Agent Memory Architectures

Short-Term Memory

(Working Memory)

- Context window

- Current conversation

- Scratchpad for reasoning

- Limited by token count

Duration: Seconds to minutes

Long-Term Memory

(Persistent Storage)

- Vector database

- Knowledge base

- User preferences

- Learned patterns

Duration: Days to permanent

AI Agent

Read/Write

Retrieve/Store

Recall/Record

Episodic Memory

(Experience Replay)

- Past task executions

- Success/failure records

- Strategy refinement

Duration: Session to permanent

Capacity: Small (4K-128K tokens) Large (millions of vectors)