TECHNICAL EVALUATION REPORT

HydraDB Performance Assessment

on the BEAM 1M Long-Term Memory Benchmark

Prepared by: HydraDB Research Team

Version: 1.0 | May 2026

1. Introduction

This report presents the results of benchmarking HydraDB on BEAM 1M, a purpose-built evaluation dataset for long-term AI memory at the one-million-token scale. BEAM 1M tests ten distinct memory capabilities including temporal reasoning, cross-session coherence, and contradiction resolution, representing the most comprehensive publicly available evaluation at this context length.

HydraDB achieved an overall average score of 82% across all ten dimensions, with standout results in Temporal Reasoning (91%), Event Ordering (92%), and Preference Following (96%). For context, these results are benchmarked against Hindsight, which has publicly claimed state-of-the-art performance on BEAM, making it the natural reference point for this evaluation.

The results validate HydraDB's core architectural approach: modeling knowledge as a versioned, time-aware graph produces measurable gains in the memory dimensions that matter most in production AI deployments.

2. About the BEAM Benchmark

2.1 Overview

BEAM (Benchmark for Evaluation of AI Memory) is a structured dataset designed to evaluate long-term memory capabilities in AI systems. It contains 100 conversations distributed across four context length tiers:

- 128K tokens
- 500K tokens
- 1M tokens
- 10M tokens

The conversations span multiple domains including general tasks, coding, and mathematics. Each conversation is constructed to include multi-turn reasoning chains, cross-session dependencies, follow-up questions, and information that must be updated, reconciled, or retrieved from temporally distant points in the context.

2.2 Memory Evaluation Dimensions

BEAM evaluates system performance across ten distinct memory dimensions, each targeting a specific failure mode observed in long-context AI systems:

- Abstention: Whether the system correctly acknowledges when information is unavailable.
- Contradiction Resolution: Whether the system resolves conflicting information appropriately.
- Event Ordering: Whether the system accurately tracks the sequence of events.
- Information Extraction: Whether relevant details are retrieved accurately from large contexts.
- Instruction Following: Whether the system adheres to user-specified instructions over time.
- Knowledge Update: Whether the system updates its knowledge when new facts supersede older ones.
- Multi-Session Reasoning: Whether the system reasons coherently across distinct sessions.
- Preference Following: Whether the system retains and applies user preferences.
- Summarization: Whether the system accurately compresses information from long contexts.
- Temporal Reasoning: Whether the system reasons correctly about time-dependent information.

3. HydraDB Architecture

HydraDB is a context engine built to address two fundamental limitations of existing memory systems: the absence of persistent cross-session memory, and the loss of structural and temporal meaning inherent in flat vector search. Rather than treating stored knowledge as isolated text fragments, HydraDB models it as a versioned, relational, time-aware graph.

The system is composed of three integrated architectural components:

Component	Description
Sliding Window Inference Pipeline	Text is segmented and enriched using surrounding context. A lightweight language model resolves references and extracts meaning, converting ambiguous statements into self-contained, independently retrievable facts.
Git-Style Versioned Knowledge Graph	Knowledge updates are appended as new graph edges rather than overwriting existing data. All historical states are preserved with timestamps, enabling accurate responses to time-sensitive queries.
Multi-Stage Retrieval Pipeline	At query time, the system expands queries into multiple forms, combines dense and sparse retrieval, traverses graph relationships, and applies multi-stage reranking to surface causally and temporally related information.

This architecture enables HydraDB to answer not only factual queries but also temporally scoped questions such as what was previously believed to be true at a given point in an interaction, a capability that is directly reflected in its benchmark results.

4. Evaluation Methodology

HydraDB was evaluated on the BEAM 1M benchmark. To enable a meaningful external comparison, we adopted the same evaluation configuration used in Hindsight's published results. This ensures the comparison is reproducible and not influenced by differences in evaluation setup. The configuration was as follows:

- Answer-generation prompts and LLM Judge prompts were taken directly from Hindsight's original published benchmark configuration, without modification.
- GPT 5.4 was used as evaluation judge across all ten memory dimensions.

Using a shared evaluation configuration means results across systems are directly comparable and any observed differences reflect genuine performance rather than methodological variation.

5. Benchmark Results

5.1 Side-by-Side Comparison

The table below presents scores for each memory dimension, along with the absolute difference between HydraDB and Hindsight. Positive values indicate dimensions where HydraDB outperforms Hindsight.

Memory Dimension	Hindsight (%)	HydraDB (%)	Difference
Abstention	90%	**89%**	**-1%**
Contradiction Resolution	59%	**66%**	**+7%**
Event Ordering	81%	**92%**	**+11%**
Information Extraction	61%	**80%**	**+19%**
Instruction Following	93%	**92%**	**-1%**
Knowledge Update	66%	**63%**	**-3%**
Multi-Session Reasoning	46%	**60%**	**+14%**
Preference Following	97%	**96%**	**-1%**
Summarization	84%	**88%**	**+4%**
Temporal Reasoning	60%	**91%**	**+31%**
Overall Average	**74%**	**82%**	**+8%**

6.2 Radar Chart: Multi-Dimensional Performance Profile

The radar chart below provides a visual profile of performance across all ten memory dimensions. The shaded area for HydraDB (dark blue) extends beyond Hindsight (mid blue) in the majority of dimensions, with the most pronounced outward expansion in Temporal Reasoning, Information Extraction, and Multi-Session Reasoning.

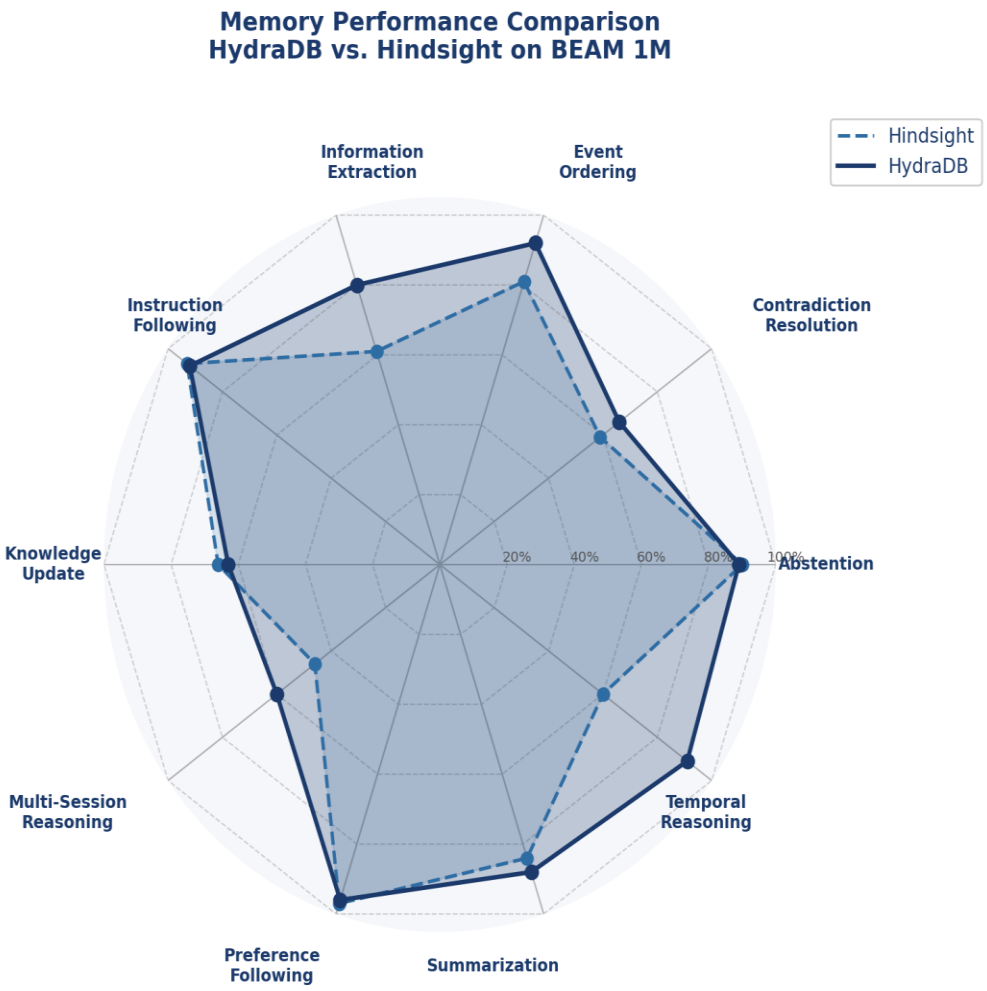

Figure 1. Radar chart comparing HydraDB and Hindsight across 10 BEAM 1M memory dimensions.

6.3 Analysis of Key Findings

HydraDB demonstrates its most substantial advantages in dimensions that require reasoning over time or across structurally distant information:

- Temporal Reasoning (+31 pp): HydraDB's versioned knowledge graph natively preserves the temporal context of every stored fact. This directly enables accurate responses to queries about when information was introduced or changed, a capability that flat retrieval systems cannot replicate.
- Information Extraction (+19 pp): The sliding window inference pipeline ensures that retrieved facts are self-contained and contextually enriched. This reduces retrieval failures caused by ambiguous or underspecified stored fragments.

- Multi-Session Reasoning (+14 pp): By preserving graph relationships across sessions rather than treating each session as an isolated retrieval pool, HydraDB maintains coherence over interactions that span long time horizons.
- Event Ordering (+11 pp): Timestamped graph edges allow the system to reconstruct event sequences with high fidelity, supporting accurate ordering queries even when events are distributed across distant parts of the context.

On five dimensions, the performance gap between HydraDB and Hindsight is within one to three percentage points. These dimensions (Abstention, Instruction Following, Knowledge Update, Preference Following) reflect capabilities that are less sensitive to temporal and structural reasoning, and where both systems perform at a broadly comparable level.

7. Conclusion

HydraDB achieves a state-of-the-art overall score of 82% on the BEAM 1M benchmark, outperforming the Hindsight baseline by 8 percentage points. The performance advantage is concentrated in memory dimensions that require temporal awareness, cross-session coherence, and retrieval of structurally related information, which are precisely the capabilities that HydraDB's architecture is designed to address.

These results validate the core architectural thesis of HydraDB: that treating stored knowledge as a versioned, time-aware graph yields measurable improvements over systems that rely on flat or stateless retrieval at scale. As AI deployments continue to operate over increasingly long interaction histories, the importance of robust long-term memory architecture will only grow.

References

[1] Mohammad Tavakoli et al. BEAM: Benchmark for Evaluation of AI Memory. 2025. arXiv:2510.27246

[2] Mohammad Tavakoli et al. BEAM Dataset Repository. GitHub, 2025. github.com/mohammadtavakoli78/BEAM

[3] HydraDB Team. HydraDB Benchmark Results. Technical Report, 2026. Google Drive

[4] HydraDB Team. HydraDB: A Context Engine for Long-Term AI Memory. Technical White Paper, 2026. benchmarks.hydradb.com/HydraDB.pdf

[5] Hindsight BEAM 1M Results. Agent Memory Benchmark, 2026. agentmemorybenchmark.ai

[6] Hindsight: Long-Term Memory for AI Systems. arXiv:2512.12818