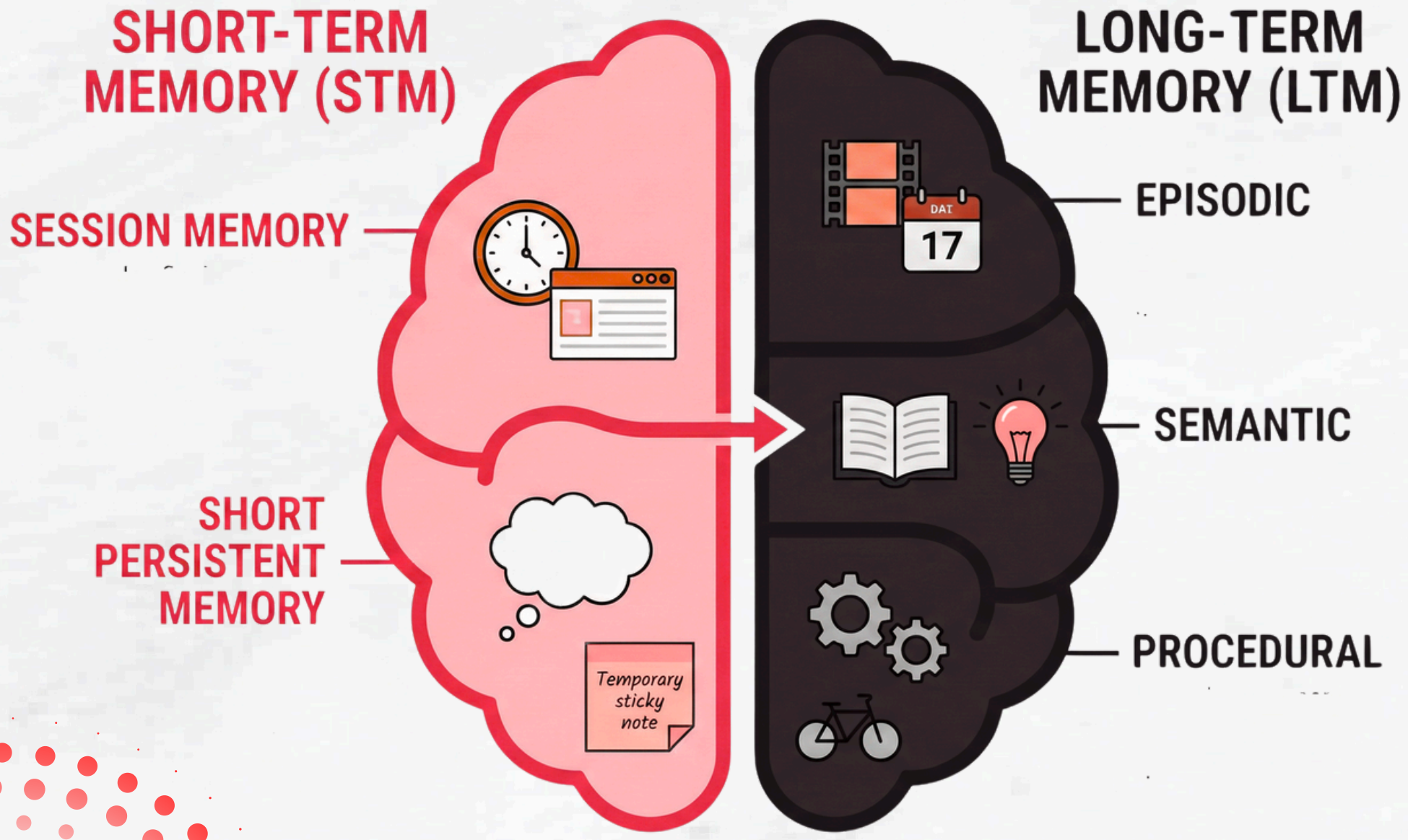
Memory Management for RAG & AI Agents

How AI systems remember, learn, and adapt



Naresh Edagotti
@PracticAI

Memory Types

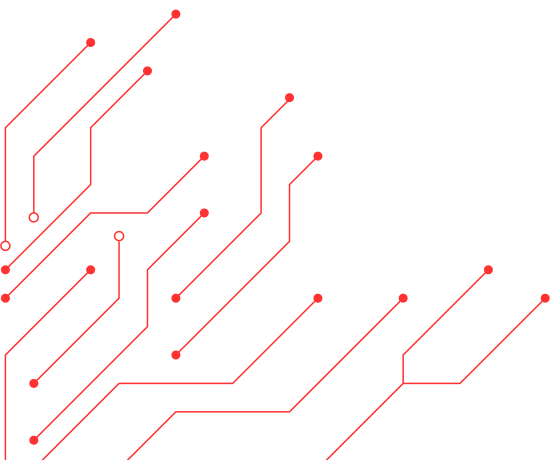
AI memory isn't magical, it mirrors human cognition through two fundamental layers:

1. **Short-Term Memory (STM):** Active, temporary context
2. **Long-Term Memory (LTM):** Persistent, retrievable knowledge

Why this matters

1. **Without STM:** Agents can't maintain conversations
2. **Without LTM:** Agents can't learn or personalize
3. **Together:** They enable truly intelligent interactions

Every memory type you hear about (conversational, episodic, semantic, procedural) falls into one of these two categories.



PracticAI

Short-Term Memory (STM)

Memory that exists only during active reasoning or ongoing tasks. It operates inside the model's context window.

STM includes:

- Session Memory
- Short-Term Persistent Memory

What it does

1. Keeps track of current conversation
2. Maintains context across turns
3. Helps resolve references ("that file", "previous answer")
4. Supports workflows happening right now

Example

User: "Extract key points from this PDF."
User next message: "Now convert them into bullets."
→ The agent remembers the extracted points temporarily.

Session Memory
(STM Type 1)

Active, temporary memory that stores everything happening in the current session until the window resets.

What it does

1. Tracks each message in the conversation
2. Maintains flow and context
3. Supports follow-up questions
4. Resets after session ends or context overflows

Example

User: “Book a cab from Bangalore to Mysore.”
User: “Make it for 6 am instead.”
→ It remembers the route (Bangalore → Mysore) from the previous turn.

Short-Term Persistent Memory (STM Type 2)

Memory that lasts beyond one session, but only for a limited duration. Acts like a temporary workspace.

What it does

1. Stores ongoing tasks, drafts, temporary preferences
2. Holds short-term goals over hours or days
3. Can be auto-promoted to long-term if needed

Example

Morning: “Start preparing slides for my ML lecture.”
Evening: “Continue the ML lecture slides.”
→ The agent recalls the draft it stored earlier.

Long-Term Memory (LTM)

Persistent knowledge that remains stored across days, weeks, or months.

LTM includes:

- Episodic Memory
- Semantic Memory
- Procedural Memory

What it does

1. Stores history, facts, and rules
2. Personalizes responses
3. Helps agents behave consistently over time
4. Enables permanent learning

Example

The system remembers the user’s work domain, writing style, preferences, and recurring tasks.

Episodic Memory
(LTM Type 1)

Memory of past events, user interactions, and time-based experiences.

What it does

1. Stores time-stamped sessions
2. Remembers past decisions and plans
3. Helps build user journey context

Example

“You created a marketing plan last month. Want to update it?”
→ The agent recalls an older session associated with the topic.

Semantic Memory
(LTM Type 2)

General knowledge memory that stores facts, rules, definitions, and domain understanding.

What it does

1. Holds product info, policies, documentation
2. Enables domain expertise
3. Supports accurate retrieval over large corpora

Example

“The GST rate for electronics is 18%.”
“Python’s zip() combines lists element-wise.”
This knowledge stays even when users change.

Procedural Memory (LTM Type 3)

Memory of how to perform tasks. Stores workflows, step-by-step actions, and automation sequences.

What it does

1. Executes repeated processes
2. Automates tasks reliably
3. Remembers rules for completing operations

Example

“Generate a weekly SEO report using the same steps as last time.”

“Follow this deployment pipeline every Friday.”

Use Cases

Healthcare AI

- STM: Current symptoms
- LTM Episodic: Medical history
- Semantic: Drug interactions
- Procedural: Treatment guidelines

E-commerce

- STM: Current browsing
- LTM: Past purchases & preferences

Customer Support

- STM: Current ticket
- LTM: Past issues, product documentation, troubleshooting flows

Memory Architecture

Short-Term Memory Layer

Session memory → Recent persistent memory

Fast access, temporary info

Long-Term Memory Layer

Episodic → Semantic → Procedural

Permanent, indexed, searchable

Flow:

User Query → STM Context → Retrieval from LTM → Merge → LLM Generation

Best Recommendations

✓ Use short-term memory for immediate context

✓ Store only important interactions into long-term

✓ Summarize long conversations before saving

✓ Use vector DBs for episodic & semantic memory

✓ Keep procedural memory modular (workflows)

✓ Prune stale memories regularly

✓ Respect privacy: user-controlled memory retention

Smart memory design = smarter RAG and agent behavior.

LIKE THIS

CONTENT?

FOLLOW FOR MORE!

NARESH EDAGOTTI

LIKE REPOST SAVE