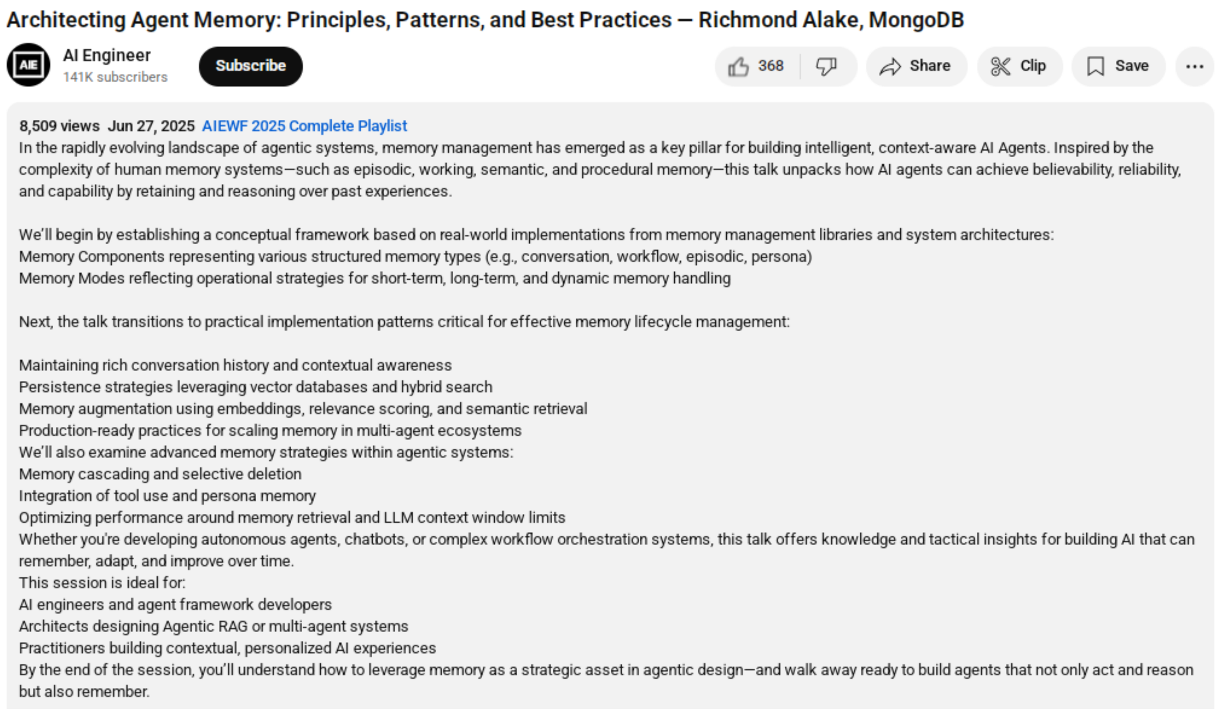


Architecting Agent Memory: Principles, Patterns, and Best Practices — Richmond Alake, MongoDB

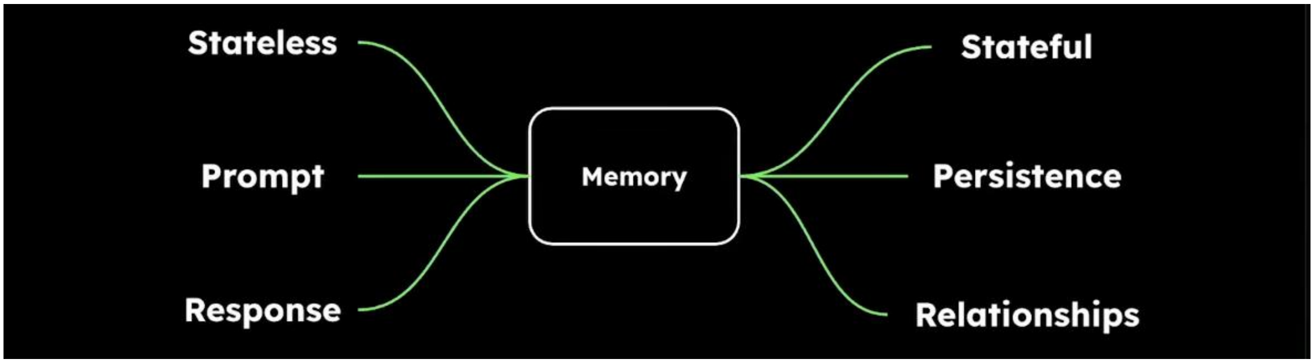
8,509 views Jun 27, 2025 AIEWF 2025 Complete Playlist

In the rapidly evolving landscape of agentic systems, memory management has emerged as a key pillar for building intelligent, context-aware AI Agents. Inspired by the complexity of human memory systems—such as episodic, working, semantic, and procedural memory—this talk unpacks how AI agents can achieve believability, reliability, and capability by retaining and reasoning over past experiences.

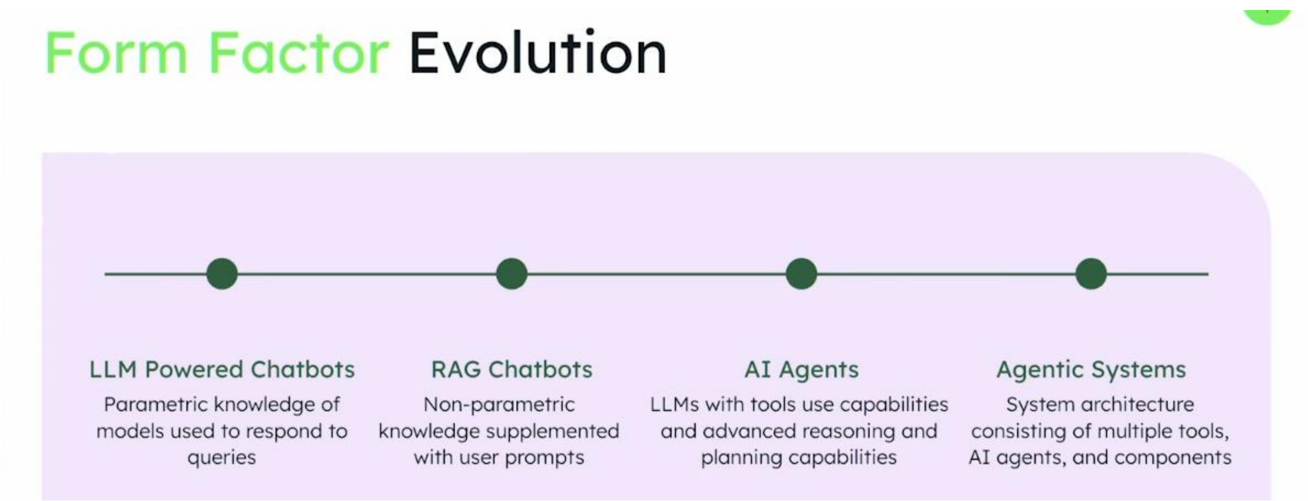
We'll begin by establishing a conceptual framework based on real-world implementations from memory management libraries and system architectures:
Memory Components representing various structured memory types (e.g., conversation, workflow, episodic, persona)
Memory Modes reflecting operational strategies for short-term, long-term, and dynamic memory handling

Next, the talk transitions to practical implementation patterns critical for effective memory lifecycle management:

Maintaining rich conversation history and contextual awareness
Persistence strategies leveraging vector databases and hybrid search
Memory augmentation using embeddings, relevance scoring, and semantic retrieval
Production-ready practices for scaling memory in multi-agent ecosystems
We'll also examine advanced memory strategies within agentic systems:
Memory cascading and selective deletion
Integration of tool use and persona memory
Optimizing performance around memory retrieval and LLM context window limits
Whether you're developing autonomous agents, chatbots, or complex workflow orchestration systems, this talk offers knowledge and tactical insights for building AI that can remember, adapt, and improve over time.
This session is ideal for:
AI engineers and agent framework developers
Architects designing Agentic RAG or multi-agent systems
Practitioners building contextual, personalized AI experiences
By the end of the session, you'll understand how to leverage memory as a strategic asset in agentic design—and walk away ready to build agents that not only act and reason but also remember.



Stateless
Prompt
Response
Memory
Stateful
Persistence
Relationships



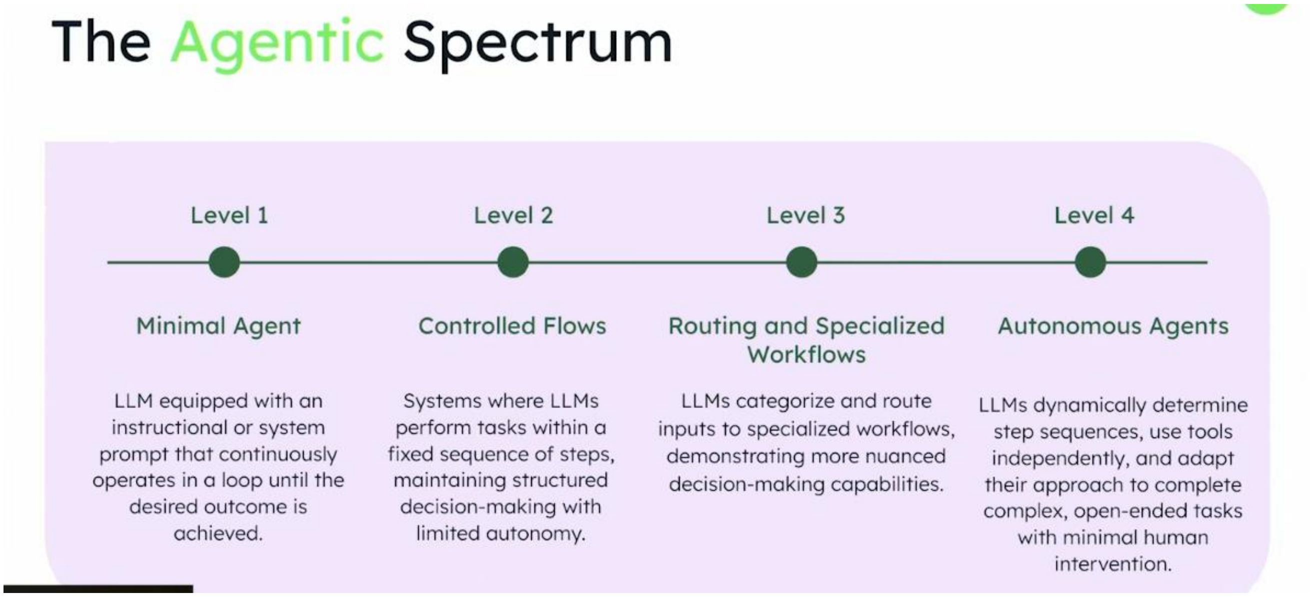
Form Factor Evolution

LLM Powered Chatbots
Parametric knowledge of models used to respond to queries

RAG Chatbots
Non-parametric knowledge supplemented with user prompts

AI Agents
LLMs with tools use capabilities and advanced reasoning and planning capabilities

Agentic Systems
System architecture consisting of multiple tools, AI agents, and components



The Agentic Spectrum

Level 1

Minimal Agent

LLM equipped with an instructional or system prompt that continuously operates in a loop until the desired outcome is achieved.

Level 2

Controlled Flows

Systems where LLMs perform tasks within a fixed sequence of steps, maintaining structured decision-making with limited autonomy.

Level 3

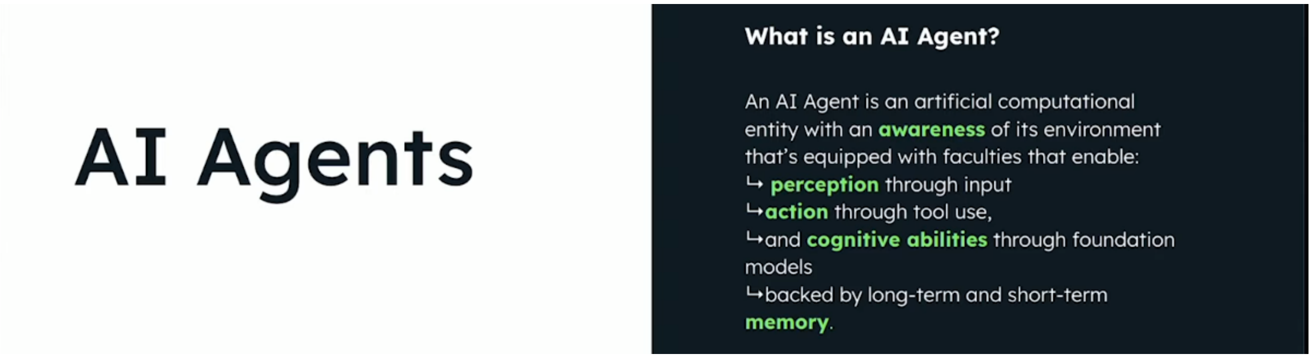
Routing and Specialized Workflows

LLMs categorize and route inputs to specialized workflows, demonstrating more nuanced decision-making capabilities.

Level 4

Autonomous Agents

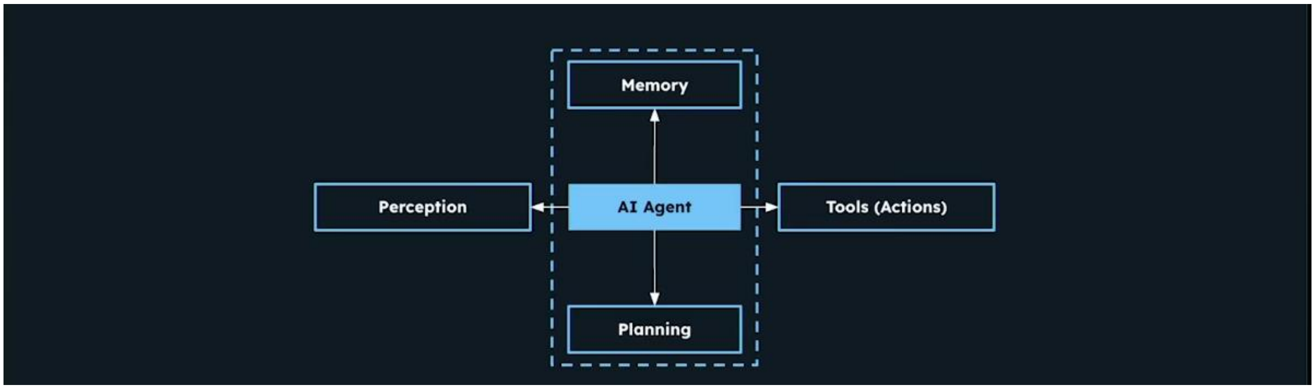
LLMs dynamically determine step sequences, use tools independently, and adapt their approach to complete complex, open-ended tasks with minimal human intervention.

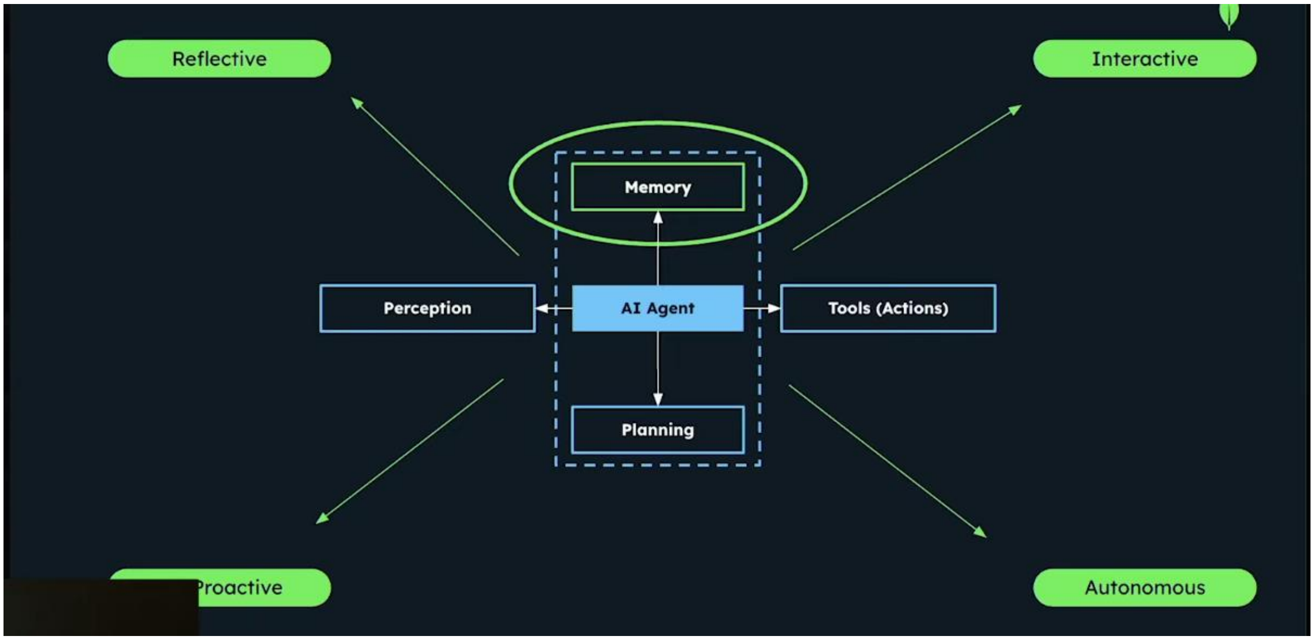


AI Agents

What is an AI Agent?

An AI Agent is an artificial computational entity with an **awareness** of its environment that's equipped with faculties that enable:
↳ **perception** through input
↳**action** through tool use,
↳and **cognitive abilities** through foundation models
↳backed by long-term and short-term **memory**.





Reflective

Interactive

Memory

Perception

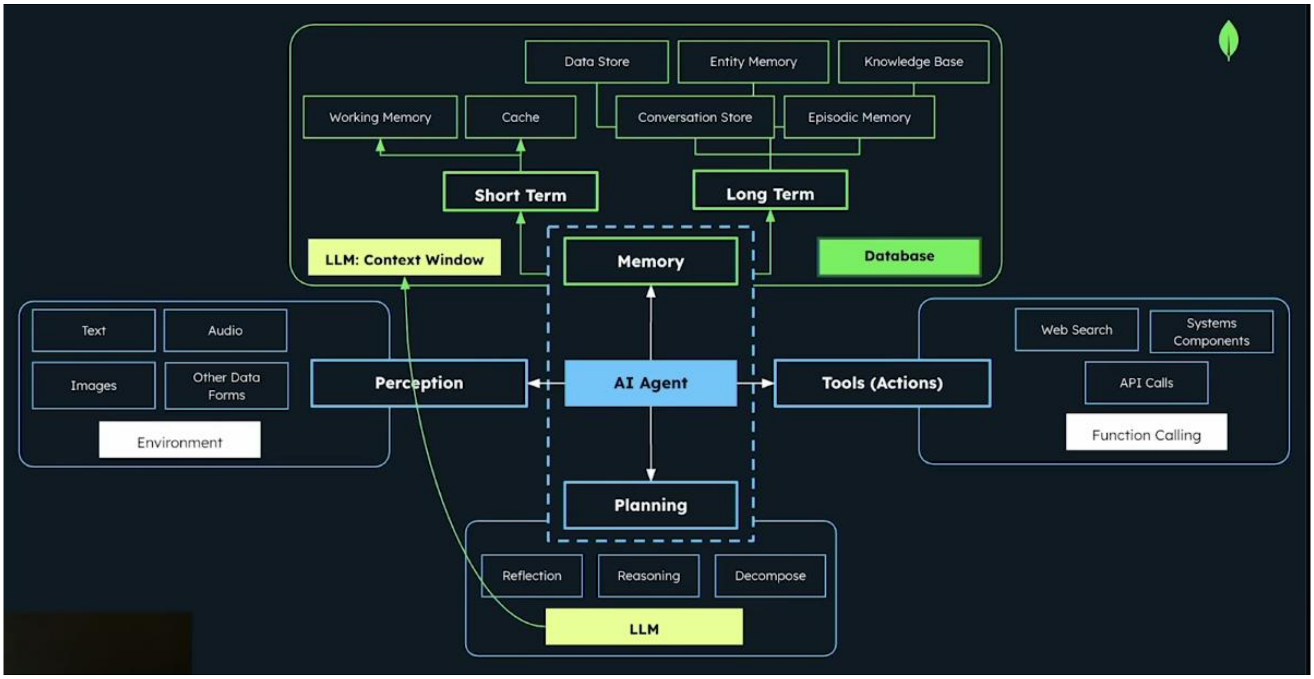
AI Agent

Tools (Actions)

Planning

[redacted]roactive

Autonomous



Data Store

Entity Memory

Knowledge Base

Working Memory

Cache

Conversation Store

Episodic Memory

Short Term

Long Term

LLM: Context Window

Memory

Database

Text

Audio

Images

Other Data Forms

Environment

Perception

AI Agent

Tools (Actions)

Web Search

Systems Components

API Calls

Function Calling

Planning

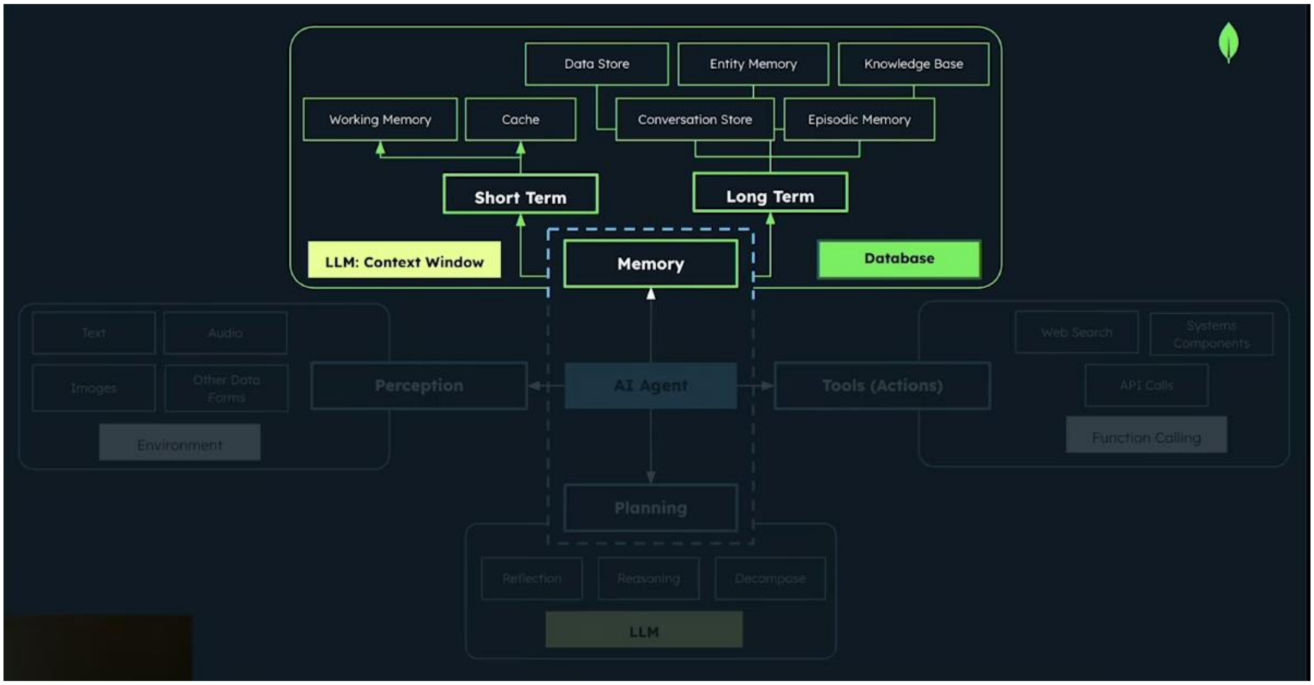
Reflection

Reasoning

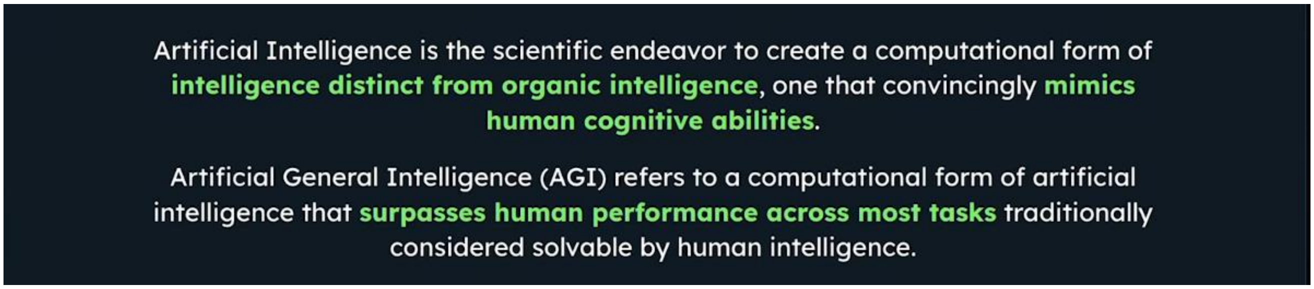
Decompose

LLM

[redacted]

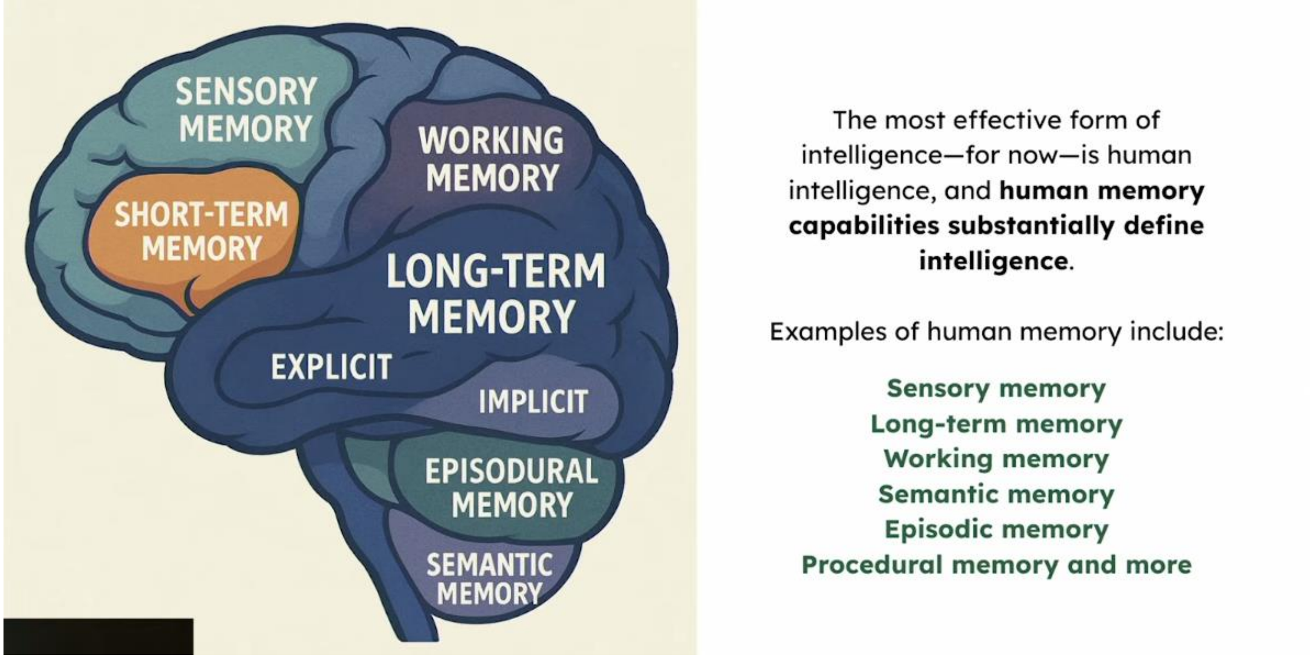


[redacted]



Artificial Intelligence is the scientific endeavor to create a computational form of **intelligence distinct from organic intelligence**, one that convincingly **mimics human cognitive abilities**.

Artificial General Intelligence (AGI) refers to a computational form of artificial intelligence that **surpasses human performance across most tasks** traditionally considered solvable by human intelligence.

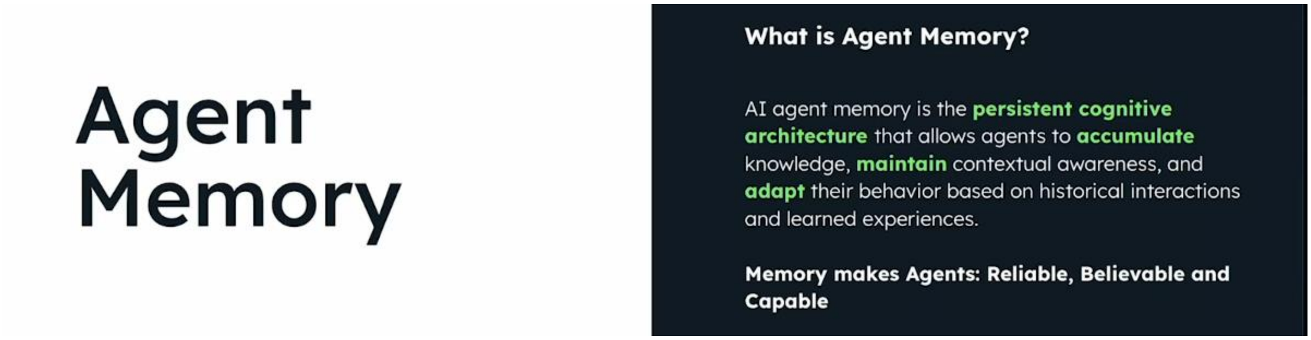


[redacted]

The most effective form of intelligence—for now—is human intelligence, and **human memory capabilities substantially define intelligence**.

Examples of human memory include:

Sensory memory
Long-term memory
Working memory
Semantic memory
Episodic memory
Procedural memory and more

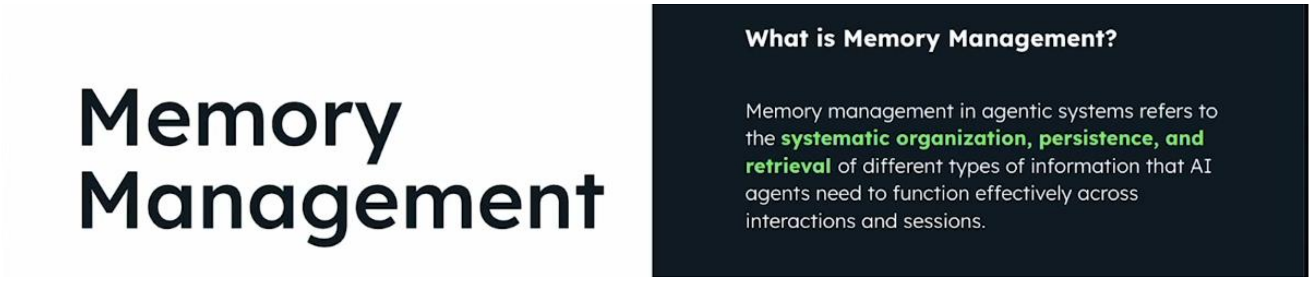


Agent Memory

What is Agent Memory?

AI agent memory is the **persistent cognitive architecture** that allows agents to **accumulate** knowledge, **maintain** contextual awareness, and **adapt** their behavior based on historical interactions and learned experiences.

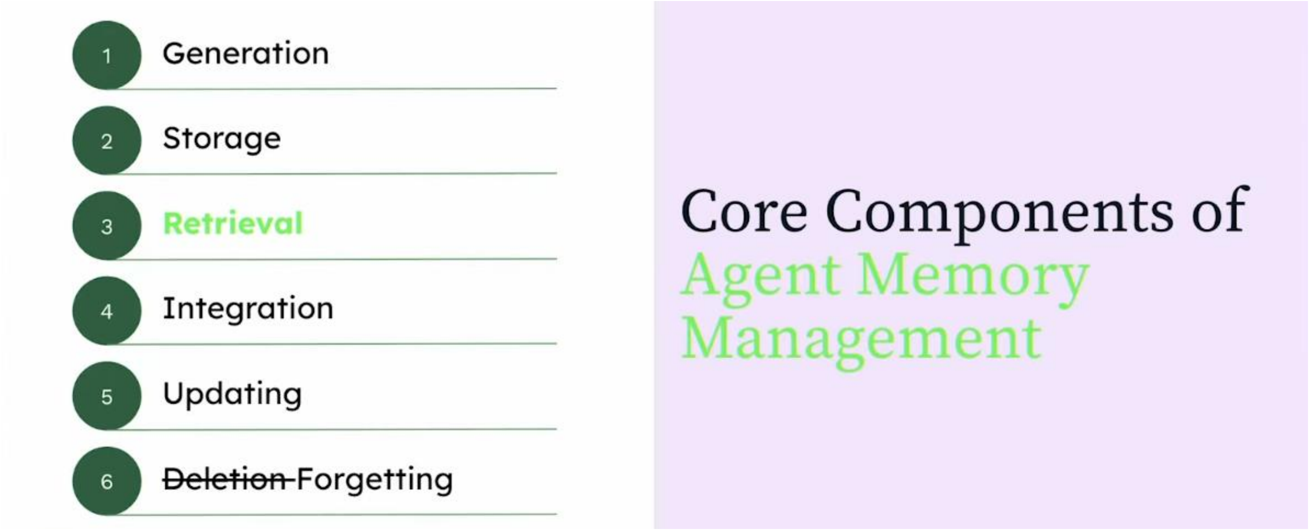
Memory makes Agents: Reliable, Believable and Capable



Memory Management

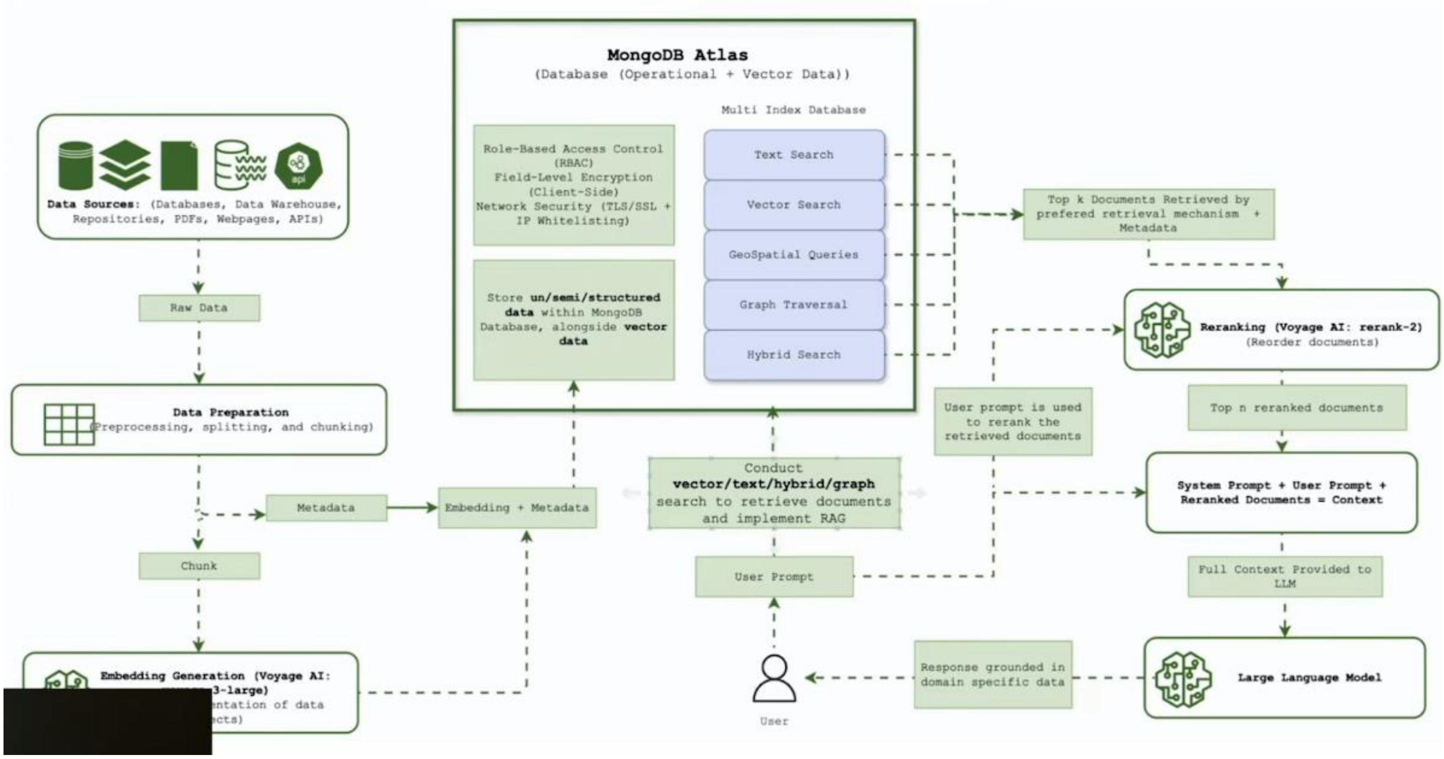
What is Memory Management?

Memory management in agentic systems refers to the **systematic organization, persistence, and retrieval** of different types of information that AI agents need to function effectively across interactions and sessions.

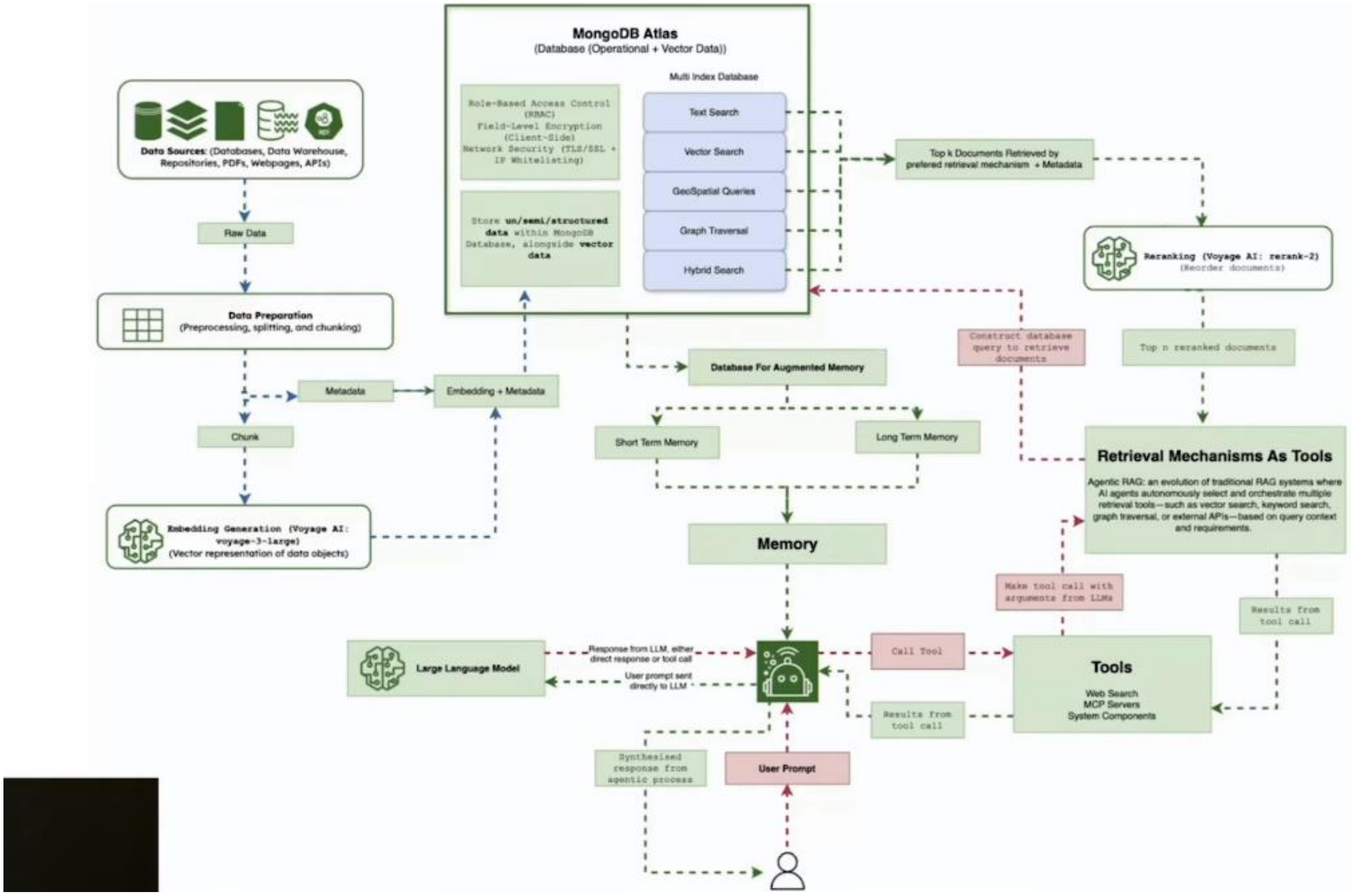


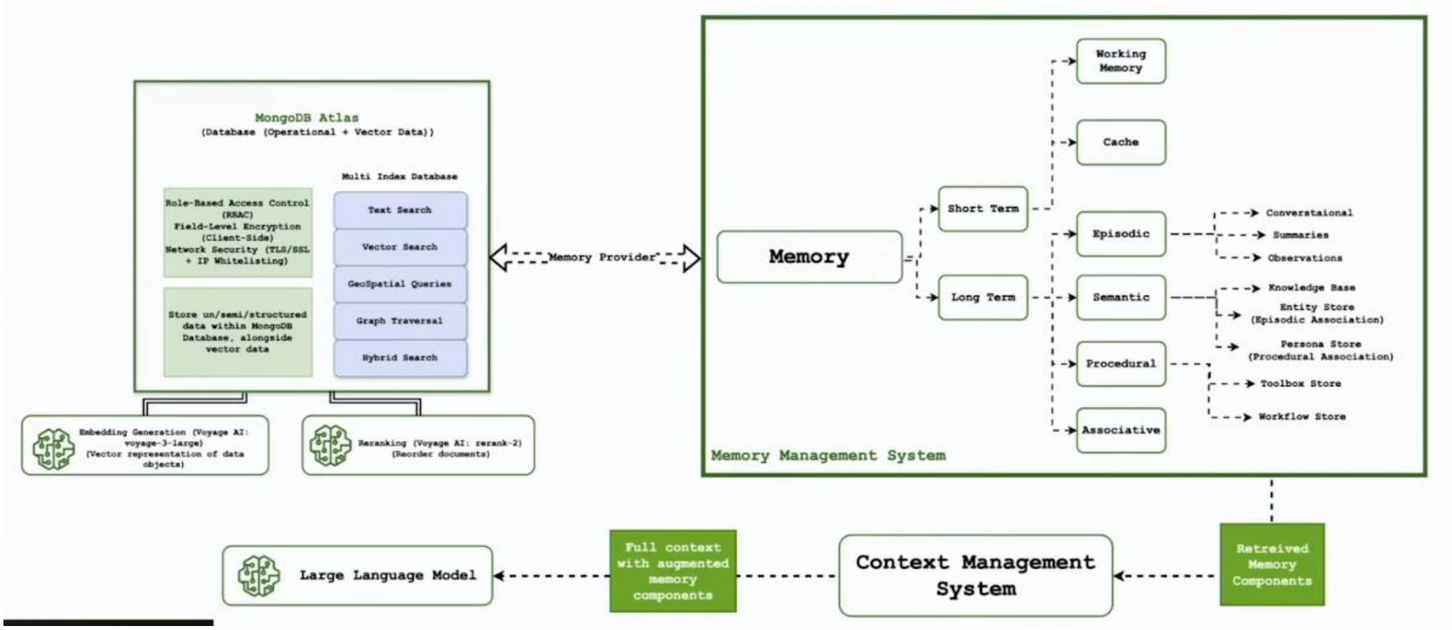
1. Generation
2. Storage
3. **Retrieval**
4. Integration
5. Updating
6. ~~Deletion~~ Forgetting

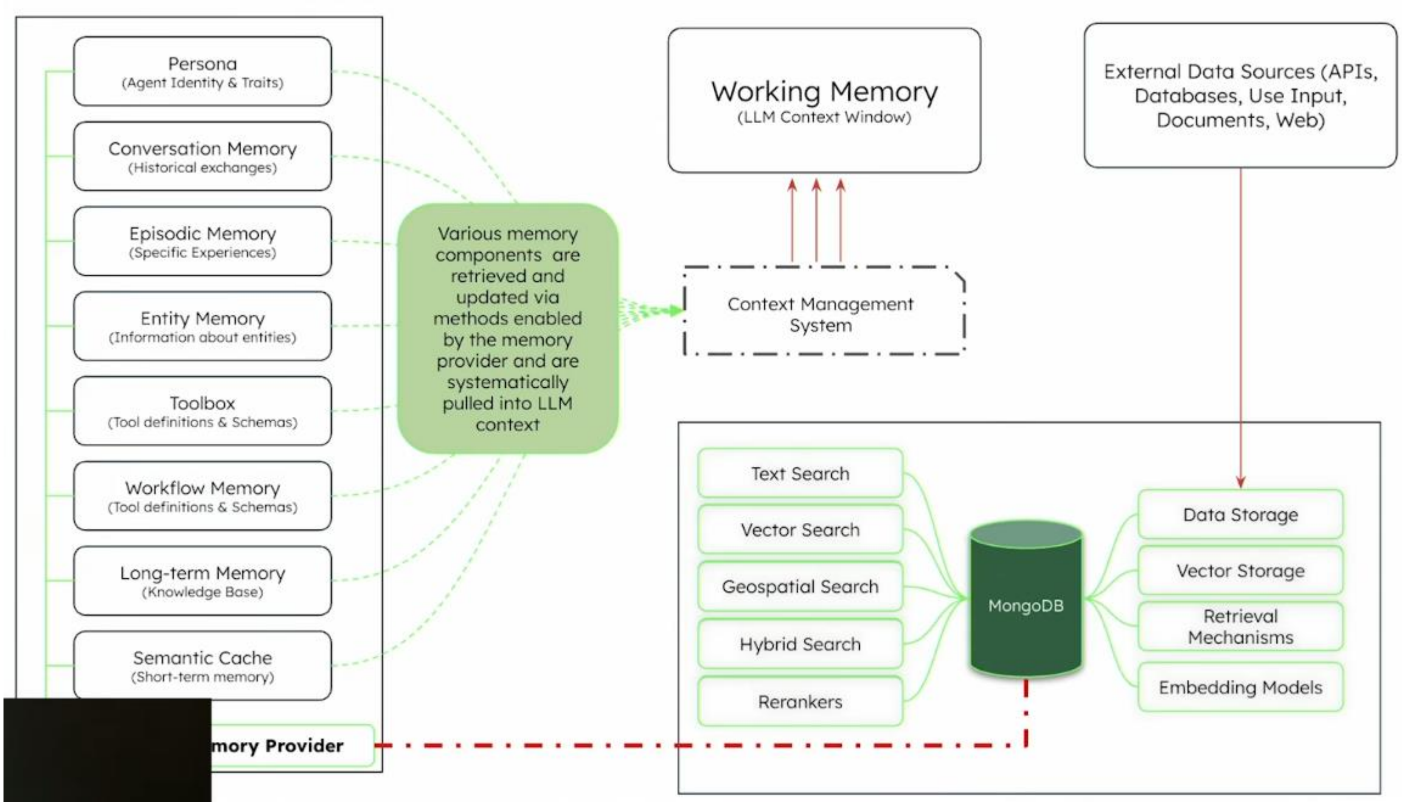
Core Components of Agent Memory Management

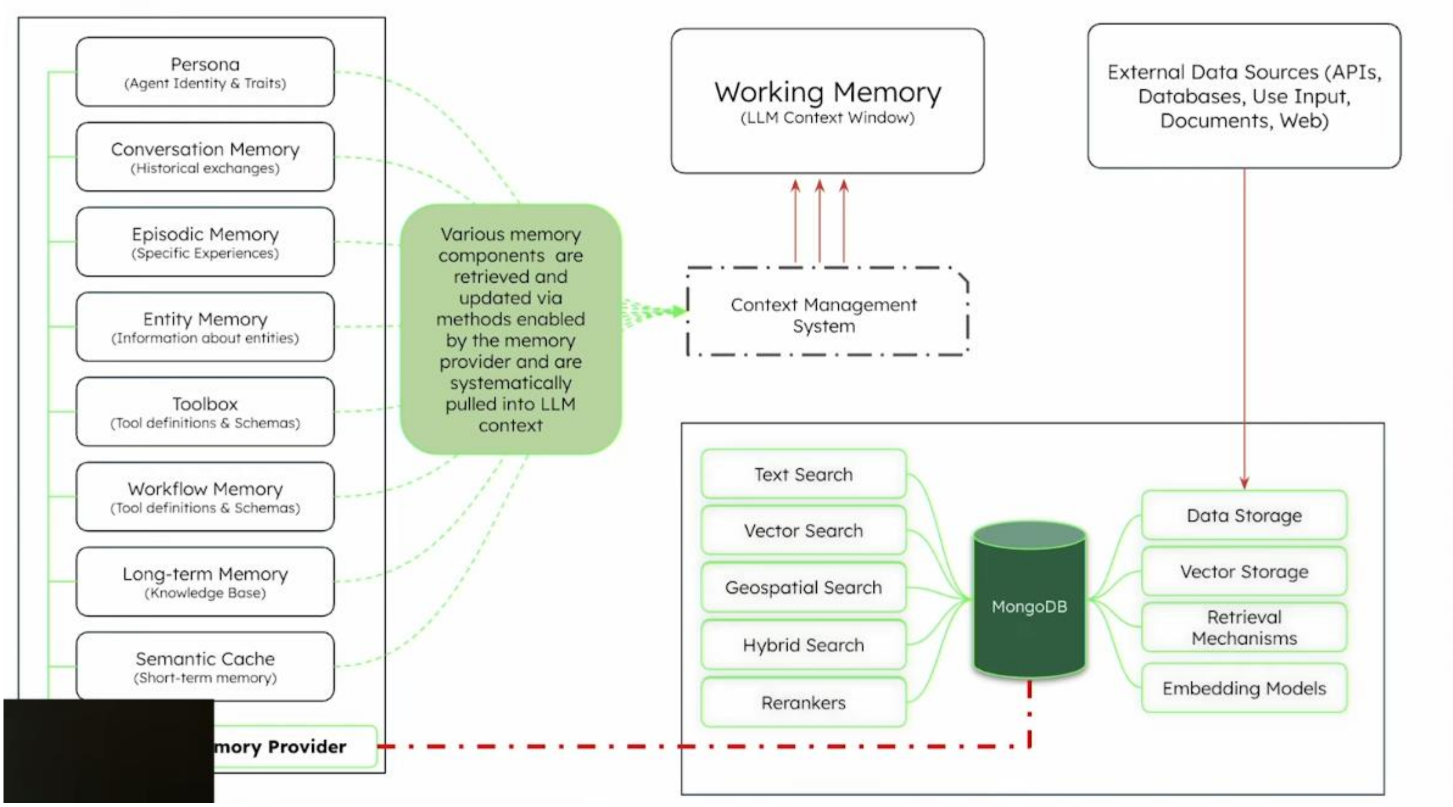

MongoDB Atlas
(Database (Operational + Vector Data))
Multi Index Database
Role-Based Access Control (RBAC)
Field-Level Encryption (Client-Side)
Network Security (TLS/SSL + IP Whitelisting)
Store un/semi/structured data within MongoDB Database, alongside vector data
Text Search
Vector Search
GeoSpatial Queries
Graph Traversal
Hybrid Search
Data Sources: (Databases, Data Warehouse, Repositories, PDFs, Webpages, APIs)
api
Raw Data
Data Preparation
(Preprocessing, splitting, and chunking)
Metadata
Embedding + Metadata
Chunk
Embedding Generation (Voyage AI:
3-large)
entation of data
ects)
Top k Documents Retrieved by prefered retrieval mechanism + Metadata
Reranking (Voyage AI: rerank-2)
(Reorder documents)
Top n reranked documents
User prompt is used to rerank the retrieved documents
System Prompt + User Prompt + Reranked Documents = Context
Full Context Provided to LLM
Conduct vector/text/hybrid/graph search to retrieve documents and implement RAG
User Prompt
Large Language Model
Response grounded in domain specific data
User

[redacted]


MongoDB Atlas
(Database (Operational + Vector Data))
Multi Index Database
Role-Based Access Control (RBAC)
Field-Level Encryption (Client-Side)
Network Security (TLS/SSL + IP Whitelisting)
Store un/semi/structured data within MongoDB Database, alongside vector data
Text Search
Vector Search
GeoSpatial Queries
Graph Traversal
Hybrid Search
Data Sources: (Databases, Data Warehouse, Repositories, PDFs, Webpages, APIs)
api
Raw Data
Data Preparation
(Preprocessing, splitting, and chunking)
Metadata
Embedding + Metadata
Chunk
Embedding Generation (Voyage AI: voyage-3-large)
(Vector representation of data objects)
Top k Documents Retrieved by prefered retrieval mechanism + Metadata
Reranking (Voyage AI: rerank-2)
(Reorder documents)
Top n reranked documents
Construct database query to retrieve documents
Database For Augmented Memory
Short Term Memory
Long Term Memory
Memory
Retrieval Mechanisms As Tools
Agentic RAG: an evolution of traditional RAG systems where AI agents autonomously select and orchestrate multiple retrieval tools—such as vector search, keyword search, graph traversal, or external APIs—based on query context and requirements.
Make tool call with arguments from LLMs
Results from tool call
Large Language Model
Response from LLM, either direct response or tool call
User prompt sent directly to LLM
Call Tool
Tools
Web Search
MCP Servers
System Components
Results from tool call
Synthesised response from agentic process
User Prompt


MongoDB Atlas
(Database (Operational + Vector Data))
Multi Index Database
Role-Based Access Control (RBAC)
Field-Level Encryption (Client-Side)
Network Security (TLS/SSL + IP Whitelisting)
Store un/semi/structured data within MongoDB Database, alongside vector data
Text Search
Vector Search
GeoSpatial Queries
Graph Traversal
Hybrid Search
Embedding Generation (Voyage AI: voyage-3-large)
(Vector representation of data objects)
Reranking (Voyage AI: rerank-2)
(Reorder documents)
Memory Provider
Memory
Short Term
Long Term
Working Memory
Cache
Episodic
Semantic
Procedural
Associative
Converstaional
Summaries
Observations
Knowledge Base
Entity Store
(Episodic Association)
Persona Store
(Procedural Association)
Toolbox Store
Workflow Store
Memory Management System
Retreived Memory Components
Context Management System
Full context with augmented memory components
Large Language Model


Persona
(Agent Identity & Traits)
Conversation Memory
(Historical exchanges)
Episodic Memory
(Specific Experiences)
Entity Memory
(Information about entities)
Toolbox
(Tool definitions & Schemas)
Workflow Memory
(Tool definitions & Schemas)
Long-term Memory
(Knowledge Base)
Semantic Cache
(Short-term memory)
[redacted]mory Provider
Various memory components are retrieved and updated via methods enabled by the memory provider and are systematically pulled into LLM context
Working Memory
(LLM Context Window)
Context Management System
External Data Sources (APIs, Databases, Use Input, Documents, Web)
Text Search
Vector Search
Geospatial Search
Hybrid Search
Rerankers
MongoDB
Data Storage
Vector Storage
Retrieval Mechanisms
Embedding Models



Persona (Agent Identity & Traits)

Conversation Memory (Historical exchanges)

Episodic Memory (Specific Experiences)

Entity Memory (Information about entities)

Toolbox (Tool definitions & Schemas)

Workflow Memory (Tool definitions & Schemas)

Long-term Memory (Knowledge Base)

Semantic Cache (Short-term memory)

[redacted]mory Provider

Various memory components are retrieved and updated via methods enabled by the memory provider and are systematically pulled into LLM context

Working Memory (LLM Context Window)

Context Management System

External Data Sources (APIs, Databases, Use Input, Documents, Web)

Text Search

Vector Search

Geospatial Search

Hybrid Search

Rerankers

MongoDB

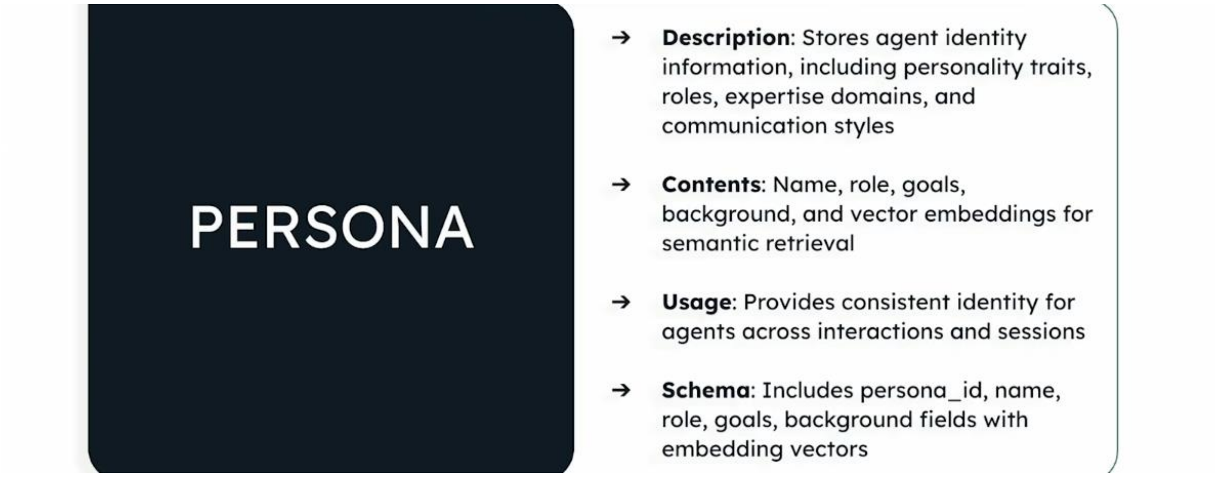
Data Storage

Vector Storage

Retrieval Mechanisms

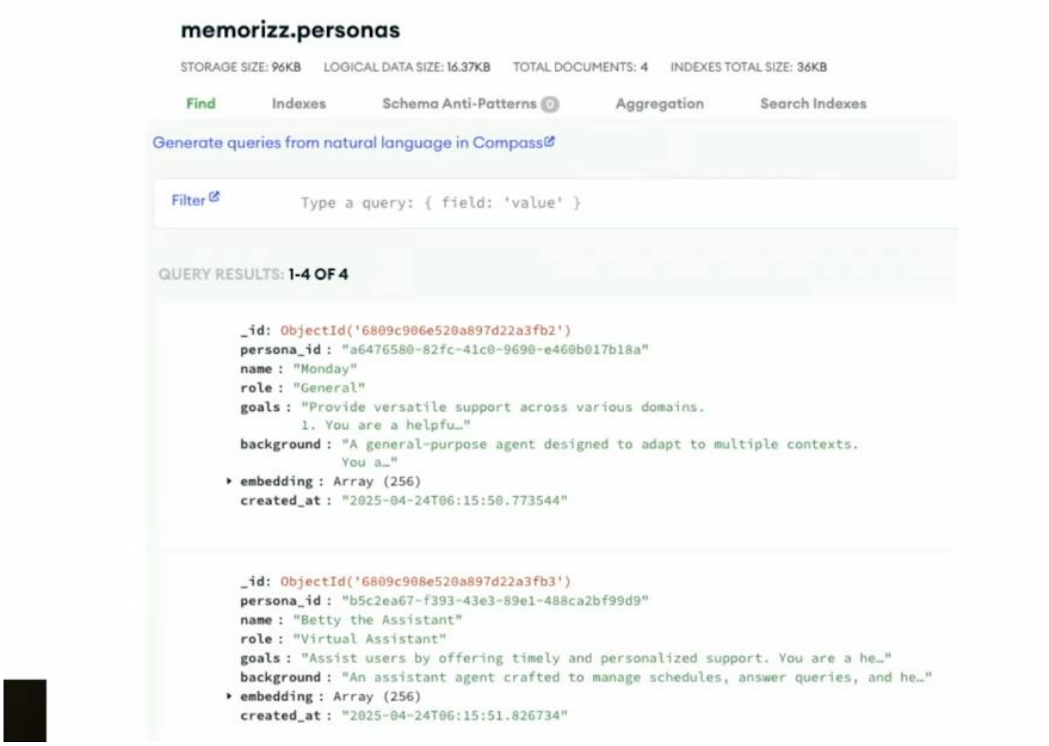
Embedding Models

Forms of Memory in AI Agents



PERSONA

- **Description**: Stores agent identity information, including personality traits, roles, expertise domains, and communication styles
- **Contents**: Name, role, goals, background, and vector embeddings for semantic retrieval
- **Usage**: Provides consistent identity for agents across interactions and sessions
- **Schema**: Includes persona_id, name, role, goals, background fields with embedding vectors



memorizz.personas

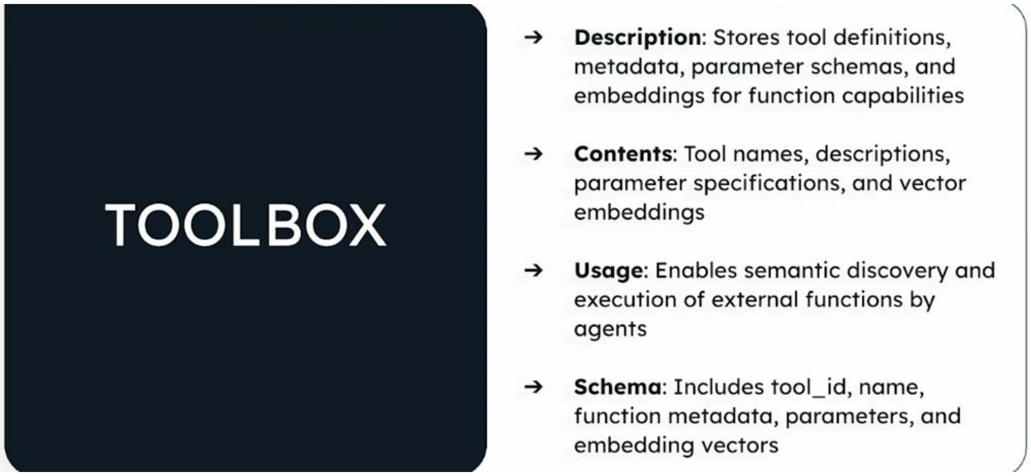
STORAGE SIZE: 96KB   LOGICAL DATA SIZE: 16.37KB   TOTAL DOCUMENTS: 4   INDEXES TOTAL SIZE: 36KB

Type a query: { field: 'value' }

QUERY RESULTS: 1-4 OF 4

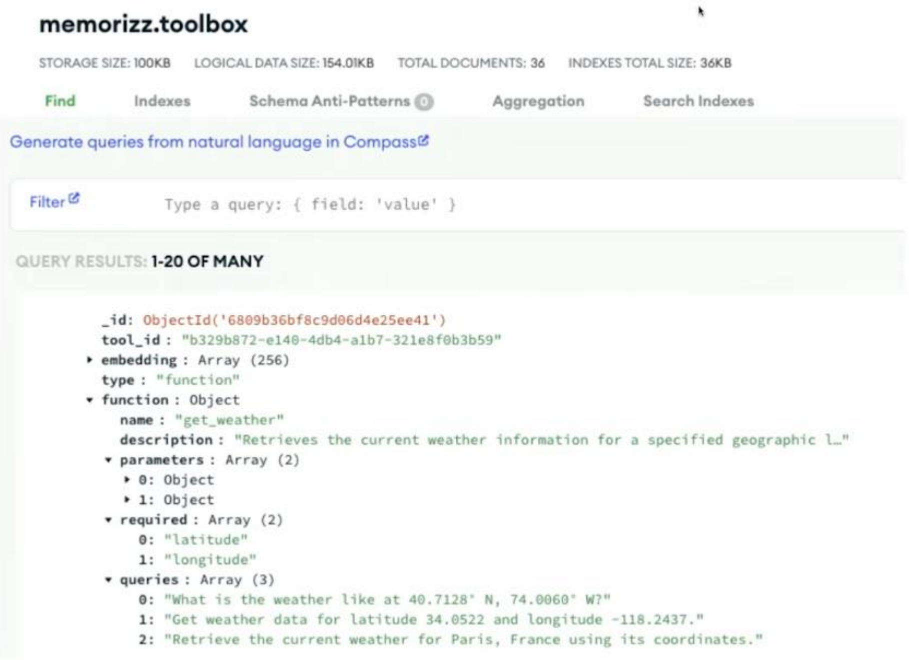
```
  _id: ObjectId('6809c906e520a897d22a3fb2')
  persona_id : "a6476580-82fc-41c0-9690-e460b017b18a"
  name : "Monday"
  role : "General"
  goals : "Provide versatile support across various domains.
          1. You are a helpfu…"
  background : "A general-purpose agent designed to adapt to multiple contexts.
               You a…"
▸ embedding : Array (256)
  created_at : "2025-04-24T06:15:50.773544"

  _id: ObjectId('6809c908e520a897d22a3fb3')
  persona_id : "b5c2ea67-f393-43e3-89e1-488ca2bf99d9"
  name : "Betty the Assistant"
  role : "Virtual Assistant"
  goals : "Assist users by offering timely and personalized support. You are a he…"
  background : "An assistant agent crafted to manage schedules, answer queries, and he…"
▸ embedding : Array (256)
  created_at : "2025-04-24T06:15:51.826734"
```

[redacted]



TOOLBOX

- → **Description**: Stores tool definitions, metadata, parameter schemas, and embeddings for function capabilities
- → **Contents**: Tool names, descriptions, parameter specifications, and vector embeddings
- → **Usage**: Enables semantic discovery and execution of external functions by agents
- → **Schema**: Includes tool_id, name, function metadata, parameters, and embedding vectors



memorizz.toolbox

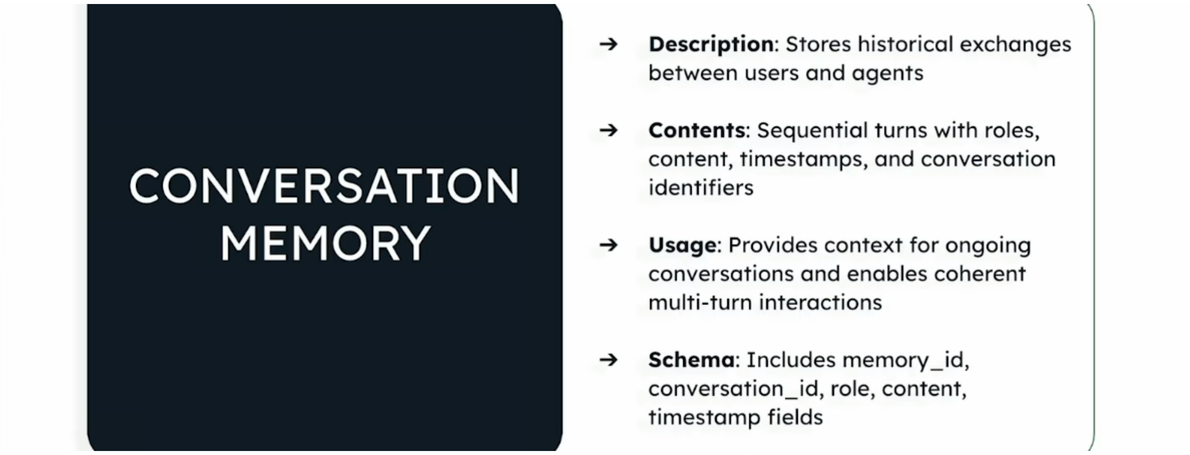
STORAGE SIZE: 100KB   LOGICAL DATA SIZE: 154.01KB   TOTAL DOCUMENTS: 36   INDEXES TOTAL SIZE: 36KB

Type a query: { field: 'value' }

QUERY RESULTS: 1-20 OF MANY

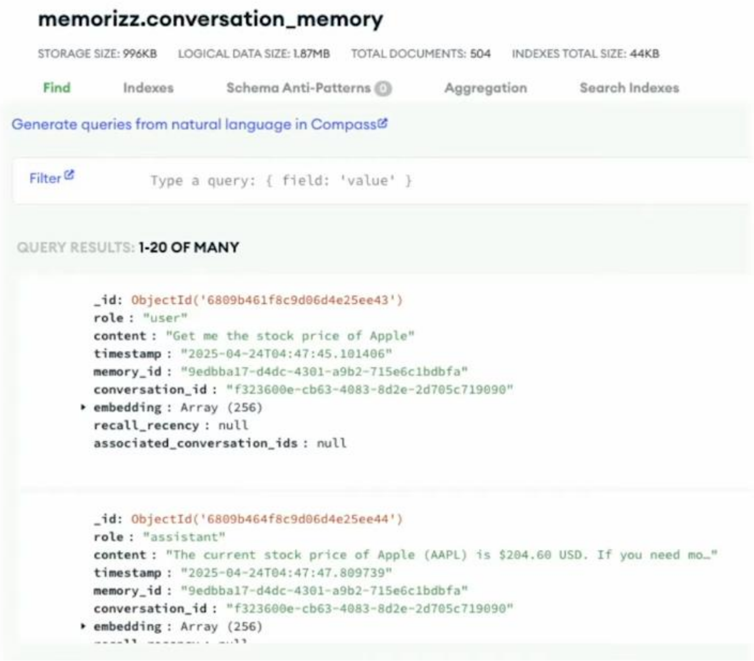
```
  _id: ObjectId('6809b36bf8c9d06d4e25ee41')
  tool_id : "b329b872-e140-4db4-a1b7-321e8f0b3b59"
▸ embedding : Array (256)
  type : "function"
▾ function : Object
    name : "get_weather"
    description : "Retrieves the current weather information for a specified geographic l…"
  ▾ parameters : Array (2)
    ▸ 0: Object
    ▸ 1: Object
  ▾ required : Array (2)
      0: "latitude"
      1: "longitude"
  ▾ queries : Array (3)
      0: "What is the weather like at 40.7128° N, 74.0060° W?"
      1: "Get weather data for latitude 34.0522 and longitude -118.2437."
      2: "Retrieve the current weather for Paris, France using its coordinates."
```

[redacted]



CONVERSATION MEMORY

- → **Description**: Stores historical exchanges between users and agents
- → **Contents**: Sequential turns with roles, content, timestamps, and conversation identifiers
- → **Usage**: Provides context for ongoing conversations and enables coherent multi-turn interactions
- → **Schema**: Includes memory_id, conversation_id, role, content, timestamp fields



memorizz.conversation_memory

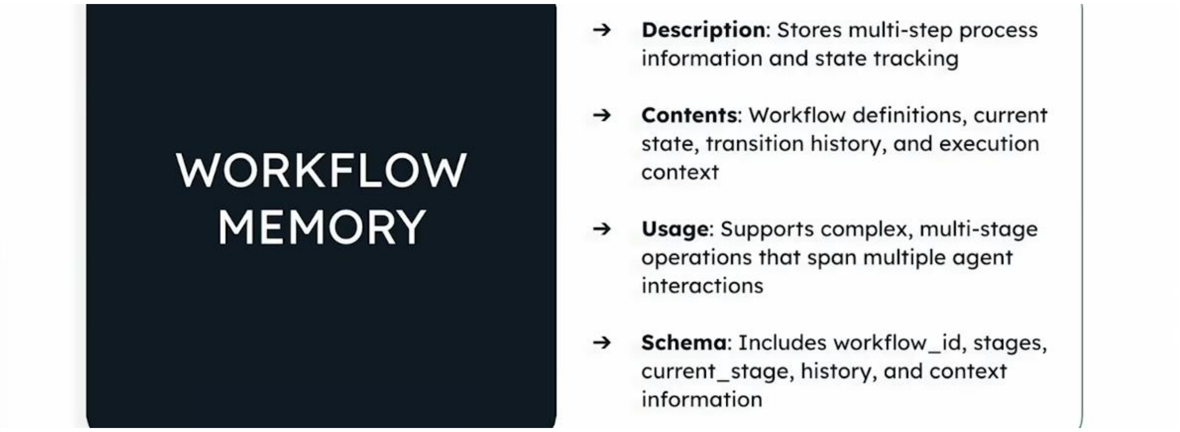
STORAGE SIZE: 996KB LOGICAL DATA SIZE: 1.87MB TOTAL DOCUMENTS: 504 INDEXES TOTAL SIZE: 44KB

Find
Indexes
Schema Anti-Patterns 0
Aggregation
Search Indexes
Generate queries from natural language in Compass
Filter
Type a query: { field: 'value' }

QUERY RESULTS: 1-20 OF MANY

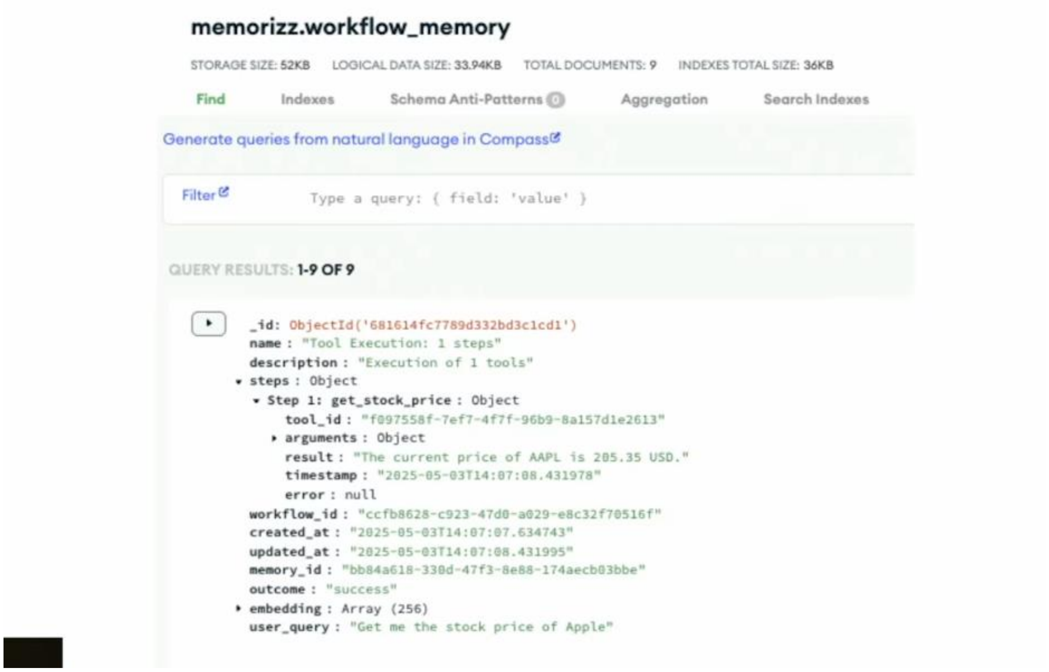
```
_id: ObjectId('6809b461f8c9d06d4e25ee43')
role : "user"
content : "Get me the stock price of Apple"
timestamp : "2025-04-24T04:47:45.101406"
memory_id : "9edbba17-d4dc-4301-a9b2-715e6c1bdbfa"
conversation_id : "f323600e-cb63-4083-8d2e-2d705c719090"
▸ embedding : Array (256)
recall_recency : null
associated_conversation_ids : null
```

```
_id: ObjectId('6809b464f8c9d06d4e25ee44')
role : "assistant"
content : "The current stock price of Apple (AAPL) is $204.60 USD. If you need mo…"
timestamp : "2025-04-24T04:47:47.809739"
memory_id : "9edbba17-d4dc-4301-a9b2-715e6c1bdbfa"
conversation_id : "f323600e-cb63-4083-8d2e-2d705c719090"
▸ embedding : Array (256)
recall_recency : null
```



WORKFLOW MEMORY

- → **Description**: Stores multi-step process information and state tracking
- → **Contents**: Workflow definitions, current state, transition history, and execution context
- → **Usage**: Supports complex, multi-stage operations that span multiple agent interactions
- → **Schema**: Includes workflow_id, stages, current_stage, history, and context information



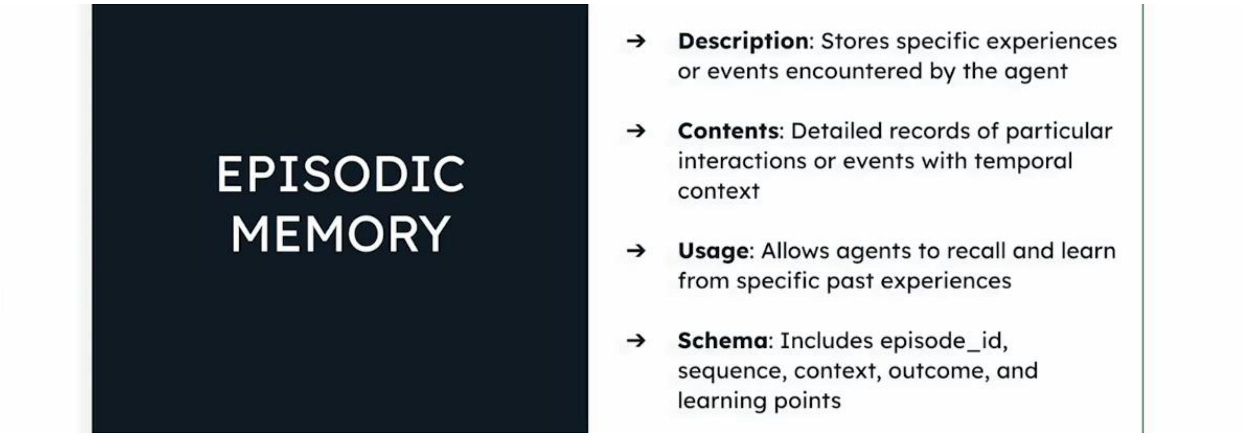
memorizz.workflow_memory

STORAGE SIZE: 52KB   LOGICAL DATA SIZE: 33.94KB   TOTAL DOCUMENTS: 9   INDEXES TOTAL SIZE: 36KB

Find
Indexes
Schema Anti-Patterns 0
Aggregation
Search Indexes
Generate queries from natural language in Compass
Filter
Type a query: { field: 'value' }

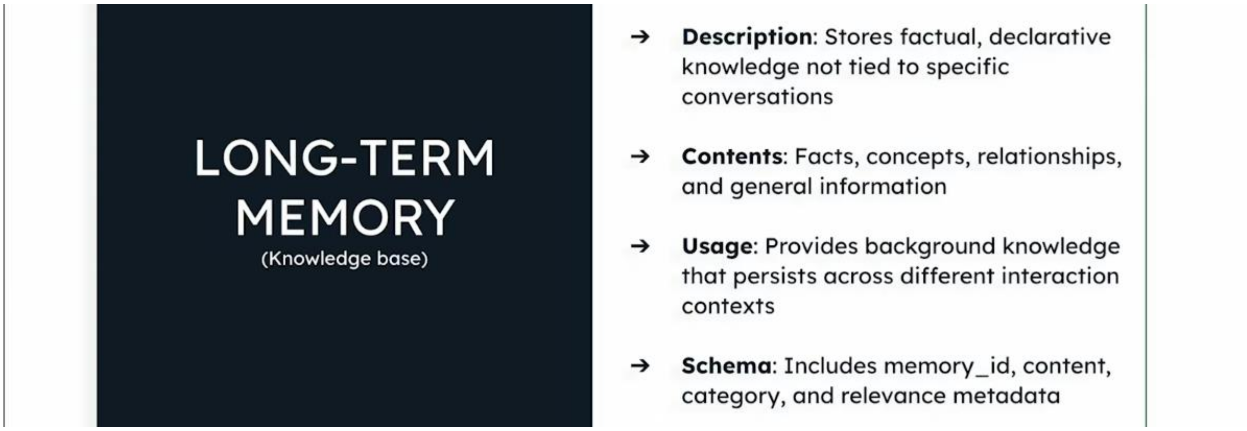
QUERY RESULTS: 1-9 OF 9

```
_id: ObjectId('681614fc7789d332bd3c1cd1')
name : "Tool Execution: 1 steps"
description : "Execution of 1 tools"
▾ steps : Object
  ▾ Step 1: get_stock_price : Object
      tool_id : "f097558f-7ef7-4f7f-96b9-8a157d1e2613"
    ▸ arguments : Object
      result : "The current price of AAPL is 205.35 USD."
      timestamp : "2025-05-03T14:07:08.431978"
      error : null
workflow_id : "ccfb8628-c923-47d0-a029-e8c32f70516f"
created_at : "2025-05-03T14:07:07.634743"
updated_at : "2025-05-03T14:07:08.431995"
memory_id : "bb84a618-330d-47f3-8e88-174aecb03bbe"
outcome : "success"
▸ embedding : Array (256)
user_query : "Get me the stock price of Apple"
```



EPISODIC MEMORY

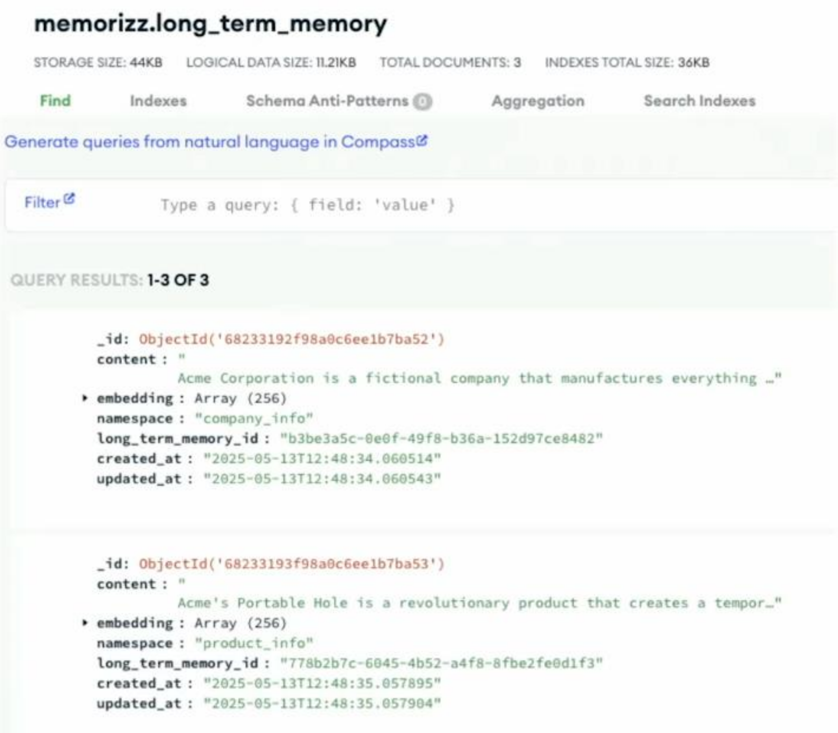
- **Description**: Stores specific experiences or events encountered by the agent
- **Contents**: Detailed records of particular interactions or events with temporal context
- **Usage**: Allows agents to recall and learn from specific past experiences
- **Schema**: Includes episode_id, sequence, context, outcome, and learning points



LONG-TERM MEMORY

(Knowledge base)

- **Description**: Stores factual, declarative knowledge not tied to specific conversations
- **Contents**: Facts, concepts, relationships, and general information
- **Usage**: Provides background knowledge that persists across different interaction contexts
- **Schema**: Includes memory_id, content, category, and relevance metadata



memorizz.long_term_memory

STORAGE SIZE: 44KB LOGICAL DATA SIZE: 11.21KB TOTAL DOCUMENTS: 3 INDEXES TOTAL SIZE: 36KB

Find
Indexes
Schema Anti-Patterns 0
Aggregation
Search Indexes
Generate queries from natural language in Compass
Filter
Type a query: { field: 'value' }

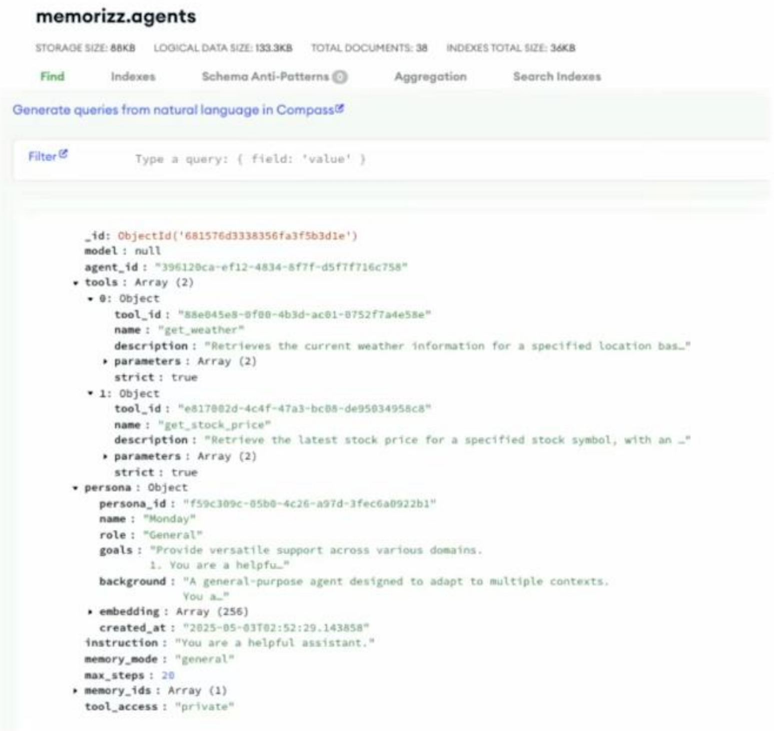
QUERY RESULTS: 1-3 OF 3

```
_id: ObjectId('68233192f98a0c6ee1b7ba52')
content : "
          Acme Corporation is a fictional company that manufactures everything …"
▸ embedding : Array (256)
namespace : "company_info"
long_term_memory_id : "b3be3a5c-0e0f-49f8-b36a-152d97ce8482"
created_at : "2025-05-13T12:48:34.060514"
updated_at : "2025-05-13T12:48:34.060543"
```

```
_id: ObjectId('68233193f98a0c6ee1b7ba53')
content : "
          Acme's Portable Hole is a revolutionary product that creates a tempor…"
▸ embedding : Array (256)
namespace : "product_info"
long_term_memory_id : "778b2b7c-6045-4b52-a4f8-8fbe2fe0d1f3"
created_at : "2025-05-13T12:48:35.057895"
updated_at : "2025-05-13T12:48:35.057904"
```

Agent Registry

→ **Description**: A store for storing facts, information and associated data with entities(humans, other agents, software, APIs) an agent interacts with during its execution.



memorizz.agents

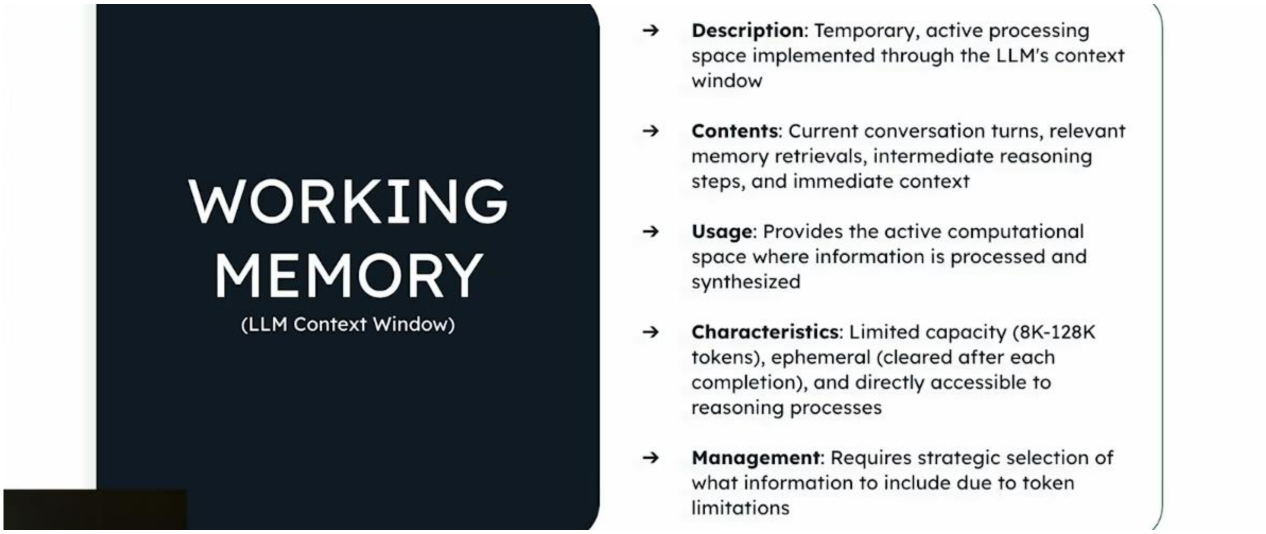
STORAGE SIZE: 88KB LOGICAL DATA SIZE: 133.3KB TOTAL DOCUMENTS: 38 INDEXES TOTAL SIZE: 36KB

Find
Indexes
Schema Anti-Patterns 0
Aggregation
Search Indexes
Generate queries from natural language in Compass
Filter
Type a query: { field: 'value' }

```
_id: ObjectId('681576d3338356fa3f5b3d1e')
model : null
agent_id : "396120ca-ef12-4834-8f7f-d5f7f716c758"
▾ tools : Array (2)
  ▾ 0: Object
      tool_id : "88e045e8-0f00-4b3d-ac01-0752f7a4e58e"
      name : "get_weather"
      description : "Retrieves the current weather information for a specified location bas…"
    ▸ parameters : Array (2)
      strict : true
  ▾ 1: Object
      tool_id : "e817002d-4c4f-47a3-bc08-de95034958c8"
      name : "get_stock_price"
      description : "Retrieve the latest stock price for a specified stock symbol, with an …"
    ▸ parameters : Array (2)
      strict : true
▾ persona : Object
    persona_id : "f59c309c-05b0-4c26-a97d-3fec6a0922b1"
    name : "Monday"
    role : "General"
    goals : "Provide versatile support across various domains.
           1. You are a helpfu…"
    background : "A general-purpose agent designed to adapt to multiple contexts.
                 You a…"
  ▸ embedding : Array (256)
    created_at : "2025-05-03T02:52:29.143858"
instruction : "You are a helpful assistant."
memory_mode : "general"
max_steps : 20
▸ memory_ids : Array (1)
tool_access : "private"
```

ENTITY MEMORY

→ **Description**: A store for storing facts, information and associated data with entities(humans, other agents, software, APIs) an agent interacts with during its execution.



WORKING MEMORY

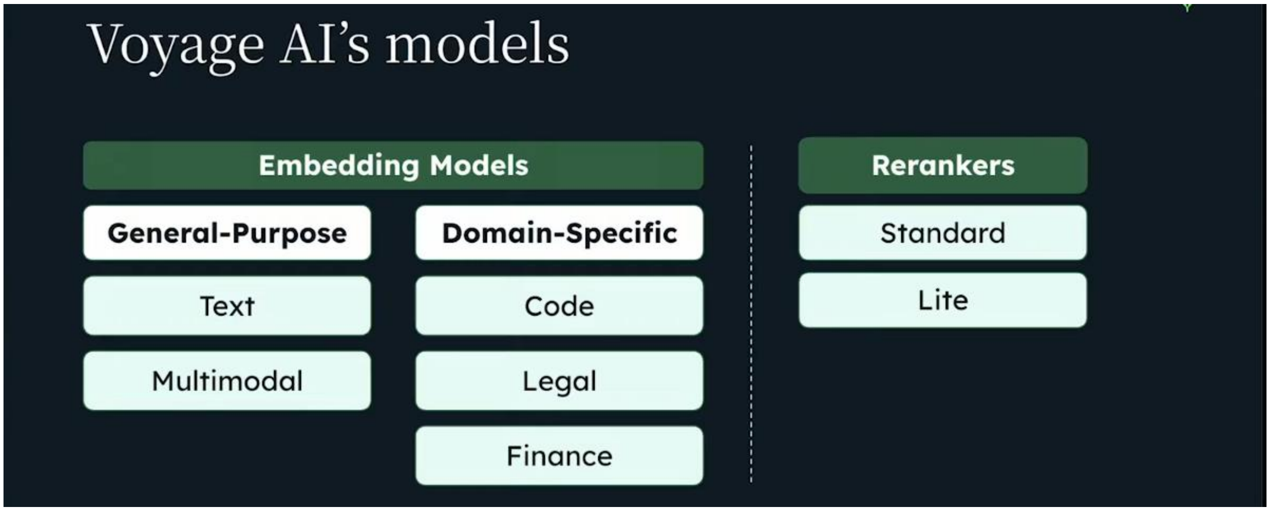
(LLM Context Window)

- → **Description**: Temporary, active processing space implemented through the LLM's context window
- → **Contents**: Current conversation turns, relevant memory retrievals, intermediate reasoning steps, and immediate context
- → **Usage**: Provides the active computational space where information is processed and synthesized
- → **Characteristics**: Limited capacity (8K-128K tokens), ephemeral (cleared after each completion), and directly accessible to reasoning processes
- → **Management**: Requires strategic selection of what information to include due to token limitations

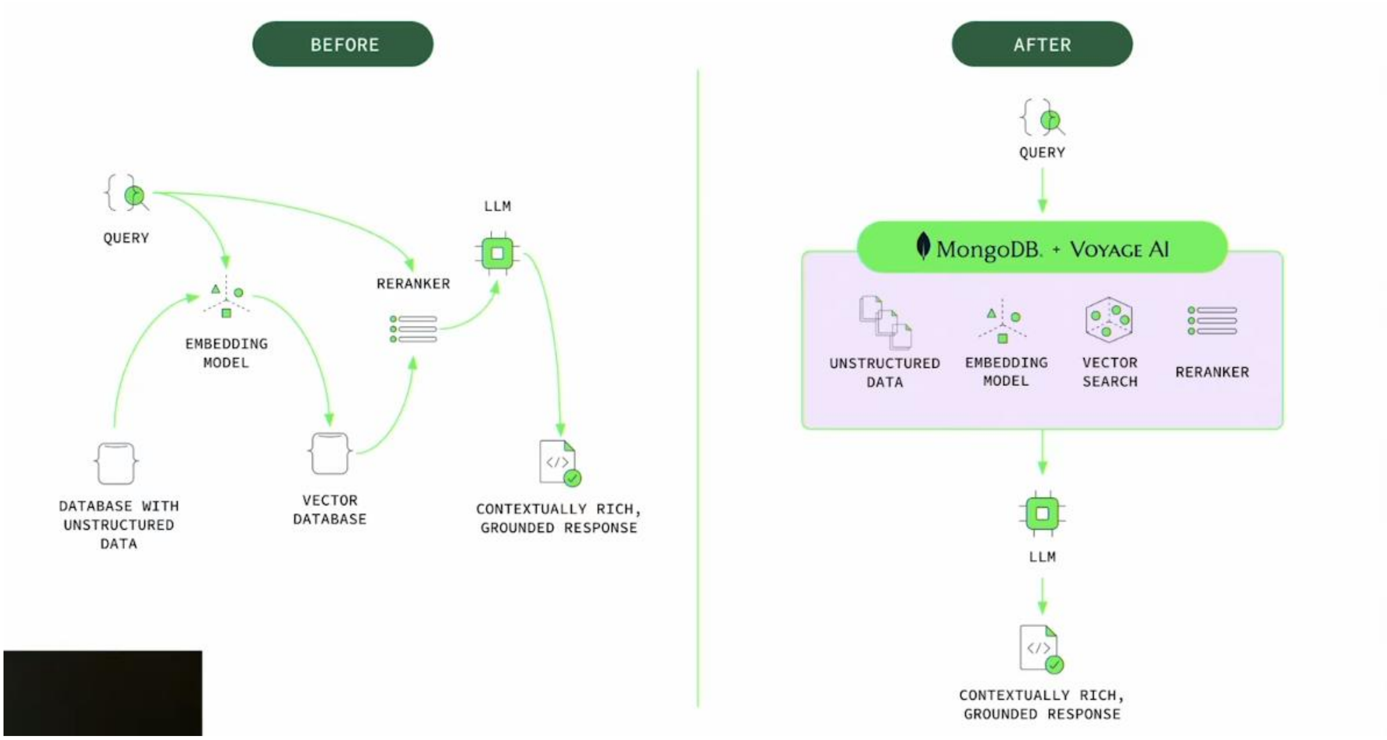
[redacted]

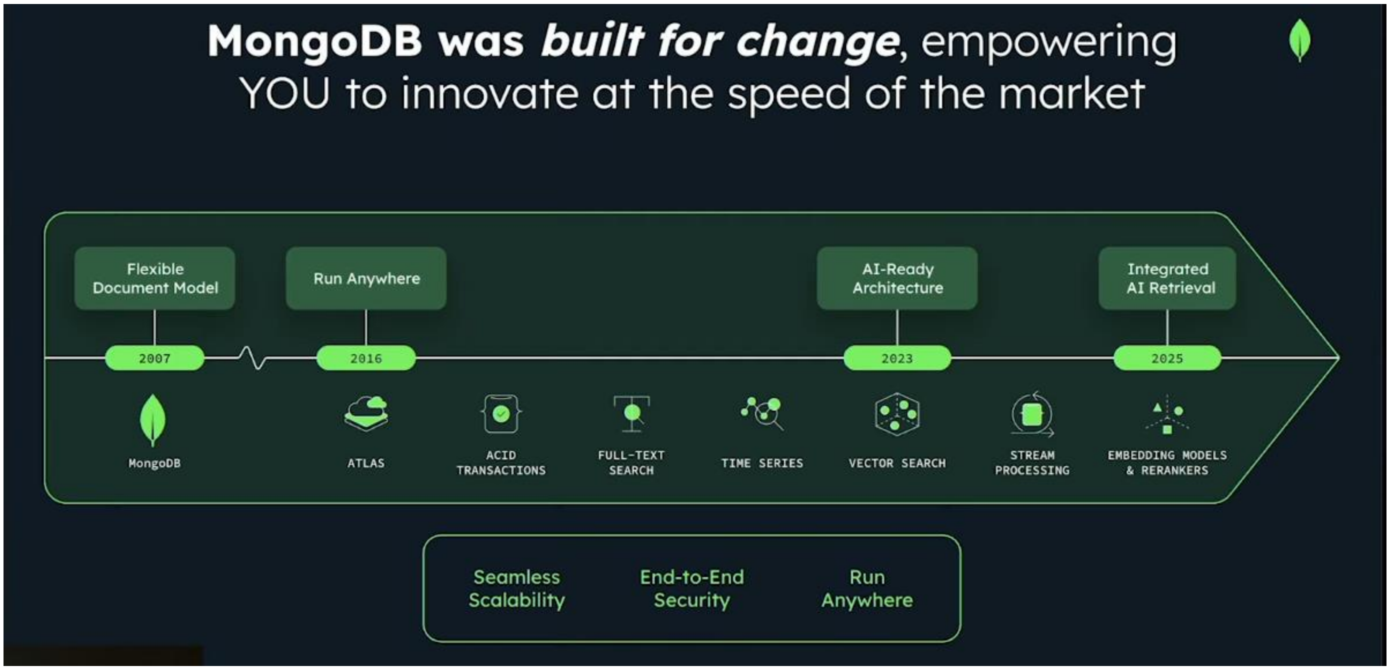
The Memory Provider For Agentic Systems: MongoDB

[redacted]



Voyage AI's models





MongoDB was ***built for change***, empowering YOU to innovate at the speed of the market

Seamless Scalability

End-to-End Security

Run Anywhere