FROM MEMORY ACCESS TO EVIDENCE SATISFACTION: DIAGNOSING AND RESOLVING DEMAND–EVIDENCE MISALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-term memory systems are typically optimized to preserve and retrieve records relevant to a later query. Yet relevance is an item-level property, whereas answering is a set-level requirement: the retrieved evidence must jointly establish the facts, relations, temporal conditions, and state constraints required by the question. We characterize this challenge as demand–evidence misalignment and introduce MemPath, an interventional diagnostic framework that traces the transformation from historical information to answer-supporting evidence across the memory lifecycle. Across representative systems, MemPath reveals that improving individual memory stages does not necessarily ensure sufficient evidence for answering, as failures can arise from insufficient access, incomplete demand coverage, or ineffective evidence utilization. Guided by this diagnosis, we propose BiDE-Mem, a demand-guided evidence construction framework that represents questions as typed demands, retrieves and organizes supporting evidence, and aligns the constructed evidence with these demands before answer generation. BiDE-Mem improves accuracy by 6.23 points on LoCoMo and achieves 91.80% overall accuracy on LongMemEval, with particularly strong results on multi-hop and multi-session reasoning. Ablations and evidence-level analyses attribute these gains to improved demand coverage and evidence organization rather than evidence quantity alone. Our results highlight that effective long-term memory requires not only accessing relevant information, but also constructing evidence that satisfies the requirements of the current question. We release our code at `https://anonymous.4open.science/r/BIDE-8507`.

1 INTRODUCTION

Long-term memory systems are commonly evaluated by whether they can retrieve information from past interactions and use it to answer a later question. This view, however, obscures a critical distinction. Retrieval relevance is an item-level property: a memory may mention the same person, event, or topic as the question. Answering is a set-level requirement: the available evidence must jointly establish the facts, relations, temporal conditions, and state constraints required by the question. Consequently, retrieving relevant memories does not imply that the resulting context is sufficient for answering.

Consider the question in Figure 1: "What activities helped Melanie relax and when?" Memories mentioning Melanie's running, pottery, or stressful work may all be topically relevant. Yet a complete answer requires more than retrieving activity-related records. The evidence must identify the activities that produced the requested effect, connect each activity to Melanie, and establish the corresponding temporal information. We refer to these answer conditions as *question demands*, denoted by $D(Q)$. Given a retrieved evidence set E, high relevance between individual records and Q does not guarantee that $E \models D(Q)$. We call this discrepancy *demand–evidence misalignment*.

Demand–evidence misalignment cannot be fully studied using only end-to-end accuracy. Memory systems can fail for multiple reasons: information may be lost during memory construction, stored evidence may become inaccessible, retrieval may omit required support, budgeted selection may discard critical dependencies, or the model may fail to effectively use otherwise sufficient evidence.

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.

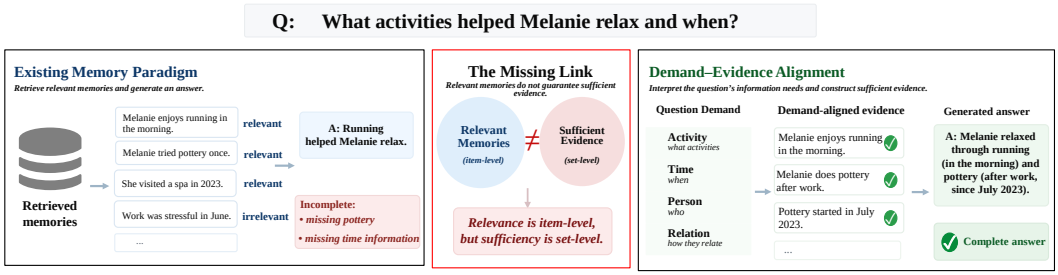


Figure 1: Relevant memories may not provide sufficient evidence.

Existing long-term memory benchmarks such as LoCoMo (Maharana et al., 2024) and LongMemEval (Wu et al., 2025) evaluate memory under multi-session interaction, temporal reasoning, knowledge updates, and long-range dependencies, but primarily measure final answer quality. Although LongMemEval examines design choices in indexing, retrieval, and reading, the root causes and propagation of individual failures across the memory pathway remain underexplored.

Recent methods have begun to address related limitations. EviMem (Li et al., 2026b) diagnoses evidence gaps after retrieval and uses them to guide iterative search, while GraphMemix (Li et al., 2026a) constructs query-aware evidence forests under a context budget. These approaches demonstrate that effective memory utilization requires more than semantic similarity. However, they do not jointly provide an interventional account of where demand–evidence misalignment emerges along the memory pathway, nor do they use an explicit, typed demand representation as a shared contract for memory access, set-level evidence construction, and pre-generation support verification.

We therefore introduce **MemPath**, an interventional diagnostic framework that traces how historical information is transformed into answer-supporting evidence. Using question demands as a shared reference, MemPath evaluates source accessibility, demand coverage, and answer recovery under oracle evidence interventions. Applying MemPath to representative memory systems reveals that accessibility, sufficiency, and utilization are distinct bottlenecks: one strong system retains over 93% of annotated sources but achieves only 85.3% answer accuracy, while verified evidence recovers 54.8% of incorrect answers, with substantially different effects across systems. These findings show that retrieving relevant memories or adding more context does not guarantee success; the available information must be organized into evidence that satisfies the question's demands.

Motivated by this diagnosis, we propose **BiDE-Mem**, a long-term memory framework with bidirectional demand–evidence alignment. BiDE-Mem represents questions as factual, relational, temporal, state, and constraint demands to guide evidence construction, and aligns the resulting evidence back to these demands before generation. BiDE-Mem improves overall accuracy by 6.23 points on LoCoMo and achieves 91.80% overall accuracy on LongMemEval, with particularly strong results on multi-hop, multi-session, temporal, and knowledge-update reasoning. Ablations and evidence-level analyses on LoCoMo attribute these gains to improved demand coverage and evidence organization rather than simply retrieving more context.

Our contributions are:

- We identify and formulate demand–evidence misalignment in long-term agent memory, separating memory relevance from evidence sufficiency.
- We introduce MemPath, an interventional diagnostic framework that traces the transformation from historical interactions to answer-supporting evidence and localizes failures through source accessibility, demand coverage, and oracle evidence interventions.
- We propose BiDE-Mem, a long-term memory framework based on bidirectional demand–evidence alignment, which uses question demands to guide memory access and evidence construction while verifying whether constructed evidence satisfies the required demands.
- We evaluate BiDE-Mem on LoCoMo and LongMemEval, demonstrating strong performance across both benchmarks, with detailed ablations and evidence-level analyses on LoCoMo.

2 MEMPATH: DIAGNOSING DEMAND–EVIDENCE MISALIGNMENT

Existing long-term memory systems optimize different stages of memory utilization, including memory construction, organization, retrieval, and context composition (Park et al., 2023; Packer et al., 2023; Zhong et al., 2024; Chhikara et al., 2025; Yang et al., 2026; Wang & Chen, 2025). However, final answer accuracy does not reveal whether failures originate from missing information, incomplete evidence construction, or ineffective evidence utilization. We introduce **MemPath**, an interventional diagnostic framework that traces how historical information is transformed into answer-supporting evidence and identifies where demand–evidence misalignment emerges.

2.1 CANONICAL MEMORY PATHWAY

MemPath represents long-term memory utilization as a sequence of transformations:

$$H \rightarrow M \rightarrow O \rightarrow C_Q \rightarrow E_Q \rightarrow A_Q, \tag{1}$$

where H denotes historical interactions, M denotes constructed memory, O denotes memory organization, C_Q denotes the retrieved candidate pool, E_Q denotes the assembled evidence context, and A_Q denotes the final answer. This pathway serves as a canonical interface for heterogeneous memory systems rather than assuming identical internal structures. For each system, observable artifacts are mapped to corresponding stages, while unavailable stages are treated as unobserved.

To evaluate whether information available at each stage can support answering, MemPath introduces Question Demands as a shared reference. For a question Q, we define:

$$D(Q) = \{d_1, \ldots, d_m\}, \tag{2}$$

where each demand d_i represents an answer requirement involving required facts, relations, temporal conditions, states, or constraints. Unlike retrieval relevance, which evaluates individual records, Question Demands characterize whether the collected evidence jointly satisfies the conditions required for answering. Conceptually, a demand is satisfied when the available evidence supports the corresponding answer condition.

Based on this representation, MemPath separates three aspects that are conflated by end-to-end accuracy: whether required information is accessible, whether assembled evidence satisfies question demands, and whether improving evidence can recover incorrect answers.

2.2 DEMAND–EVIDENCE DIAGNOSIS

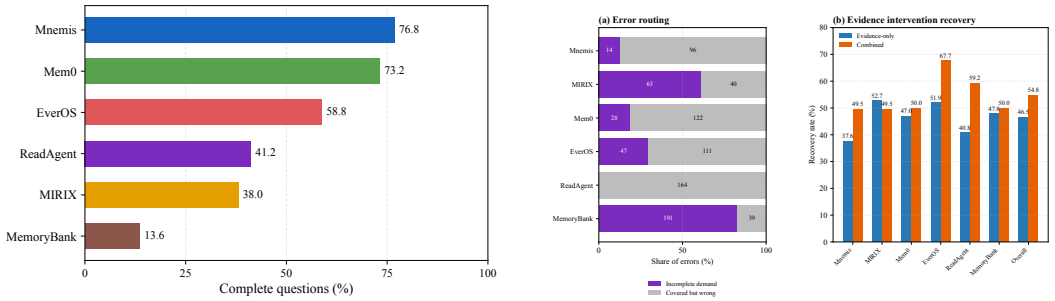
MemPath analyzes demand–evidence misalignment from three complementary perspectives.

Lifecycle Accessibility. We first examine whether required information successfully propagates through the memory pathway. We measure accessibility at different stages by tracking whether supporting information reaches the retrieved candidates and final evidence context. Table 1 summarizes the memory organization, retrieval behavior, evidence construction, and final performance of representative systems. Detailed metric definitions, source annotation procedures, and implementation details are provided in Appendix B.

Table 1: Lifecycle-level analysis of long-term memory systems.

System	Memory Organization	Retrieval Budget	Retrieval Coverage	Evidence Budget	Prompt Coverage	Accuracy
Mnemis	Graph/event memory	62.33 units	93.68%	62.33 units	95.85%	85.26
MIRIX	Episodic memory + tools	39.95 items	80.80%	Tool evidence	78.77%	83.51
Mem0	Semantic memory	60.00 items	90.37%	60.00 items	90.17%	80.39
EverOS	Episode memory	10 episodes	75.96%	Episode context	75.96%	80.19
ReadAgent	Gist + page hierarchy	2.44 pages	73.63%	48.96 units	99.87%	74.48
MemoryBank	Retention entries	Top-5 entries	25.85%	5 entries	25.93%	29.61

The results reveal diverse accessibility patterns across systems. MemoryBank suffers primarily from insufficient information access, with retrieval and prompt coverage of only 25.85% and 25.93%, re-



(a) Question-level demand completeness.

(b) Evidence intervention.

Figure 2: Demand–evidence diagnosis with MemPath. (a) Question-level demand completeness. (b) Error routing and answer recovery under evidence intervention.

spectively, which limits the amount of usable evidence available for answering. In contrast, ReadAgent substantially increases information availability through hierarchical expansion and achieves 99.87% prompt coverage, but its answer accuracy remains at 74.48%, showing that evidence availability alone does not guarantee successful answering. Meanwhile, Mnemis retrieves over 93% of annotated supporting information and achieves high prompt coverage (95.85%), yet still obtains only 85.26% accuracy, indicating that accessible information may still require effective evidence organization and utilization before answer generation.

Demand Completeness. Lifecycle accessibility indicates whether supporting information reaches later stages of the memory pathway, but does not determine whether the assembled evidence jointly satisfies all requirements of a question. For a question Q, we define demand-level coverage as

$$\mathrm{Cov}(E_Q, D(Q)) = \frac{1}{|D(Q)|} \sum_{d_i \in D(Q)} \mathbb{I}[E_Q \models d_i]. \tag{3}$$

A question is considered *demand-complete* when $\mathrm{Cov}(E_Q, D(Q)) = 1$, i.e., when all of its annotated demands are satisfied.

Figure 2a reports the percentage of demand-complete questions on the reviewed diagnostic subset of 250 questions containing 425 Question Demands. The results show that source accessibility and demand completeness are not equivalent. Systems may expose substantial supporting information while still failing to assemble evidence that jointly satisfies all answer requirements. ReadAgent provides a clear example: despite near-complete prompt-level source accessibility, only a substantially smaller fraction of questions are demand-complete, showing that context expansion alone does not guarantee sufficient evidence for answering.

Evidence Intervention. MemPath evaluates whether evidence insufficiency contributes to answer failures. For incorrect answers, we first distinguish between demand-incomplete cases and demand-complete but wrong cases, and then replace the original evidence context with a demand-complete oracle evidence package while keeping the reader and answer generation process unchanged.

This intervention isolates the contribution of evidence quality from reader capability. Figure 2b jointly shows error routing by demand completeness and answer recovery under evidence intervention. Completing missing or improperly organized evidence can recover a substantial fraction of incorrect answers, while recovery varies significantly across systems. These differences indicate that memory systems fail at different points along the evidence construction pathway.

2.3 Diagnostic Findings

MemPath reveals that improving individual memory stages does not necessarily produce sufficient evidence for answering. Retrieval accessibility, context availability, demand completeness, and final accuracy capture different aspects of memory utilization, but none alone determines whether the assembled evidence is sufficient for answering.

Across systems, failures emerge from different stages of evidence construction: some systems fail because required information cannot be accessed, some because available information does not jointly satisfy question demands, and others because demand-complete evidence is not effectively used for generation. These findings show that source arrival does not imply demand satisfaction, and demand satisfaction does not imply answer success. They motivate a shift from stage-specific optimization toward demand-aware evidence construction, where question requirements guide memory access and evidence organization, and the evidence is checked against requirements before answer.

3 BiDE-Mem: Bidirectional Demand–Evidence Alignment

3.1 Overview

The diagnosis in Section 2 shows that retrieving relevant memories does not guarantee sufficient evidence for answering. Existing long-term memory systems mainly optimize memory construction, organization, and retrieval, while answer requirements are rarely represented explicitly. BiDE-Mem therefore formulates long-term memory utilization as a demand-guided evidence construction process. Additional implementation details and prompt contracts are provided in Appendix C and D.

Given a question Q, BiDE-Mem constructs an explicit demand representation to guide evidence acquisition and organization, and aligns the evidence back to these demands before answer generation:

$$Q \rightarrow D(Q) \rightarrow C_Q \rightarrow (E_Q, Z_Q) \rightarrow A_Q, \tag{4}$$

where $D(Q)$ denotes Question Demands, C_Q the retrieved candidate memories, E_Q the constructed evidence package, Z_Q the demand–evidence alignment states, and A_Q the final answer. Unlike relevance-based retrieval, BiDE-Mem uses Question Demands as constraints throughout evidence construction and verifies whether the assembled evidence satisfies the required conditions.

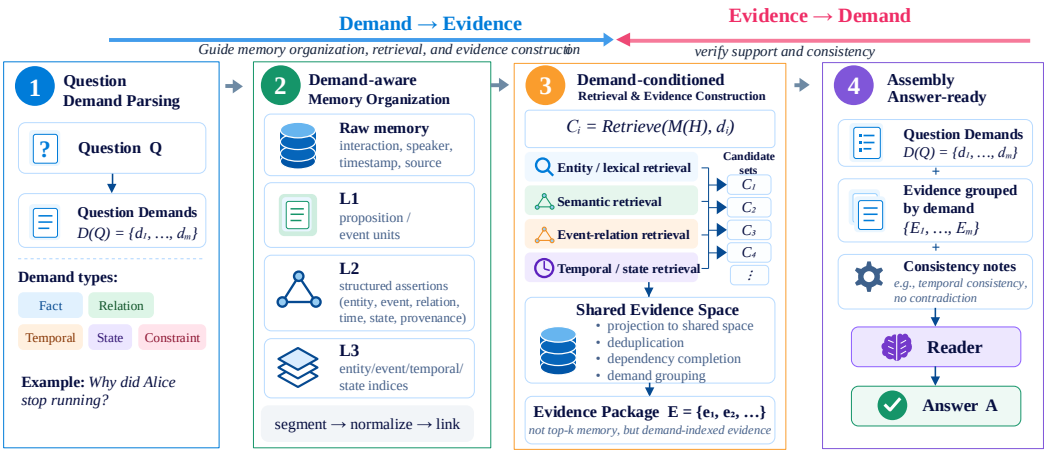


Figure 3: Overview of BiDE-Mem.

As illustrated in Figure 3, BiDE-Mem performs bidirectional demand–evidence alignment. In the demand-to-evidence direction, the predefined demand informs memory organization, while question-specific demands guide evidence acquisition and construction. In the evidence-to-demand direction, constructed evidence is aligned back to the demands to verify support and consistency before generation, distinguishing merely relevant memories from evidence sufficient for answering.

3.2 Question Demand Representation

Following MemPath, BiDE-Mem treats Question Demands as answer-oriented conditions rather than retrieval objectives. Instead of representing a question as a single retrieval query, it explicitly models the evidence requirements for producing a correct answer:

$$D(Q) = \{d_1, \ldots, d_m\}. \tag{5}$$

Each demand d_i specifies an answer requirement involving entities or events, relations, temporal conditions, states, or constraints. We categorize them into five operational types: *Fact*, *Relation*, *Temporal*, *State*, and *Constraint*, covering explicit facts, event dependencies, temporal ordering, state transitions, and answer restrictions.

Unlike query decomposition, which separates a question into retrieval intents, Question Demands represent answer conditions that may require joint evidence from multiple memories. A demand may depend on several records, while a single memory may support multiple demands. Question Demands therefore serve both evidence acquisition and evidence sufficiency assessment.

3.3 Demand-guided Evidence Acquisition

Given Question Demands, BiDE-Mem retrieves candidate memories for individual requirements rather than treating the question as a single retrieval signal. For each demand d_i:

$$C_i = \text{Retrieve}(\mathcal{M}, d_i), \qquad C_Q = \bigcup_{i=1}^{m} C_i, \tag{6}$$

where C_Q denotes the candidate pool for question Q. Retrieval combines lexical and semantic matching, entity and event associations, and temporal or state-aware access, with different demand types activating different evidence paths. Factual demands can retrieve explicit assertions, relational demands can follow entity and event connections, and temporal or state demands can access relevant updates across sessions.

Retrieved candidates retain provenance and structural associations so that complementary records can be combined when a demand requires multi-record support. Demand-guided acquisition therefore focuses on collecting candidates that jointly support the heterogeneous requirements of the question rather than only maximizing individual relevance.

3.4 Demand-guided Evidence Organization

Retrieved candidates do not directly constitute sufficient evidence. BiDE-Mem therefore organizes them into an answer-ready evidence package guided by Question Demands. The memory representation is demand-aware by design, preserving factual, relational, temporal, state, and provenance information required for later access.

Given the reranked Top-K candidates, BiDE-Mem reconstructs compact evidence records while preserving their original provenance, including content, speaker, timestamps, source identifiers, retrieval rank, and session/turn order. Redundant representations are removed while relationships required for reasoning are retained.

The resulting evidence preserves event continuity, temporal ordering, and state changes when needed for interpretation. Records describing the same event remain connected, while distinct events and temporal versions remain distinguishable. The evidence package is therefore a structured context for multi-record reasoning rather than an unordered collection of retrieved memories.

BiDE-Mem then aligns the constructed evidence back to each demand d_i. Each demand is assigned one of four alignment states: satisfied, partially supported, conflicting, or unmet. These states form Z_Q and capture whether the evidence supports the corresponding requirements while preserving entity, relation, temporal, and state consistency, allowing unsupported or inconsistent requirements to be identified before generation.

Finally, the Reader generates the answer from the question and the aligned evidence package:

$$A_Q = \text{Reader}\left(Q, \text{Serialize}(E_Q, Z_Q)\right), \tag{7}$$

where E_Q denotes the final evidence and Z_Q its alignment with the Question Demands. The Reader generates the answer from this aligned package rather than independently resolving missing evidence from raw retrieved memories.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We evaluate BiDE-Mem on LoCoMo and LongMemEval. On LoCoMo, we use Categories 1–4 with 1,540 answerable questions and compare against MemoryBank, ReadAgentLee et al. (2024), MIRIX, Mem0, EverMemOSHu et al. (2026), and MnemisTang et al. (2026). All LoCoMo systems use GPT-5.4-mini as the shared answer-generation model. On LongMemEval, we follow the official 500-question evaluation setting and compare with reported baseline results; BiDE-Mem uses Gemma 4 31B and Llama 3.1 8B Instruct for memory processing and GPT-5.4-mini for answer generation and evaluation. Final answers are evaluated using the corresponding benchmark protocols. Detailed dataset settings, model configurations, baseline protocols, prompts, and implementation details are provided in the Appendix A.

4.2 MAIN RESULTS

LoCoMo Results. Table 2 summarizes the performance comparison on LoCoMo.

Table 2: Performance comparison on LoCoMo answerable questions (Cat1–Cat4). Input tokens are averaged per question.

Method	Overall	Multi-hop	Temporal	Open-domain	Single-hop	Input Tokens
MemoryBank	29.61	26.24	28.04	28.13	31.51	708.76
ReadAgent	74.48	71.99	60.44	41.67	84.42	4117.46
Mem0	80.39	68.79	78.19	58.33	87.63	3037.61
MIRIX	83.51	75.18	83.18	55.21	89.66	3182.15
EverOS	80.19	75.89	78.50	64.58	84.07	2864.51
Mnemis	85.26	69.15	82.87	60.42	94.41	13370.13
BiDE-Mem Top10	90.13	84.75	88.79	66.67	95.12	1334.91
BiDE-Mem Top20	**91.49**	**87.59**	**90.34**	**67.71**	**95.96**	2152.04

BiDE-Mem Top20 achieves 91.49% overall accuracy on LoCoMo, outperforming the strongest overall baseline, Mnemis (85.26%), by 6.23 percentage points. The improvement is consistent across all four categories. The largest gain occurs on multi-hop reasoning, where BiDE-Mem reaches 87.59%, compared with the best baseline result of 75.89%. BiDE-Mem Top10 already achieves 90.13% overall accuracy with substantially fewer input tokens than several strong baselines, while Top20 further improves performance across all categories. These results indicate that increasing the evidence budget is beneficial when additional evidence is selected and organized according to question demands, rather than simply increasing the amount of retrieved context.

LongMemEval Results. Table 3 reports the results on LongMemEval. Baseline numbers correspond to the reported results of the respective systems under their original evaluation settings.

BiDE-Mem achieves 91.80% overall accuracy on LongMemEval, slightly exceeding the strongest reported baseline, Mnemis, at 91.60%. The improvements are mainly observed on multi-session reasoning, temporal reasoning, and knowledge-update tasks, where evidence must be integrated, temporally resolved, or updated across interactions. This pattern is consistent with our focus on demand-guided evidence construction, while the smaller overall margin reflects the already strong performance of Mnemis on several single-session categories.

4.3 ABLATION STUDY

We conduct ablation experiments on LoCoMo using the Top20 configuration in Table 4.

Table 3: Performance comparison on LongMemEval.

Method	SS-User	Multi-Session	SS-Pref.	Temporal	Update	SS-Assist.	Overall
MemoryBank	17.86	42.86	79.17	16.67	19.23	43.33	32.67
ReadAgent	12.50	67.86	8.33	33.33	15.38	53.33	27.72
MIRIX	72.85	30.07	53.33	25.56	52.56	63.63	43.49
Mem0	94.30	66.90	86.70	75.90	87.20	96.40	80.80
EverMemOS	100.00	78.50	96.70	71.20	87.20	78.60	82.00
Mnemis	98.60	86.50	**100.00**	86.50	93.60	**100.00**	91.60
BiDE-Mem	98.57	**87.97**	90.00	**87.22**	**94.87**	**100.00**	**91.80**

Table 4: Ablation study of BiDE-Mem.

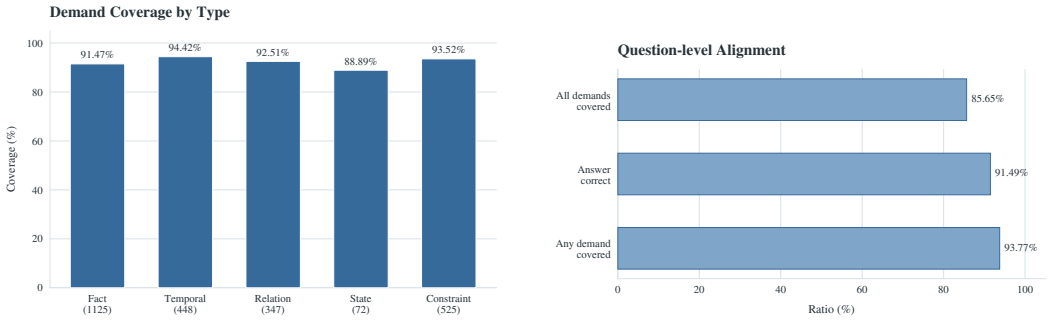
Variant	Any	Exact	Macro Recall	Accuracy	ΔAcc.
Full	95.05	86.39	91.34	91.49	–
w/o Explicit Demand	79.95	65.95	72.99	78.70	-12.79
w/o Structured Memory Organization	90.82	79.82	85.67	86.04	-5.45
w/o Designed Retrieval & Reranking	69.47	56.05	62.18	68.83	-22.66
w/o Post-retrieval Processing	95.05	86.39	91.34	84.55	-6.94

Removing designed retrieval and reranking causes the largest degradation, reducing accuracy from 91.49% to 68.83%, together with substantial drops in Any, Exact, and Macro Recall. This shows that demand-aware evidence construction still depends on reliable access to supporting memories. Removing explicit demand modeling lowers accuracy to 78.70%, confirming that Question Demands provide useful guidance beyond treating the question as a single retrieval query. Removing structured memory organization further reduces accuracy to 86.04%, indicating the benefit of preserving entity, event, temporal, and state structure.

Notably, removing post-retrieval processing leaves retrieval metrics unchanged but decreases accuracy from 91.49% to 84.55%. Thus, even when the same supporting memories remain accessible, answer quality degrades without organizing them into an answer-ready evidence package, directly supporting the distinction between memory access and evidence construction.

4.4 Demand–Evidence Alignment Analysis

To examine whether BiDE-Mem reduces demand–evidence misalignment, we analyze evidence coverage at both the demand and question levels under the Top20 budget. Unlike the 250-question, 425-demand subset used for the cross-system MemPath diagnosis, this analysis covers all 1,540 LoCoMo questions and 2,517 annotated Question Demands.



(a) Demand coverage by type.

(b) Question-level alignment.

Figure 4: Analysis of demand–evidence alignment in BiDE-Mem.

Figure 4a reports evidence coverage across different types. Across 2,517 demands, BiDE-Mem achieves 92.49% overall coverage, with consistently high coverage on factual, temporal, relational,

and constraint. State demands obtain relatively lower coverage at 88.89%, suggesting that evidence involving dynamic state changes remains more difficult to preserve and align over long interactions.

Figure 4b further summarizes question-level alignment. BiDE-Mem achieves 93.77% any-demand coverage and 85.65% complete demand coverage, alongside a final answer accuracy of 91.49%. These results show that high demand-level coverage does not collapse directly into answer accuracy. Some questions can still be answered correctly without satisfying every annotated demand under our coverage criterion, while some failures remain even when the required evidence is available.

Among the 131 incorrect answers, 95 involve incomplete demand coverage, indicating that incomplete demand support remains the dominant observed failure mode under our coverage criterion. The remaining errors occur despite complete demand coverage, showing that evidence availability alone does not fully determine answer correctness. These residual errors suggest that further improvements should focus not only on retrieving and constructing evidence, but also on reasoning over already sufficient evidence.

5 Related Work

Long-term memory for LLM agents. Long-term memory enables LLM agents to preserve and reuse information beyond limited context windows. Existing approaches improve memory construction, updating, organization, and retrieval through reflection, hierarchical structures, and structured memory representations (Park et al., 2023; Zhong et al., 2024; Chhikara et al., 2025; Yang et al., 2026; Wang & Chen, 2025). Long-context methods further explore compression and selective access over interaction histories. However, existing methods mainly optimize memory access and representation, while whether retrieved memories form sufficient evidence for answering remains less explored.

Retrieval and evidence reasoning. Retrieval-augmented generation and dense retrieval improve evidence access by retrieving relevant information from large collections (Lewis et al., 2020; Karpukhin et al., 2020). Multi-hop question answering demonstrates that complex questions require combining multiple evidence pieces and reasoning over relationships (Yang et al., 2018; Welbl et al., 2018). Recent studies investigate evidence quality beyond relevance, including evidence sufficiency and gap-driven retrieval (Li et al., 2026b), query-aware evidence validation (Yang et al., 2026), and complementary evidence construction (Li et al., 2026a). However, these studies focus on specific retrieval or evidence mechanisms rather than tracing demand–evidence alignment throughout the memory pipeline.

Evaluation of long-term memory systems. Recent benchmarks evaluate long-term memory under multi-session interaction, temporal reasoning, and multi-hop settings. LoCoMo evaluates conversational memory across diverse reasoning categories (Maharana et al., 2024), while LongMemEval measures long-term information extraction, updates, temporal reasoning, and multi-session reasoning (Wu et al., 2025). Analysis frameworks such as A-TMA separate memory retrieval and answer-level failures (Shi et al., 2026). Nevertheless, a general framework for diagnosing where demand–evidence misalignment emerges across heterogeneous memory systems remains limited.

6 Conclusion

This work studies long-term memory from the perspective of question demands and evidence sufficiency. We introduce MemPath to diagnose demand–evidence misalignment by tracing information accessibility, demand coverage, and evidence effectiveness across the memory pathway. Guided by this diagnosis, BiDE-Mem integrates demand representation, structured memory access, evidence construction, and demand–evidence alignment into a unified pipeline. Experiments on LoCoMo and LongMemEval demonstrate strong performance, particularly on multi-hop, multi-session, temporal, and knowledge-update tasks. Overall, our results show that accessing relevant memories does not necessarily yield sufficient evidence for answering, motivating a shift from retrieval-oriented memory optimization toward demand-aware evidence construction.

AI USE STATEMENT

In this work, we used generative AI models as components of the proposed method, including memory construction, demand-conditioned query representation, answer generation, and answer evaluation. We also used generative AI tools for assisting manuscript refinement and improving presentation clarity.

AI-assisted components were run under fixed configurations. The authors manually reviewed the resulting experimental artifacts and analyses, verified the reported results and claims, and take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the datasets, evaluation protocols, model configurations, baseline settings, and implementation details in the main paper and Appendix. Our experiments use the publicly available LoCoMo and LongMemEval benchmarks and follow their corresponding evaluation protocols. The Appendix further documents the MemPath diagnostic procedure, BiDE-Mem configuration, demand annotations, retrieval and evidence settings, and prompt contracts. Source code, evaluation scripts, and configurations are provided in the anonymous repository accompanying this submission.

REFERENCES

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory. In *ECAI 2025: 28th European Conference on Artificial Intelligence*, pp. 2993–3000, 2025. doi: 10.3233/FAIA251160.

Chuanrui Hu, Xingze Gao, Zuyi Zhou, Dannong Xu, Yi Bai, Xintong Li, Hui Zhang, Tong Li, Chong Zhang, Lidong Bing, et al. Evermemos: A self-organizing memory operating system for structured long-horizon reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 45836–45853, 2026.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 6769–6781, 2020.

Kuang-Huei Lee, Xinyun Chen, Hiroki Furuta, John Canny, and Ian Fischer. A human-inspired reading agent with gist memory of very long contexts. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 26396–26415, 2024.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Geng Li, Yuhao Wang, Dong Li, Jianye Hao, and Yuxin Peng. Graphmemix: Query-aware evidence forests for long-term multimodal agent memory. *arXiv preprint arXiv:2608.26983*, 2026a.

Yuyang Li, Yime He, Zeyu Zhang, and Dong Gong. Evimem: Evidence-gap-driven iterative retrieval for long-term conversational memory. *arXiv preprint arXiv:2604.27695*, 2026b.

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, 2024.

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

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Zitong Shi, Yixuan Tang, and Anthony Kum Hoe Tung. A-tma: Decoupling state-aware memory failures in long-term agent memory. *arXiv preprint arXiv:2607.01935*, 2026.

Zihao Tang, Xin Yu, Ziyu Xiao, Zengxuan Wen, Zelin Li, Jiaxi Zhou, Hualei Wang, Haohua Wang, Haizhen Huang, Weiwei Deng, et al. Mnemis: Dual-route retrieval on hierarchical graphs for long-term llm memory. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23914–23928, 2026.

Yu Wang and Xi Chen. Mirix: Multi-agent memory system for llm-based agents. *arXiv preprint arXiv:2507.07957*, 2025.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302, 2018.

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.

Wei Yang, Bryce Kan, Shixuan Li, Li Li, Yuehan Qin, Jiate Li, Paul Bogdan, and Jesse Thomason. Ramem: Contextual reinstatement for long-term agentic memory. *arXiv preprint arXiv:2606.22844*, 2026.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2369–2380, 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.

A Experimental Details

A.1 Dataset and Evaluation Setup

We evaluate BiDE-Mem on LoCoMo and LongMemEval. LoCoMo is used for both the main evaluation and the MemPath diagnostic analysis. We focus on Categories 1–4, resulting in 1,540 answerable questions. LongMemEval is used as an additional benchmark following its official 500-question evaluation setting and task definitions.

For LoCoMo, we use two demand annotation sets for different analyses. The **MemPath diagnostic subset** contains 250 questions and 425 manually reviewed Question Demands and is used for the cross-system diagnosis in Section 2, including demand completeness, error routing, and evidence intervention. The **full demand analysis** covers all 1,540 LoCoMo questions and contains 2,517 annotated Question Demands. It is used for the full-scale demand–evidence alignment analysis of BiDE-Mem in Section 4.4. The 425-demand subset supports detailed diagnosis across heterogeneous memory systems, while the 2,517-demand inventory supports complete analysis of BiDE-Mem over the full LoCoMo evaluation set.

A.2 LongMemEval Configuration

For LongMemEval, BiDE-Mem follows a staged processing pipeline for memory construction, structured extraction, answer generation, and evaluation. The main model configuration is summarized in Table 6.

Table 5: Dataset statistics and usage in our experiments.

Setting	Usage	Scale	Annotation
LoCoMo	Main evaluation	1,540 questions	Gold source evidence
MemPath diagnostic subset	Six-system diagnosis	250 questions	425 reviewed demands
Full demand analysis	BiDE-Mem analysis	1,540 questions	2,517 demands
LongMemEval	Main evaluation	500 questions	Official evaluation protocol

Table 6: Model configuration used for BiDE-Mem on LongMemEval.

Stage	Model	Function
Preprocessing	Gemma 4 31B	Content parsing and chunk processing
Extraction and analysis	Llama 3.1 8B Instruct	L1 mentions, L2 facts, entity relations, and source linking
Answer generation	GPT-5.4-mini	Final answer generation
Judge	GPT-5.4-mini	Evaluation with the official judging prompt

Gemma 4 31B is used to parse the interaction history and construct the initial memory representation. Llama 3.1 8B Instruct performs structured extraction and analysis, including L1 mentions, L2 factual units, entity–relation information, and source linking. GPT-5.4-mini generates the final answer from the constructed evidence package and is also used with the official LongMemEval judging prompt for evaluation.

For baseline comparison on LongMemEval, we use the reported results of the corresponding systems under their original evaluation settings rather than rerunning all baselines with a shared answer-generation model.

A.3 Baseline Evaluation

For the controlled LoCoMo comparison, we evaluate BiDE-Mem against six representative long-term memory systems: Mnemis, MIRIX, Mem0, EverOS, ReadAgent, and MemoryBank.

The evaluated systems adopt different memory representations, retrieval interfaces, and answer-generation procedures. We therefore preserve each system's native memory representation, retrieval procedure, evidence interface, and answer prompt rather than replacing them with a unified memory pipeline. All systems use GPT-5.4-mini as the shared answer-generation model for the LoCoMo evaluation. Final answers are evaluated using the same LoCoMo Judge protocol.

For MemPath diagnosis, system-specific adapters record observable intermediate artifacts and map them to the common diagnostic interface without modifying the underlying memory representation, retrieval behavior, evidence interface, or answer prompt.

Table 7: Baseline systems and evaluated artifacts on LoCoMo.

Method	Evaluated Artifact	Retrieval Interface
Mnemis	Reconstruction v2	Memory retrieval
MIRIX	Tool-aware evaluation trace	Tool-based retrieval
Mem0	Local reconstruction	Memory retrieval
EverOS	Frozen evaluation run	Hybrid retrieval
ReadAgent	Page hierarchy reconstruction	Page expansion
MemoryBank	Retention-based memory evaluation	Entry retrieval

A.4 BiDE-Mem Configuration on LoCoMo

For the LoCoMo experiments, BiDE-Mem uses GPT-5.4-mini for LLM-based memory construction, demand-conditioned query representation, and final answer generation. Retrieval and evidence con-

struction combine multiple retrieval channels, reranking, and budget-controlled evidence selection. Final answers are evaluated using the same Judge protocol as the baseline systems.

The main implementation configurations are summarized in Table 8.

Table 8: Implementation configuration of BiDE-Mem on LoCoMo.

Component	Configuration
LLM backbone	GPT-5.4-mini
Dense retriever	BGE-M3
Sparse retriever	SPLADE
Reranker	Qwen3-8B reranker
Candidate pool	128 candidates
Evidence budget	Top10 / Top20
Reader	GPT-5.4-mini
Judge	GPT-5.4-mini

All LLM-based components use temperature 0. Candidates are collected from multiple retrieval channels, merged, deduplicated, reranked, and selected under the specified evidence budget. The Top10 and Top20 configurations differ only in the final evidence budget unless otherwise stated.

Experiments are conducted on an NVIDIA A100 80GB PCIe GPU with CUDA 12.2 and NVIDIA driver 535.309.01.

B MemPath Diagnostic Framework Details

B.1 System Adaptation to the Diagnostic Interface

MemPath provides a unified diagnostic interface for analyzing how heterogeneous long-term memory systems transform question-relevant historical information into answer-supporting evidence. Following the formulation in the main text, the operational diagnostic interface is

$$Q \rightarrow D(Q) \rightarrow C_Q \rightarrow (E_Q, Z_Q) \rightarrow A_Q, \tag{8}$$

where $D(Q)$ denotes the Question Demands used as the common diagnostic reference, C_Q denotes the candidate information exposed by the system, E_Q denotes the final evidence context available for answering, Z_Q records the relationship between the observed evidence and the annotated demands, and A_Q denotes the system answer.

The internal memory construction and organization mechanisms preceding C_Q remain system-specific. Existing long-term memory systems expose different memory representations, retrieval mechanisms, and observable intermediate states. MemPath therefore does not require identical architectures. Instead, we construct system-specific adapters that map available artifacts into the diagnostic interface while preserving the original behavior of each system.

Table 9: System-specific artifacts mapped to the MemPath diagnostic interface.

System	Memory Artifact	Retrieval Artifact
Mnemis	Memory units and event structures	Retrieved memory records
MIRIX	Episodic memory and tool traces	Tool-based retrieval outputs
Mem0	Stored memory entries	Retrieved memory records
EverOS	Available memory traces	Retrieved memory context
ReadAgent	Page-level memory hierarchy	Expanded reading pages
MemoryBank	Retained memory entries	Retrieved memory items

The adapters only collect observable intermediate artifacts and associated metadata. They do not modify the memory construction, retrieval, Reader, answer prompt, or answer-generation procedure of the evaluated systems.

B.2 Question Demand Annotation

MemPath defines a Question Demand as an atomic answer requirement that the available evidence must satisfy. To support demand-level diagnosis, we construct a manually reviewed demand annotation set from 250 LoCoMo questions, containing 425 Question Demands.

Each question is decomposed into one or more atomic demands according to the information required by the question and reference answer. Each demand is further associated with supporting source evidence. The annotation procedure consists of:

1. identifying answer-bearing requirements from the question and reference answer;
2. decomposing these requirements into atomic Question Demands;
3. associating each demand with the source evidence required to support it.

Each demand is represented by a structured record containing its semantic requirement and evidence dependencies.

Table 10: Question Demand annotation schema.

Field	Description
Demand type	Fact, Relation, Temporal, State, or Constraint
Target entity/event	Entity or event involved in the requirement
Relation requirement	Required relationship or dependency
Temporal condition	Required time scope or ordering constraint
State condition	Required state or transition information
Supporting sources	Source IDs providing required evidence

The 425-demand annotation set is used only for the cross-system MemPath diagnosis. Separately, the full-scale BiDE-Mem analysis uses the 2,517-demand inventory described in Appendix A.1. These evaluation annotations are not provided to BiDE-Mem during memory construction, retrieval, evidence construction, or answer generation.

B.3 Diagnostic Evaluation Protocol

MemPath evaluates memory systems from three complementary perspectives: source accessibility, demand completeness, and evidence intervention.

Source-level accessibility. Source-level analysis examines whether annotated supporting information successfully propagates into the observable retrieval and answer contexts of each system. Given a set of annotated supporting sources, we measure whether these sources appear in the retrieved candidate set and whether they remain available in the final context used for answering. These measurements correspond to the retrieval and prompt coverage statistics reported in the main text.

Demand completeness. Demand-level analysis evaluates whether the observed evidence provides the source support required by each Question Demand. Let A_i denote the annotated supporting source set for demand d_i, and let S denote the source set observable at the evaluated stage.

The any-anchor criterion considers a demand accessible when at least one supporting source is available:

$$z_i^{\text{any}} = \mathbb{I}[A_i \cap S \neq \emptyset]. \tag{9}$$

The all-support criterion requires all annotated supporting sources for the demand to be available:

$$z_i^{\text{all}} = \mathbb{I}[A_i \subseteq S]. \tag{10}$$

The any-anchor criterion measures whether a demand has at least one accessible evidence anchor, whereas the all-support criterion provides the stricter operationalization used to assess complete demand support. For a question Q, the source-grounded demand coverage is

$$\mathrm{Cov}(S, D(Q)) = \frac{1}{|D(Q)|} \sum_{d_i \in D(Q)} z_i^{\mathrm{all}}. \tag{11}$$

A question is considered *demand-complete* when all of its annotated demands satisfy the all-support criterion, i.e., when $\mathrm{Cov}(S, D(Q)) = 1$. This question-level completeness measure is used in the MemPath analysis reported in Figure 2a.

Evidence intervention. For incorrect answers, MemPath first distinguishes demand-incomplete cases from demand-complete but wrong cases. We then intervene on the evidence while keeping the Reader and answer-generation procedure unchanged. In the **Evidence-only** setting, the original answer context is replaced with verified supporting evidence. In the **Combined** setting, missing verified evidence is added to the information already available to the system, and the resulting evidence is reorganized into a coherent answer-ready context before being passed to the Reader. All intervention outputs are evaluated using the same answer-level Judge protocol.

The corresponding recovery rate for an intervention j is

$$\mathrm{Recovery}(j) = \frac{1}{|\mathcal{F}_j|} \sum_{Q \in \mathcal{F}_j} \mathbb{I}[A_Q^{(j)} \text{ is correct}], \tag{12}$$

where $\mathcal{F}_j$ denotes the set of originally incorrect questions evaluated under intervention j and $A_Q^{(j)}$ denotes the answer produced after the intervention.

C BiDE-Mem Implementation Details

C.1 Memory Assertion Construction

BiDE-Mem constructs structured memories from raw interactions to support demand-aware retrieval and evidence construction. The memory organization is demand-aware by design rather than reconstructed for each question: the predefined Question Demand framework determines which factual, relational, temporal, state, and provenance information should be preserved so that future question-specific demands can access it reliably.

Rather than storing only compressed summaries, BiDE-Mem preserves fine-grained assertions together with their original provenance. Given an interaction history, the memory construction process transforms raw records into a hierarchical representation:

$$\text{Raw Interaction} \rightarrow \text{L1 Proposition} \rightarrow \text{L2 Assertion} \rightarrow \text{L3 Index}. \tag{13}$$

The construction process first identifies clause-level proposition units and then converts them into structured searchable assertions. Each assertion preserves semantic information required for later retrieval and reasoning, including entities, events, relations, values, semantic roles, polarity, temporal information, state information, and provenance.

The detailed construction contract and output format are provided in Appendix D.

The constructed memory maintains traceability from structured assertions back to the original interaction records. This design preserves the information structures anticipated by the demand framework while avoiding unsupported cross-session summaries and enabling retrieved evidence to be traced back to its original source.

Table 11: Hierarchical memory representation in BiDE-Mem.

Level	Representation	Function
Raw	Original conversation records	Preserve source information
L1	Clause-level proposition units	Capture atomic statements
L2	Structured assertions	Support retrieval and reasoning
L3	Entity/event/temporal/state indices	Enable structured access

C.2 QUERYFRAME IMPLEMENTATION OF QUESTION DEMANDS

Question Demand, denoted by $D(Q)$ in the main text, is the conceptual representation of the answer requirements associated with a question. In the implementation, these requirements are instantiated through a structured object referred to as **QueryFrame**. QueryFrame is therefore not a separate methodological representation; it is the implementation-level realization of the Question Demand formulation.

Given a question Q, the QueryFrame compiler converts the requirements implied by the question into structured fields used by downstream memory access and evidence construction. Unlike conventional query rewriting approaches that mainly extract semantic topics or retrieval keywords, QueryFrame represents answer-oriented conditions such as relation dependencies, temporal constraints, answer roles, and required evidence structures.

The compiler contract is described in Appendix D.2.

Table 12: QueryFrame implementation fields for Question Demands.

Field	Description
Subject	Target entity or event
Observer	Information holder or perspective entity
Evaluated subject	Entity being evaluated or described
Relation family	Required relation type
Relation direction	Direction of dependency
Answer role	Expected answer form
Temporal constraint	Required temporal condition
Location constraint	Required spatial condition
Evidence shape	Required evidence structure
List protection	Prevent omission of required items

The QueryFrame representation does not contain answers, retrieved memories, reference answers, gold demand annotations, or external knowledge. It only operationalizes the information requirements implied by the input question.

C.3 DEMAND-GUIDED RETRIEVAL AND EVIDENCE CONSTRUCTION

Given the QueryFrame implementation of $D(Q)$, BiDE-Mem retrieves candidate memories through multiple complementary retrieval channels. Different channels target lexical, semantic, entity/event, structural, temporal, and typed evidence signals required by different Question Demands.

For each demand d_i, the implementation follows the retrieval formulation introduced in the main text:

$$C_i = \mathrm{Retrieve}(\mathcal{M}, d_i), \qquad C_Q = \bigcup_{i=1}^{m} C_i. \tag{14}$$

Candidates returned by the different access paths are merged into a unified candidate pool, deduplicated, reranked, and selected according to the evidence budget.

Table 13: Retrieval channels used in BiDE-Mem.

Channel	Retrieval Target
Clause retrieval	Proposition-level evidence
Multivector retrieval	Semantic similarity
Occurrence retrieval	Entity/event mentions
SPLADE retrieval	Lexical matching
Structural retrieval	Relation-dependent evidence
Typed-L2 retrieval	Structured assertions

After reranking, the Top-K records are reconstructed into compact evidence units using their original information. The final evidence package preserves:

- source identifier;
- speaker information;
- timestamp;
- original text;
- optional image caption;
- retrieval rank;
- source session and turn order.

Redundant retrieval wrappers are removed while the relationships required for reasoning are preserved. Rather than directly concatenating retrieved memories, BiDE-Mem organizes selected records so that evidence associated with the same event remains connected, temporal order and state changes remain visible, and distinct events or temporal versions remain distinguishable. This produces the structured evidence package E_Q used in the subsequent evidence-to-demand alignment stage.

C.4 Evidence–Demand Alignment

After evidence construction, BiDE-Mem aligns the selected evidence back to the Question Demands before answer generation. This stage implements the evidence-to-demand direction of the bidirectional alignment described in Section 3.1.

Let $E_Q = \{e_1, \ldots, e_n\}$ denote the constructed evidence package. At the evidence-unit level, BiDE-Mem records whether individual evidence units support particular demands. These pairwise links are then aggregated into a demand-level alignment state:

$$S(d_i, E_Q) \in \{\text{satisfied}, \text{partial}, \text{conflicting}, \text{unmet}\}. \tag{15}$$

The collection of demand-level support and consistency states forms Z_Q. Table 14 summarizes the four alignment states.

Table 14: Demand–evidence alignment states.

State	Description
Satisfied	Evidence sufficiently supports the demand
Partial	Evidence supports only part of the requirement
Conflicting	Evidence contains inconsistent support
Unmet	Required supporting evidence is not identified

The alignment considers:

- entity identity consistency;

- relation dependency consistency;
- temporal constraint consistency;
- state consistency.

This stage does not introduce external knowledge or independently generate an answer. Its purpose is to determine how the constructed evidence supports the requirements of the question and to identify incomplete or inconsistent support before generation. The answer-ready package therefore contains both the organized evidence E_Q and its demand–evidence alignment information Z_Q, corresponding to the main pipeline

$$Q \rightarrow D(Q) \rightarrow C_Q \rightarrow (E_Q, Z_Q) \rightarrow A_Q. \tag{16}$$

The serialized (E_Q, Z_Q) package is finally provided to the Reader for one-pass answer generation.

D Prompt Contracts and Examples

BiDE-Mem uses two structured LLM-based transformations: memory assertion construction and the QueryFrame implementation of Question Demands. Instead of relying on unconstrained generation, both transformations follow fixed contracts that define their inputs, outputs, and generation restrictions. These contracts keep the resulting representations traceable and compatible with downstream retrieval and evidence construction.

D.1 Memory Writer V3 Contract

The Memory Writer transforms raw interaction records into structured memory assertions. Its objective is to preserve searchable information required by the demand-aware memory schema while maintaining provenance to the original interaction.

The input consists of an immutable original record together with deterministic clause spans. The output is a structured assertion object used to construct the hierarchical memory representation described in Appendix C.1.

Table 15: Memory Writer V3 contract.

Component	Description
Input	Original interaction and clause spans
Output	Structured memory assertion object
Atomic unit	Clause-level proposition
Required information	Entity, relation, value, role, polarity, time
Traceability	Source clause provenance
Restriction	No unsupported cross-record summaries

The Memory Writer follows several constraints:

- every clause is processed;
- coordinated propositions are split into independent assertions;
- negation, preference, causal, result, temporal, and state information are preserved;
- each assertion remains linked to its source clause;
- unsupported inference is prohibited.

D.2 QueryFrame V2 Compiler Contract

The QueryFrame compiler implements the Question Demand representation described in Section 3.2 and Appendix C.2. It transforms an input question into the structured fields used to instantiate $D(Q)$ during inference.

The compiler only encodes question requirements. It does not perform memory retrieval, evidence construction, or answer generation.

The compiler follows the following restrictions:

- it does not answer the input question;
- it does not access memory records or retrieval results;
- it does not use reference answers or gold demand annotations;
- it does not introduce external knowledge;
- it only represents information requirements implied by the question.

The generated QueryFrame is then used as the implementation-level interface between Question Demands and downstream demand-guided retrieval and evidence construction.

D.3 Example Transformations

This section illustrates the two structured transformations used in BiDE-Mem.

Memory assertion construction. Given an interaction:

> Alice stopped enjoying jazz after moving to Boston.

The Memory Writer produces a structured assertion:

```
{
  "subject": "Alice",
  "relation": "preference_change",
  "value": "stopped enjoying jazz",
  "condition": "after moving to Boston",
  "provenance": "source_clause_12"
}
```

The assertion preserves the state transition, temporal condition, and source provenance required for later retrieval.

QueryFrame construction. Given the question:

> When did Alice stop enjoying jazz?

The QueryFrame compiler produces:

```
{
  "relation_family": "state_change",
  "answer_role": "time",
  "temporal_constraint": "end_time",
  "evidence_shape": "transition"
}
```

The resulting QueryFrame is the implementation-level realization of the corresponding Question Demands. It guides retrieval toward evidence describing the relevant state transition and its temporal condition rather than only retrieving semantically similar memories.