: Case study of current-state retrieval under evolving memory.

Retrieval–Generation Error Analysis

Description: For this diagnostic analysis, retrieval is counted as a gold-anchor hit when at least one exact annotated gold-turn anchor appears among the top-3 retrieved source memories. Generation is counted as correct when the automatic judge returns yes.

1. Gold-anchor hit + correct generation

Query:	When did Caroline go to the LGBTQ support group?
Gold:	7 May 2023
Top-3 source memories:	support-group turn; related follow-up; related follow-up
Answer:	Caroline went to the LGBTQ support group on 7 May 2023, the day before the 8 May session.
Interpretation:	The relevant source memory is retrieved and the relative temporal expression is **correctly resolved.**

2. Gold-anchor hit + wrong generation

Query:	What fields would Caroline be likely to pursue in her education?
Gold:	Psychology, counseling certification
Top-3 source memories:	support-group turn; education/counseling turn; related follow-up
Answer:	Counseling or mental health fields.
Failure:	Relevant evidence is retrieved, but generation **omits psychology and the counseling-certification detail.**

3. Gold-anchor miss + correct generation

Query:	Would Caroline pursue writing as a career option?
Gold:	Likely no; she wants to be a counselor
Top-3 source memories:	education/counseling turn; semantically related turn; related turn
Answer:	Caroline is interested in counseling or mental health, not writing.
Interpretation:	Although the exact gold anchor is missed, a **semantically related** source memory provides sufficient evidence for a correct answer.

4. Gold-anchor miss + wrong generation

Query:	What is Caroline's relationship status?
Gold:	Single
Top-3 source memories:	support-group turn; related follow-up; related turn
Answer:	The retrieved contexts do not mention Caroline's relationship status.
Failure:	Retrieval does not recover **relationship-related evidence**, leading the generator to abstain.

Figure 25: Case study of retrieval and generation errors.