A Traceable Hierarchical Memory System for Long-Term Research Meeting Support

NATIONAL YANG MING CHIAO TUNG UNIVERSITY

Students: Shang-Jung Tsai, Yu-Jen Chen, Yi-Hsin Lee

Advisor: Yu-Chun Yen

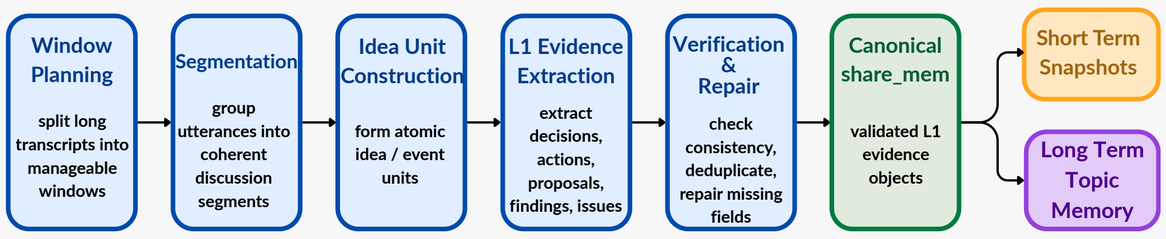
Motivation

- Research ideas evolve across meetings, but current AI systems often treat meetings as isolated text.
- Short chat history and flat RAG miss decision updates, outdated directions, and evidence trails.
- Full-transcript prompting preserves more context, but is costly, noisy, and hard to inspect or correct.
- How can long-term research meeting memory be made traceable, scalable, and correctable?

Our Approach

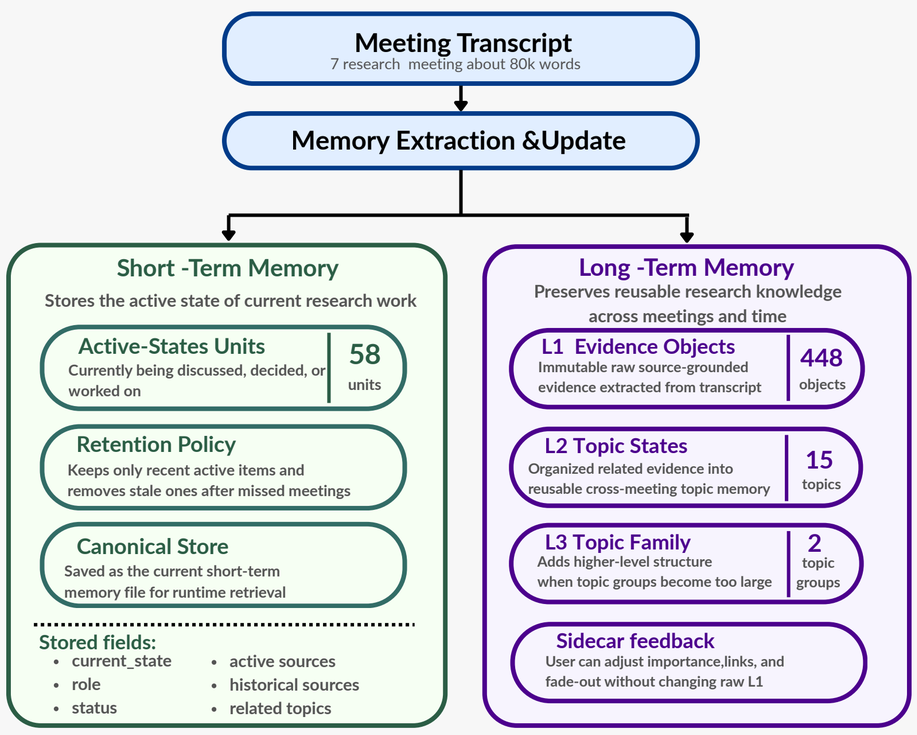
- We introduce a hierarchical memory system that turns research meeting transcripts into reusable and inspectable memory.
- Key idea:
  We augment research meeting memory to:
  - Extract share_mem evidence from raw transcripts
  - Build short/long-term memory for recent and evolving context
  - Retrieve structured context for grounded, inspectable answers

Memory Extraction



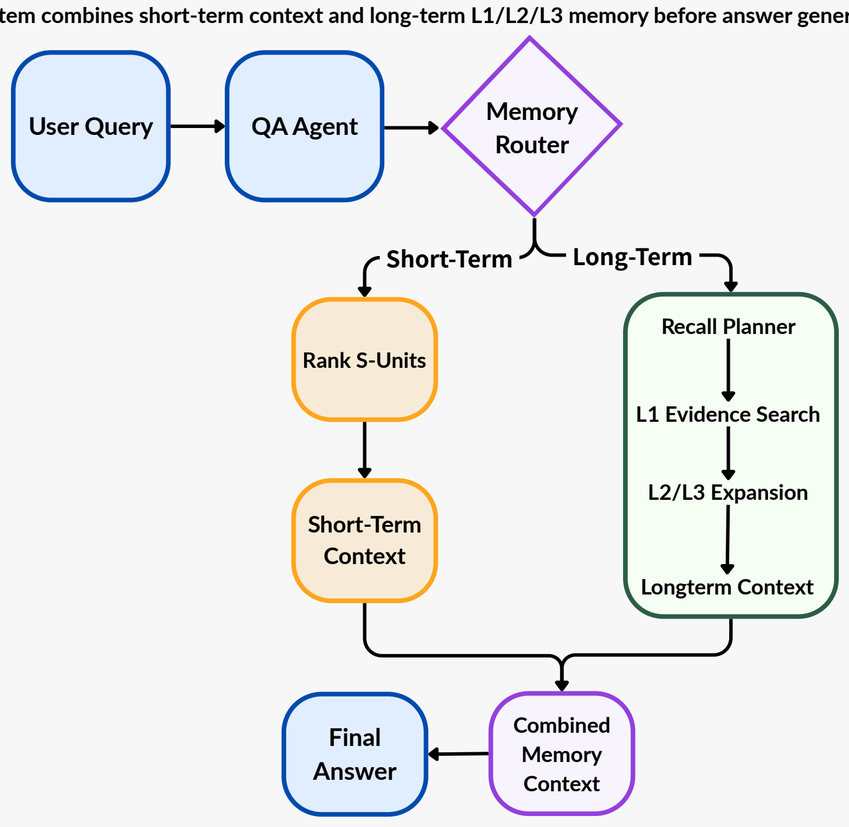
Memory Architecture

Short-term memory tracks active state; long-term memory preserves reusable topic structure.



Retrieval Flow

The system combines short-term context and long-term L1/L2/L3 memory before answer generation.



Key Hypotheses

- Traceable evidence improves answer grounding and reliability.
- Topic-state memory improves long-term reasoning.
- Inspectable retrieval increases human trust and controllability.

Evaluation Results

- We evaluate 14 questions across 5 research-memory question types using a 1–5 LLM-judge rubric.
- Final score = 0.7 × base quality + 0.3 × type-specific ability.
- Success = Final Quality ≥ 4.0, with Completeness and Type-Specific Score both ≥ 3.

Quality–Cost Comparison

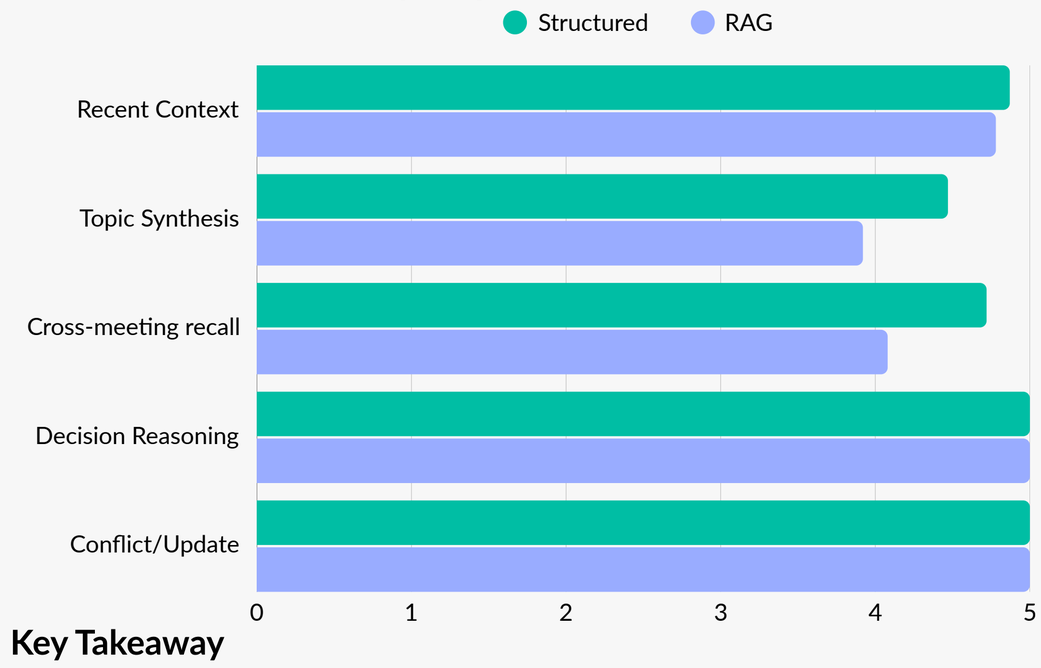
Method	Success Rate	Final Quality	Tokens
Structured memory	**92.9%**	**4.79**	**14.1K**
RAG Baseline	71.4%	4.51	21.3K
Full transcript	92.9%	4.89	81.1K

Note: Online cost excludes reusable one-time L1 construction.
One-time L1 construction: 3.08M tokens; break-even vs Full Transcript after ~46 questions.

Key Takeaway

- **Structured Memory achieves the same success rate as Full Transcript and higher quality than RAG.**
- **It approaches Full Transcript quality while using only ~17% of its online tokens.**

Where Structured Memory Helps



Key Takeaway

- **Structured Memory matches or exceeds RAG across all question types.**
- **The largest gains appear in cross-meeting recall and topic synthesis, where questions require long-term research context.**

Contributions

- Hierarchical memory: L1 evidence + short-term active state + L2/L3 topic memory.
- Evidence-linked topic context: retrieves L1 evidence first, then attaches linked L2/L3 topic state for cross-meeting reasoning.
- Inspectable correction: adjust L1 retrieval priority without rewriting raw memory.

Future Work

- Scale evaluation with more questions, a larger dataset, and human validation.
- Improve retrieval ranking to reduce missing-evidence cases.
- Study human-AI collaborative memory inspection and correction through the Memory Observatory.