# JOURNEY

# Memory Engineering

## The Discipline Behind Memory Augmented Agents

# Richmond Alake

Director, AI Developer Experience
Oracle Database

## Agent Memory and Memory Engineering

On a mission to make Memory Engineering a recognized discipline in AI. It's how I think about the science of helping agents remember, reason, and act — and it's what I build every day through MemoRizz and my #100DaysOfAgentMemory series.

# Richmond Alake

Director, AI Developer Experience
Oracle Database

**Writer | Educator | Speaker**

Written 200+ articles, clocked over a million views, and spoken at conferences around the world. My goal is simple — take the hardest ideas in AI and make them accessible to every developer.

# Richmond Alake

Director, AI Developer Experience
Oracle Database

## AI Developer

I don't just talk about this stuff. I build it. From open-source tools to supporting developers in large enterprise organization in developing and deploying AI systems, I bring an engineer's mindset to everything I do.

● **JOURNEY**

○ Ecosystem

○ Form Factor

○ Disciplines

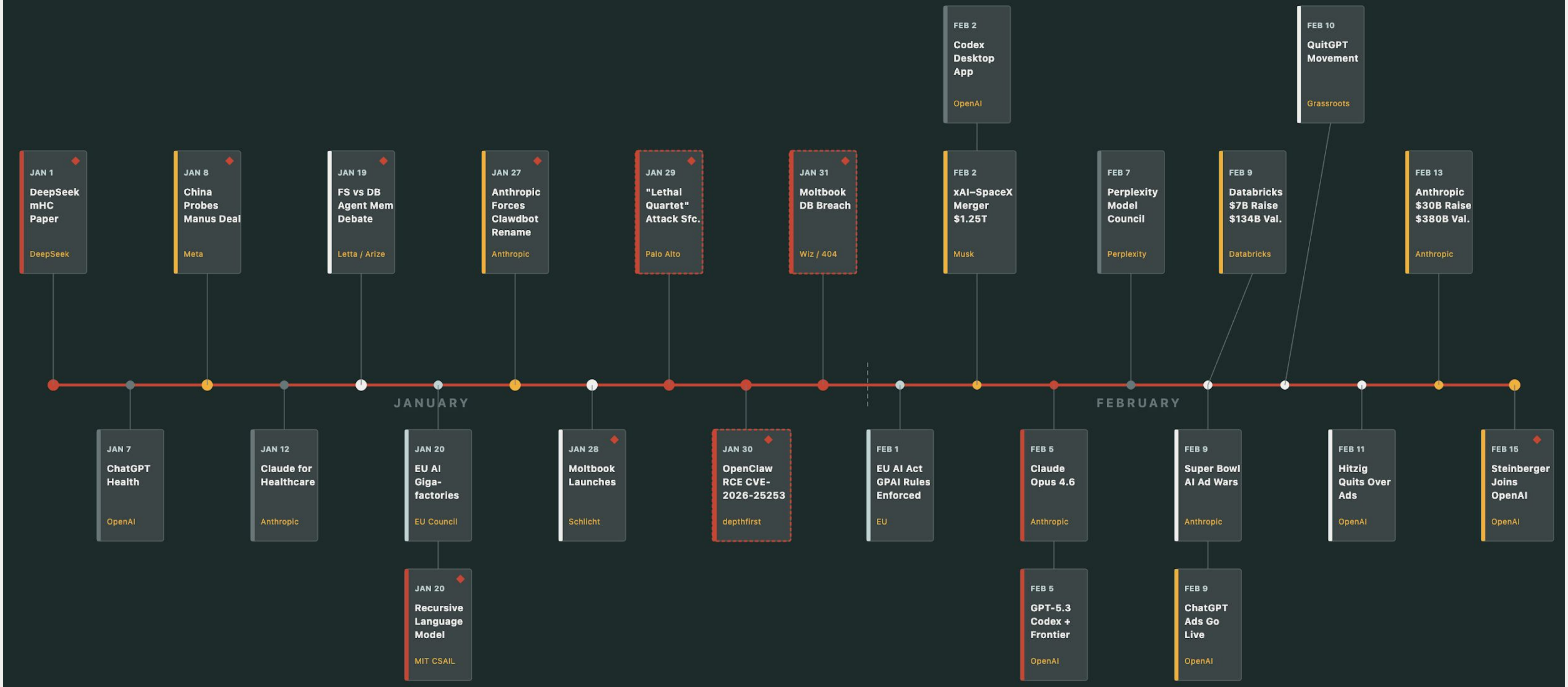
○ 
# JOURNEY

● Ecosystem

○ Form Factor

○ Disciplines

# AI Industry Timeline — January to February 2026



JANUARY

JAN 1 ◆ DeepSeek mHC Paper — DeepSeek

JAN 7 ChatGPT Health — OpenAI

JAN 8 ◆ China Probes Manus Deal — Meta

JAN 12 Claude for Healthcare — Anthropic

JAN 19 ◆ FS vs DB Agent Mem Debate — Letta / Arize

JAN 20 EU AI Giga-factories — EU Council

JAN 20 ◆ Recursive Language Model — MIT CSAIL

JAN 27 ◆ Anthropic Forces Clawdbot Rename — Anthropic

JAN 28 ◆ Moltbook Launches — Schlicht

JAN 29 ◆ "Lethal Quartet" Attack Sfc. — Palo Alto

JAN 30 ◆ OpenClaw RCE CVE-2026-25253 — depthfirst

JAN 31 ◆ Moltbook DB Breach — Wiz / 404

FEBRUARY

FEB 1 EU AI Act GPAI Rules Enforced — EU

FEB 2 Codex Desktop App — OpenAI

FEB 2 xAI–SpaceX Merger $1.25T — Musk

FEB 5 Claude Opus 4.6 — Anthropic

FEB 5 GPT-5.3 Codex + Frontier — OpenAI

FEB 7 Perplexity Model Council — Perplexity

FEB 9 Databricks $7B Raise $134B Val. — Databricks

FEB 9 Super Bowl AI Ad Wars — Anthropic

FEB 9 ChatGPT Ads Go Live — OpenAI

FEB 10 QuitGPT Movement — Grassroots

FEB 11 Hitzig Quits Over Ads — OpenAI

FEB 13 Anthropic $30B Raise $380B Val. — Anthropic

FEB 15 ◆ Steinberger Joins OpenAI — OpenAI

🟥 Model ◼ Product 🟨 Business ⬚ Security ◻ Policy ⬜ Culture ◆ Agent Memory Relevance

Data as of Feb 16, 2026 | Richmond Alake

# IT IS ONLY FEBRUARY

*The speed of development of AI this past few years is just breathtaking. I know everyone, whether we pretend or not, everyone deep in your heart is feeling that anxiety of there's just too much to read, too many blogs, too many news, too many model releases, too many and and that sense of anxiety is is speaks of our time is that this technology is just moving as at a breathtaking speed.*

*So that's that gives me a lot of excitement but also keeps me very grounded about how little I know. You know this famous saying I don't know anything and someone like me even feels that and I want you to at least hear this from me and recognize we all feel that and don't give up your learning and continue to be curious*

Fei-Fei Li

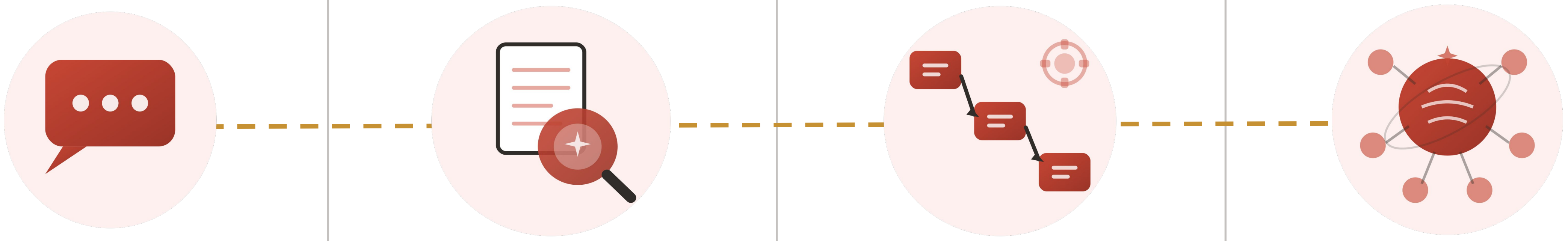
# JOURNEY

○ Ecosystem

● Form Factor

○ Disciplines

# Form Factors and Maturity Levels



## LLM Chatbots

Single-turn or multi-turn **conversational interfaces** powered by large language models. Minimal context retention, best suited for Q&A, summarization, and general assistance

## RAG Applications

LLM **responses grounded** in your organization's data through retrieval-augmented generation. Connects models to internal knowledge bases for accurate, domain-specific answers.

## LLM-Driven Workflows

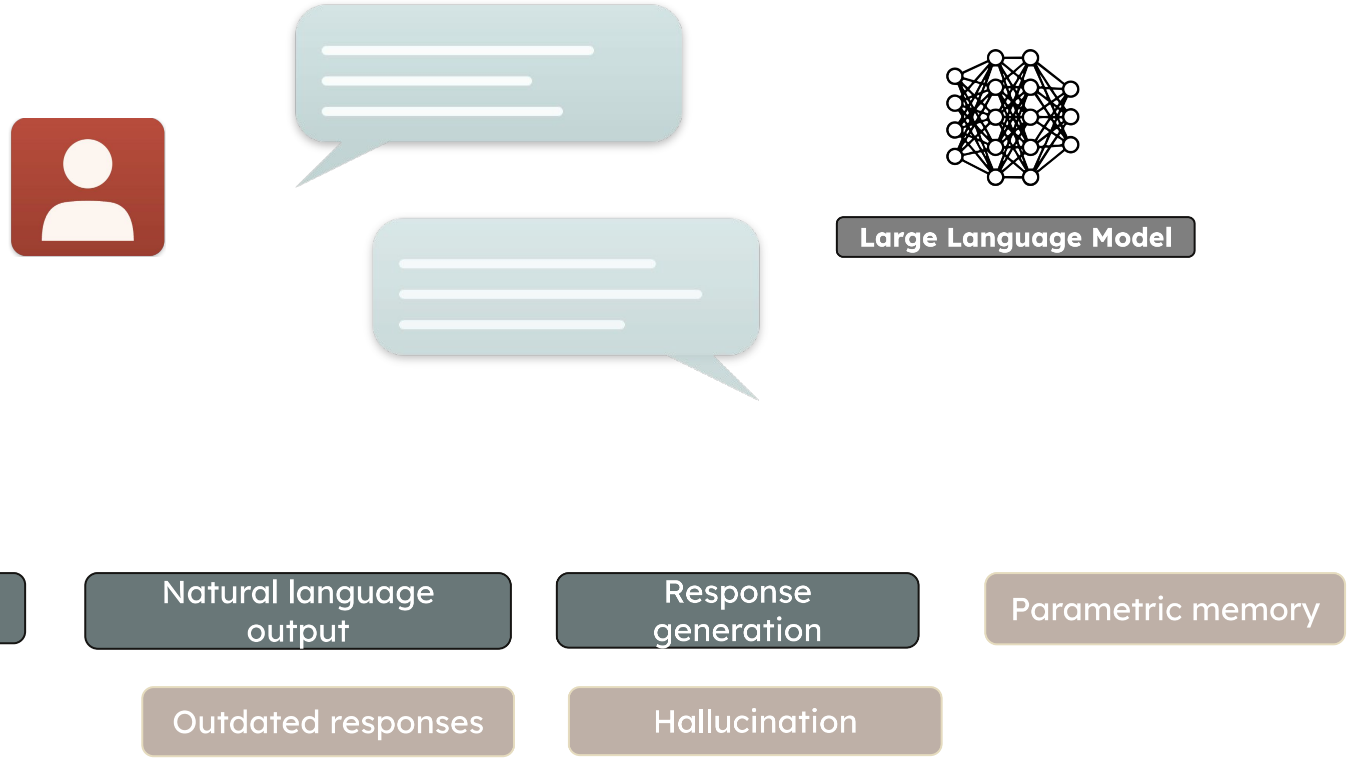
Multi-step processes orchestrated by LLMs within human-defined logic. **Automates** structured tasks like document processing, approvals, and data transformation with predictable outputs.

## Agentic AI

**Autonomous** systems that plan, reason, use tools, and make decisions with minimal human intervention. Maintain persistent memory and adapt dynamically to accomplish complex goals

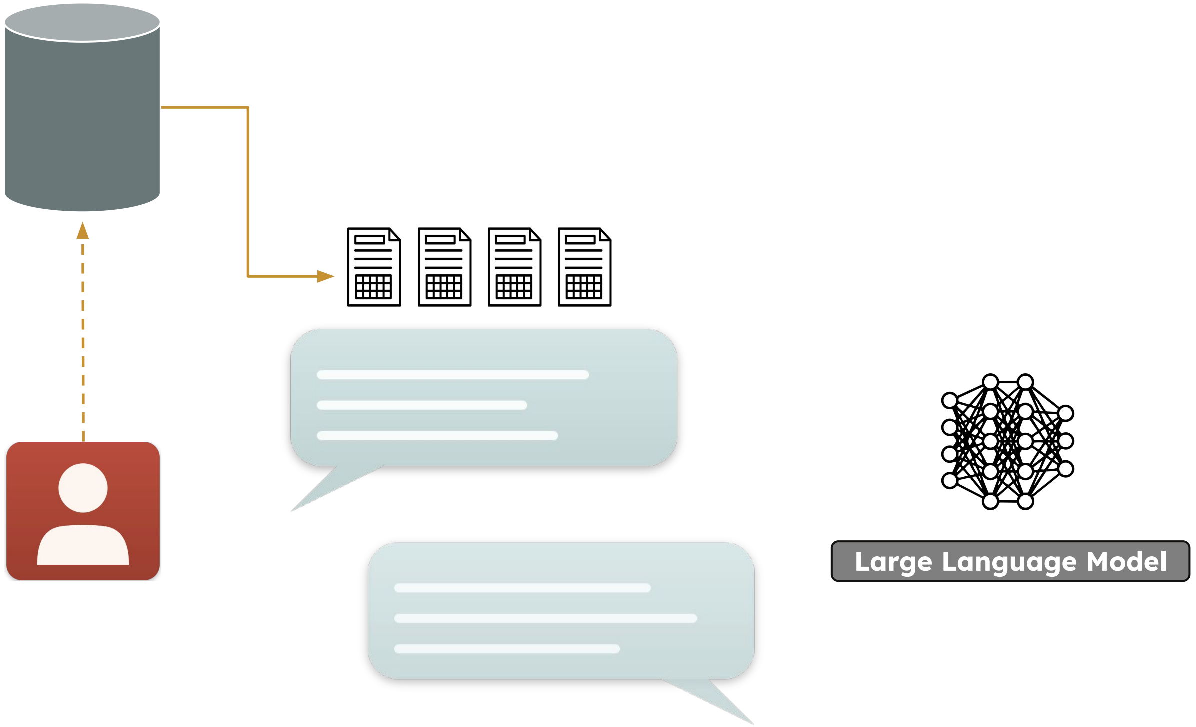
## LLM Chatbots

Single-turn or multi-turn **conversational interfaces** powered by large language models. Minimal context retention, best suited for Q&A, summarization, and general assistance



## RAG Applications

LLM **responses grounded** in your organization's data through retrieval-augmented generation. Connects models to internal knowledge bases for accurate, domain-specific answers.



Retrieval-Augmented Generation (RAG)

Grounded responses

Knowledge bases

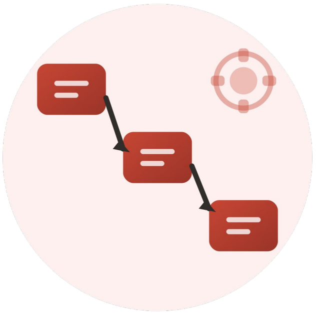
Domain-specific

Knowledge Bases

Single-turn only

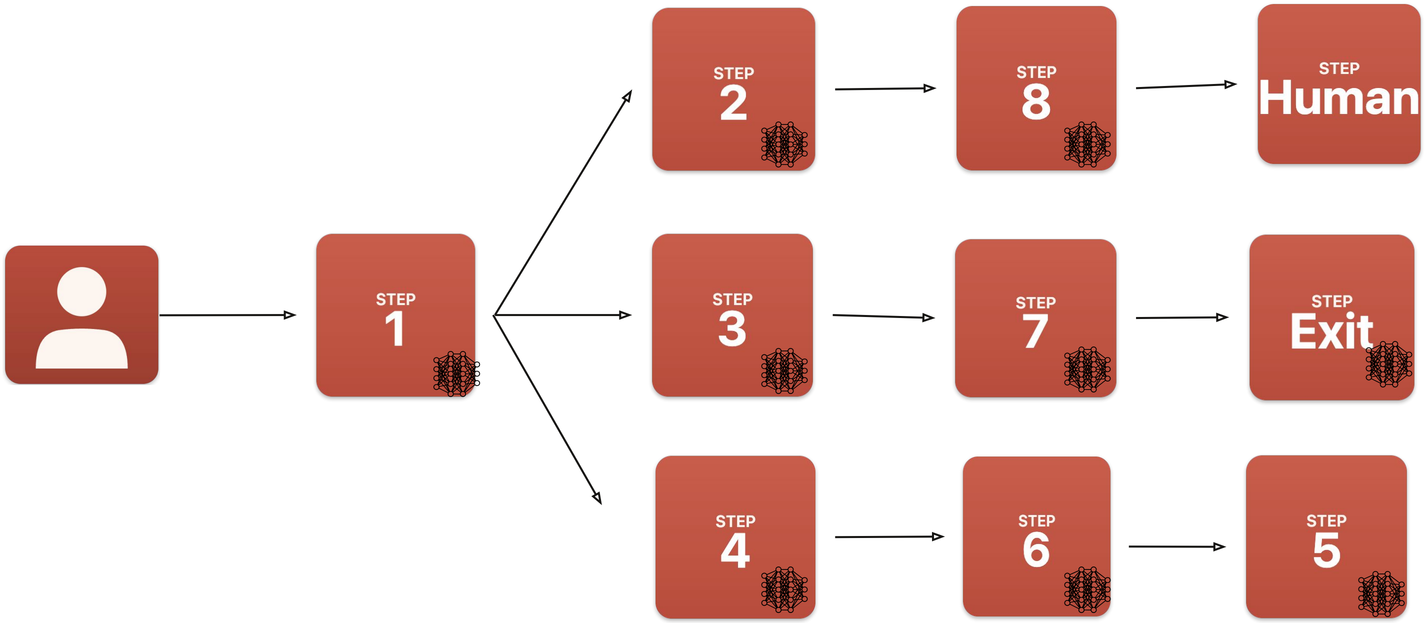
No Decision Flow

Stateless



## LLM-Driven Workflows

Multi-step processes orchestrated by LLMs within human-defined logic. **Automates** structured tasks like document processing, approvals, and data transformation with predictable outputs.



Multi-step processes

Orchestrated

Human-defined logic

Domain-specific

Knowledge Bases

Reactive only

Rigid/Inflexible

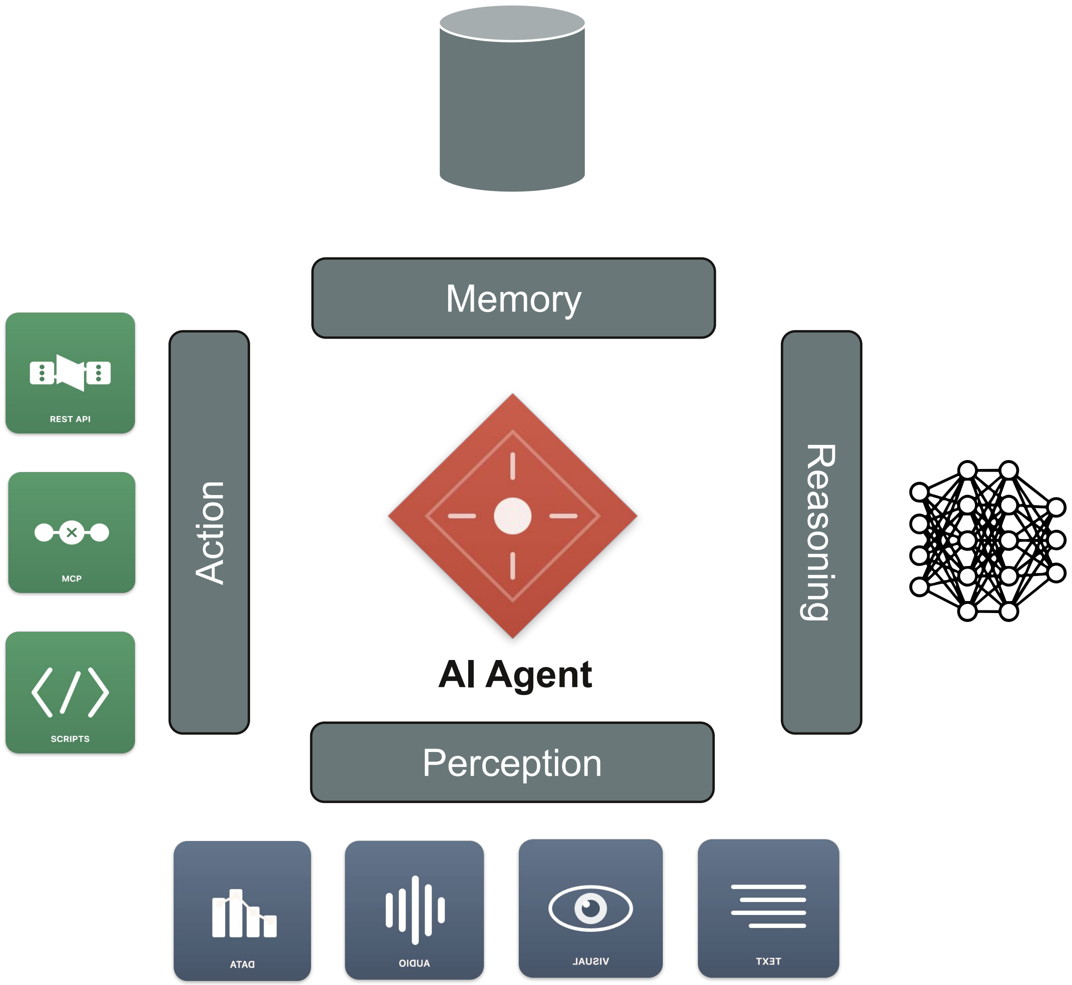
Predictable outputs

No autonomy

No learning

## Agentic AI

**Autonomous** systems that plan, reason, use tools, and make decisions with minimal human intervention. Maintain persistent memory and adapt dynamically to accomplish complex goals



# JOURNEY

○ Ecosystem

○ Form Factor

● Disciplines

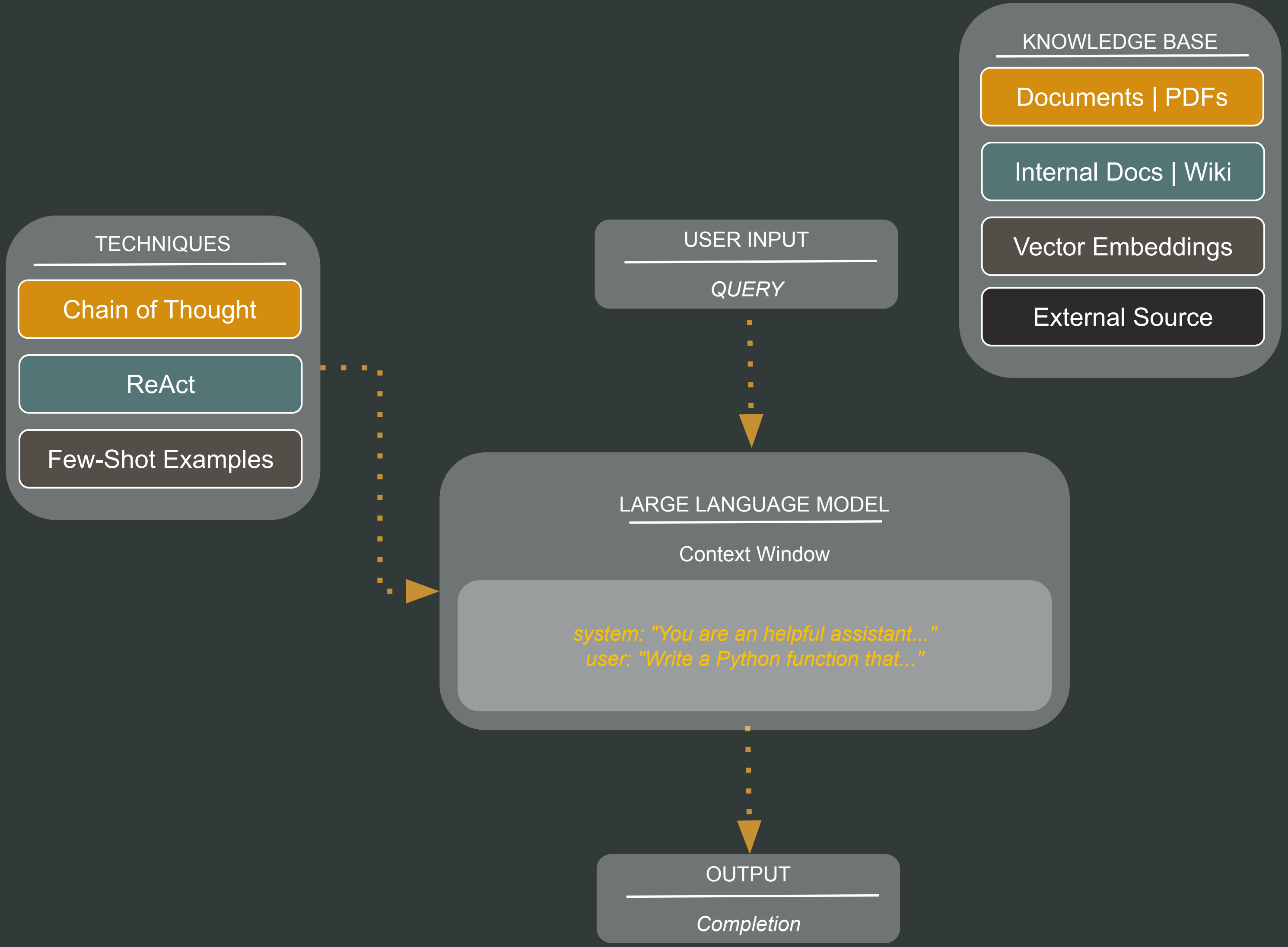
Prompt Engineering

Context Engineering

Memory Engineering ?

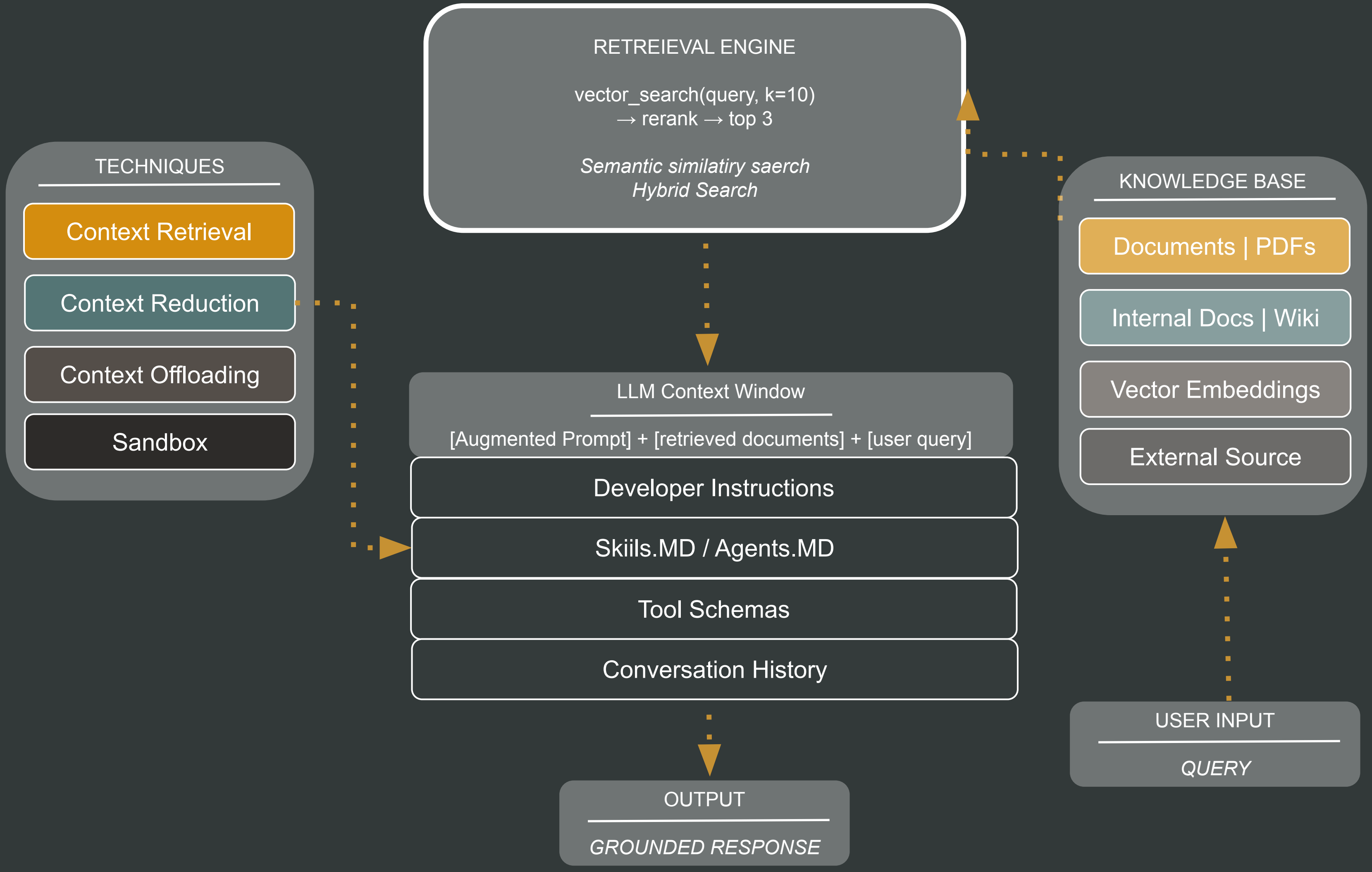
# Prompt Engineering

The utilization of **linguistic patterns and language modification to maximize** the input provided into an LLM in order to illicit a desired behavior and output



# Context Engineering

The practice of selecting, structuring, and prioritizing tokens within a finite context window, optimizing the signal-to-noise ratio of inputs to reliably produce desired model behavior.



RETREIEVAL ENGINE

vector_search(query, k=10)
→ rerank → top 3

*Semantic similatiry saerch*
*Hybrid Search*

TECHNIQUES

- Context Retrieval
- Context Reduction
- Context Offloading
- Sandbox

KNOWLEDGE BASE

- Documents | PDFs
- Internal Docs | Wiki
- Vector Embeddings
- External Source

LLM Context Window

[Augmented Prompt] + [retrieved documents] + [user query]

- Developer Instructions
- Skiils.MD / Agents.MD
- Tool Schemas
- Conversation History

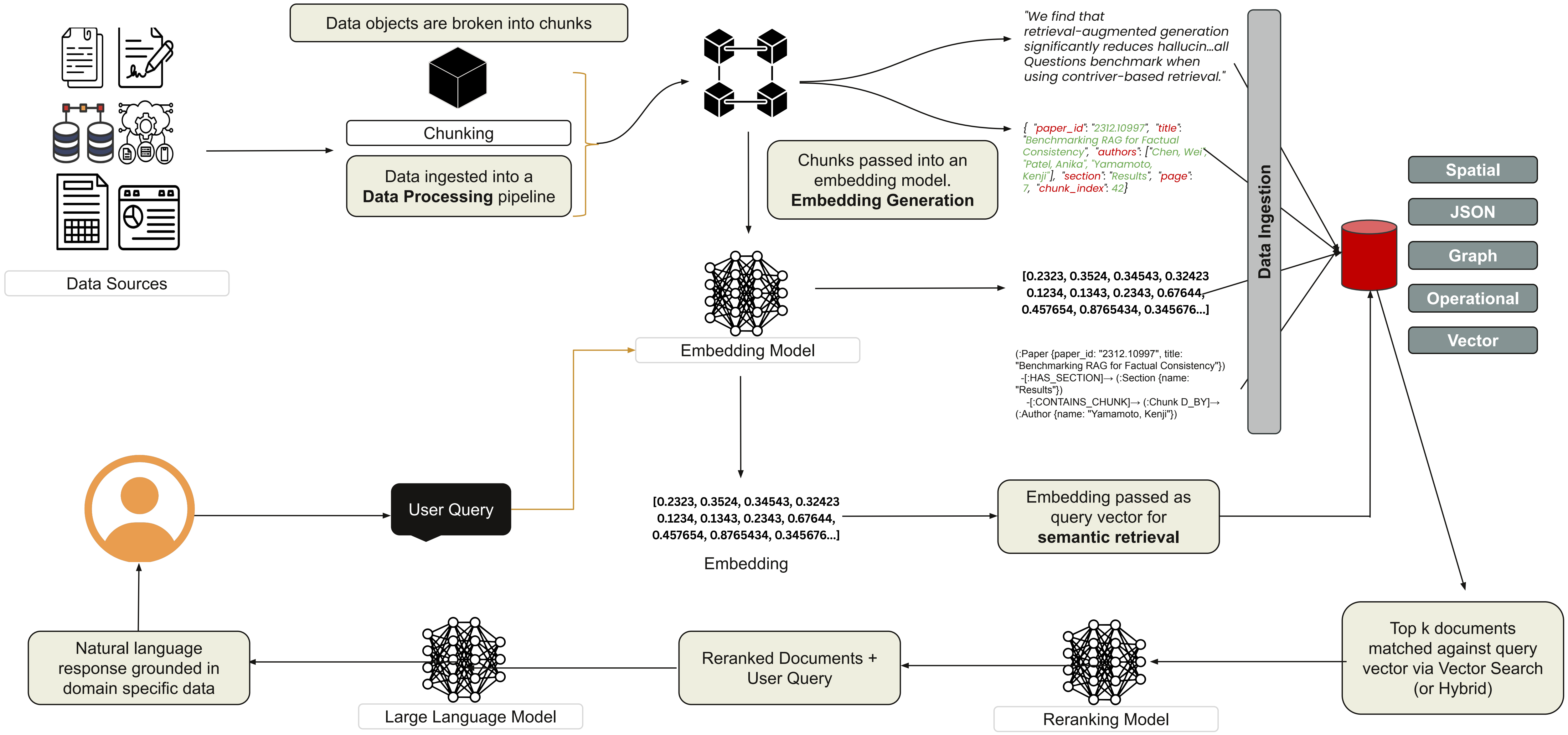
USER INPUT

*QUERY*

OUTPUT

*GROUNDED RESPONSE*

# Agent Memory



Data Sources

Data objects are broken into chunks

Chunking

Data ingested into a **Data Processing** pipeline

Chunks passed into an embedding model. **Embedding Generation**

Embedding Model

*"We find that retrieval-augmented generation significantly reduces hallucin...all Questions benchmark when using contriver-based retrieval."*

```
{ "paper_id": "2312.10997", "title": "Benchmarking RAG for Factual Consistency", "authors": ["Chen, Wei", "Patel, Anika", "Yamamoto, Kenji"], "section": "Results", "page": 7, "chunk_index": 42}
```

**[0.2323, 0.3524, 0.34543, 0.32423 0.1234, 0.1343, 0.2343, 0.67644, 0.457654, 0.8765434, 0.345676...]**

```
(:Paper {paper_id: "2312.10997", title:
"Benchmarking RAG for Factual Consistency"})
  -[:HAS_SECTION]→ (:Section {name:
"Results"})
    -[:CONTAINS_CHUNK]→ (:Chunk D_BY]→
(:Author {name: "Yamamoto, Kenji"})
```

**Data Ingestion**

**Spatial**

**JSON**

**Graph**

**Operational**

**Vector**

User Query

**[0.2323, 0.3524, 0.34543, 0.32423 0.1234, 0.1343, 0.2343, 0.67644, 0.457654, 0.8765434, 0.345676...]**

Embedding

Embedding passed as query vector for **semantic retrieval**

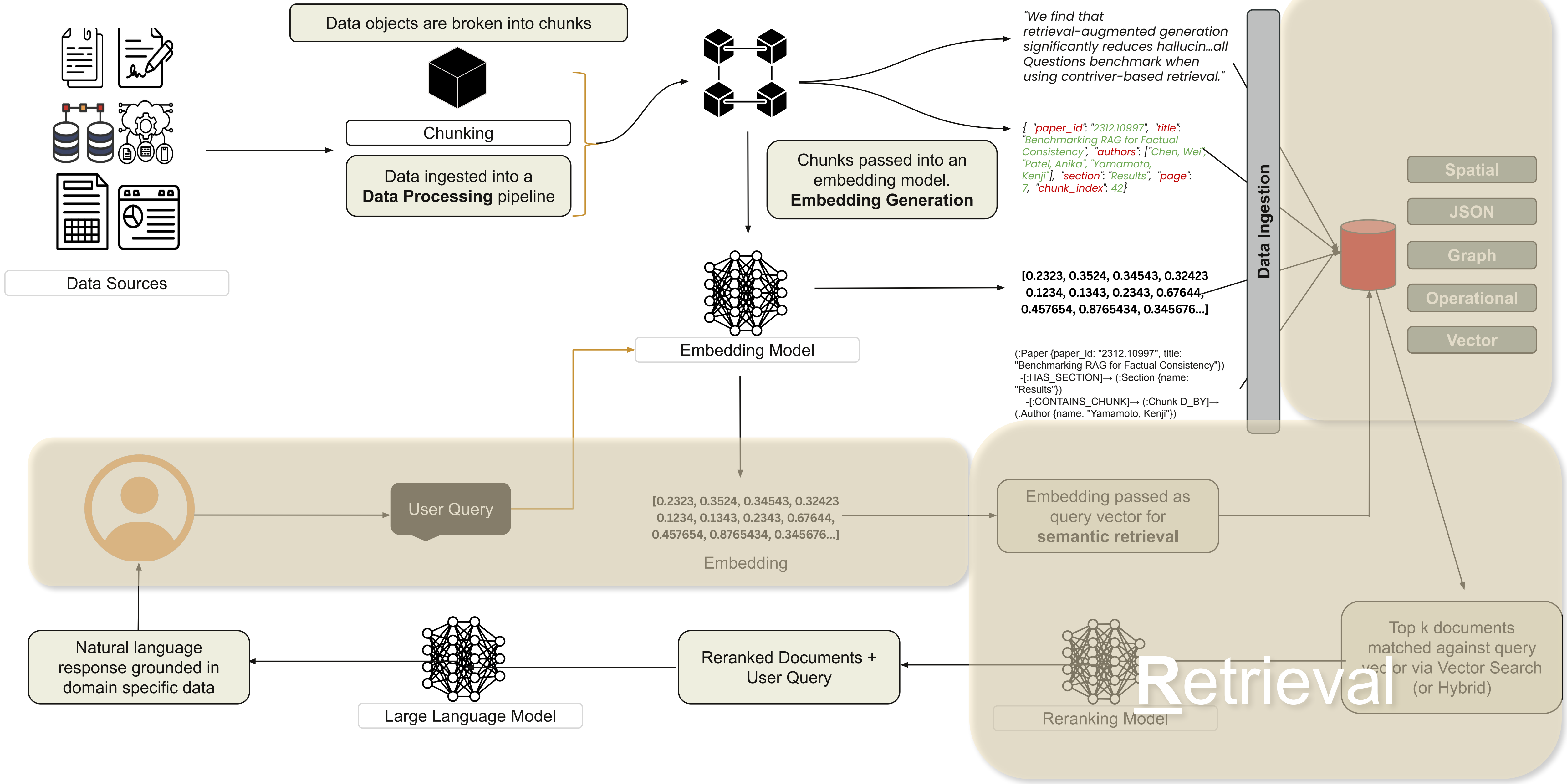
Top k documents matched against query vector via Vector Search (or Hybrid)

Reranking Model

Reranked Documents + User Query

Large Language Model

Natural language response grounded in domain specific data



Data objects are broken into chunks

Chunking

Data ingested into a **Data Processing** pipeline

Data Sources

Chunks passed into an embedding model. **Embedding Generation**

Embedding Model

*"We find that retrieval-augmented generation significantly reduces hallucin…all Questions benchmark when using contriver-based retrieval."*

```
{  "paper_id": "2312.10997",  "title": "Benchmarking RAG for Factual Consistency",  "authors": ["Chen, Wei", "Patel, Anika", "Yamamoto, Kenji"],  "section": "Results",  "page": 7,  "chunk_index": 42}
```

**[0.2323, 0.3524, 0.34543, 0.32423 0.1234, 0.1343, 0.2343, 0.67644, 0.457654, 0.8765434, 0.345676...]**

```
(:Paper {paper_id: "2312.10997", title:
"Benchmarking RAG for Factual Consistency"})
  -[:HAS_SECTION]→ (:Section {name:
"Results"})
    -[:CONTAINS_CHUNK]→ (:Chunk D_BY]→
(:Author {name: "Yamamoto, Kenji"})
```

Data Ingestion

Spatial

JSON

Graph

Operational

Vector

User Query

[0.2323, 0.3524, 0.34543, 0.32423 0.1234, 0.1343, 0.2343, 0.67644, 0.457654, 0.8765434, 0.345676...]

Embedding

Embedding passed as query vector for **semantic retrieval**

Top k documents matched against query vector via Vector Search (or Hybrid)

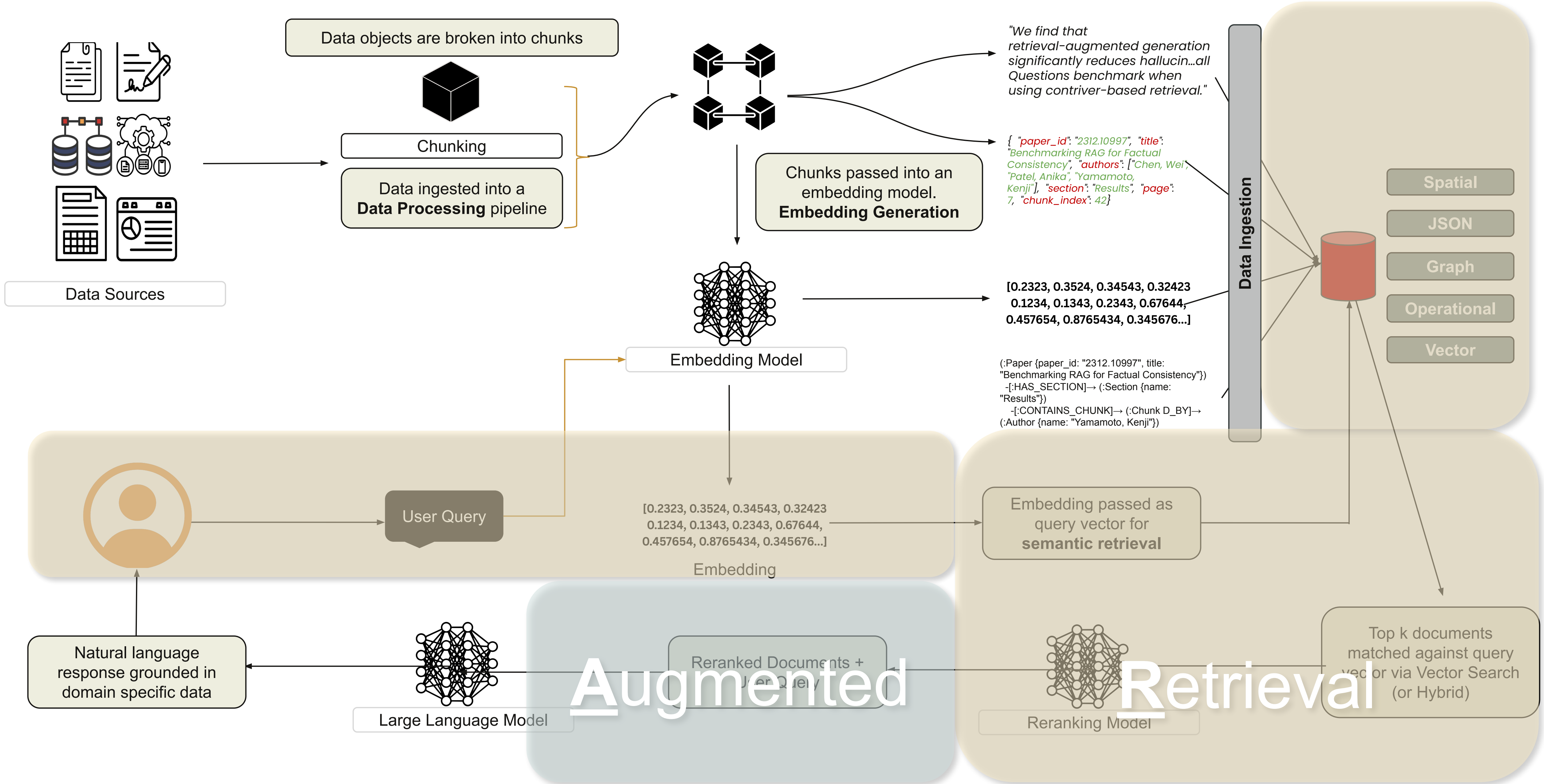
Reranking Model

Retrieval

Reranked Documents + User Query

Large Language Model

Natural language response grounded in domain specific data



Data Sources

Data objects are broken into chunks

Chunking

Data ingested into a **Data Processing** pipeline

Chunks passed into an embedding model. **Embedding Generation**

Embedding Model

*"We find that retrieval-augmented generation significantly reduces hallucin…all Questions benchmark when using contriver-based retrieval."*

```
{ "paper_id": "2312.10997", "title": "Benchmarking RAG for Factual Consistency", "authors": ["Chen, Wei", "Patel, Anika", "Yamamoto, Kenji"], "section": "Results", "page": 7, "chunk_index": 42}
```

**[0.2323, 0.3524, 0.34543, 0.32423 0.1234, 0.1343, 0.2343, 0.67644, 0.457654, 0.8765434, 0.345676...]**

```
(:Paper {paper_id: "2312.10997", title:
"Benchmarking RAG for Factual Consistency"})
  -[:HAS_SECTION]→ (:Section {name:
"Results"})
    -[:CONTAINS_CHUNK]→ (:Chunk D_BY]→
(:Author {name: "Yamamoto, Kenji"})
```

**Data Ingestion**

Spatial

JSON

Graph

Operational

Vector

User Query

[0.2323, 0.3524, 0.34543, 0.32423 0.1234, 0.1343, 0.2343, 0.67644, 0.457654, 0.8765434, 0.345676...]

Embedding

Embedding passed as query vector for **semantic retrieval**

Top k documents matched against query vector via Vector Search (or Hybrid)

Reranking Model

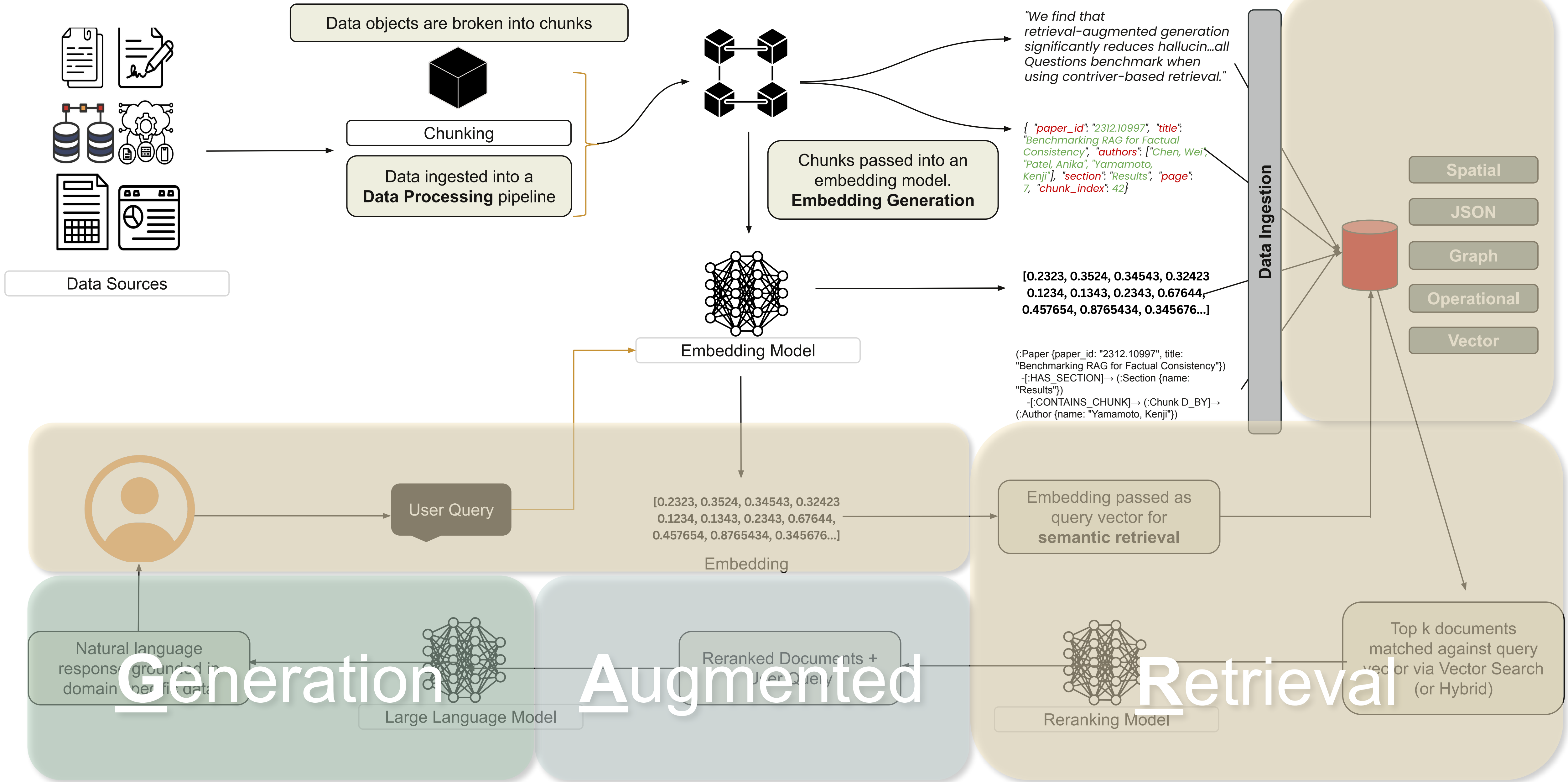
Retrieval

Reranked Documents + User Query

Augmented

Large Language Model

Natural language response grounded in domain specific data



Data Sources

Data objects are broken into chunks

Chunking

Data ingested into a **Data Processing** pipeline

Chunks passed into an embedding model. **Embedding Generation**

Embedding Model

*"We find that retrieval-augmented generation significantly reduces hallucin…all Questions benchmark when using contriver-based retrieval."*

```
{ "paper_id": "2312.10997", "title": "Benchmarking RAG for Factual Consistency", "authors": ["Chen, Wei", "Patel, Anika", "Yamamoto, Kenji"], "section": "Results", "page": 7, "chunk_index": 42}
```

**[0.2323, 0.3524, 0.34543, 0.32423 0.1234, 0.1343, 0.2343, 0.67644, 0.457654, 0.8765434, 0.345676...]**

```
(:Paper {paper_id: "2312.10997", title:
"Benchmarking RAG for Factual Consistency"})
  -[:HAS_SECTION]→ (:Section {name:
"Results"})
    -[:CONTAINS_CHUNK]→ (:Chunk D_BY]→
(:Author {name: "Yamamoto, Kenji"})
```

**Data Ingestion**

Spatial

JSON

Graph

Operational

Vector

User Query

[0.2323, 0.3524, 0.34543, 0.32423 0.1234, 0.1343, 0.2343, 0.67644, 0.457654, 0.8765434, 0.345676...]

Embedding

Embedding passed as query vector for **semantic retrieval**

Top k documents matched against query vector via Vector Search (or Hybrid)

Reranking Model

Reranked Documents + User Query

Large Language Model

Natural language response grounded in domain specific data

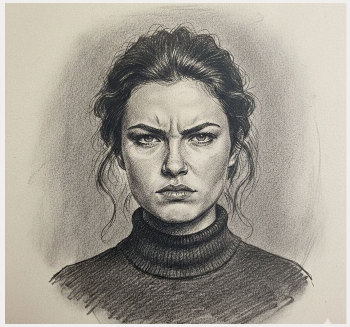
Generation

Augmented

Retrieval

### David Hubel and Torsten Wiesel

Neurophysiologists who won the 1981 Nobel Prize for mapping the visual cortex, discovering how neurons process images. They identified simple/complex cells that detect visual edges



### System 1

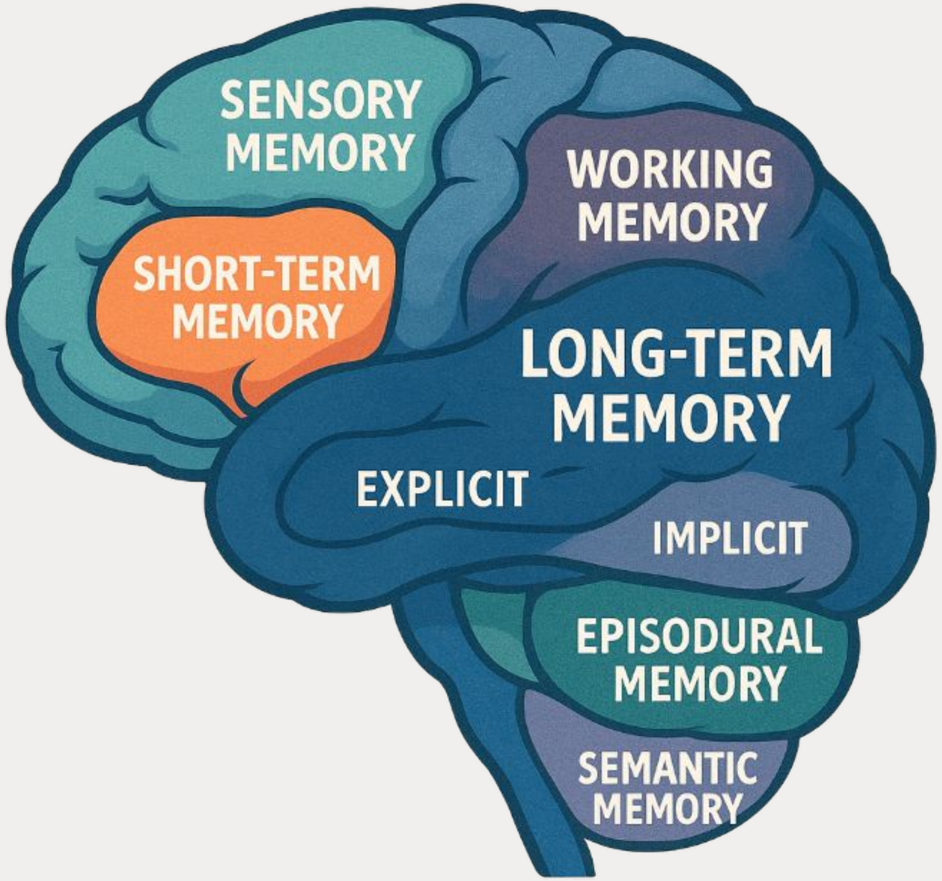
Operates quickly, effortlessly, and without conscious deliberation. It handles things like recognizing faces, understanding simple sentences, driving on an empty road, or catching a ball. It relies on heuristics, pattern recognition, and intuition.

### System 2

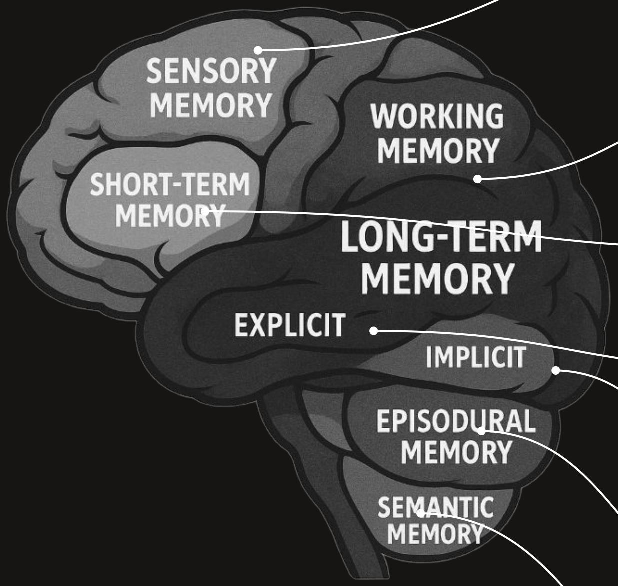
Slow, deliberate, and analytical. It's what you engage when solving a complex math problem, comparing two products with different features, or carefully planning a project.

### Lilienthal's Glider in Flight

Otto Lilienthal's work on flight was deeply inspired by birds. He studied their wing mechanics and used their mechanisms of flight as the foundation for building controllable flying vehicles. His book, Birdflight as the Basis of Aviation, went on to become a key influence on the Wright brothers' own pioneering work.


SENSORY MEMORY
SHORT-TERM MEMORY
WORKING MEMORY
LONG-TERM MEMORY
EXPLICIT
IMPLICIT
EPISODURAL MEMORY
SEMANTIC MEMORY

Humans have a good ability to **retain**, **recall** and **reuse** information over **short and long period** of time.



The brief after image you see when you look away from a bright light

Remembering a phone number just long enough to dial it

Where you parked your car at the mall today

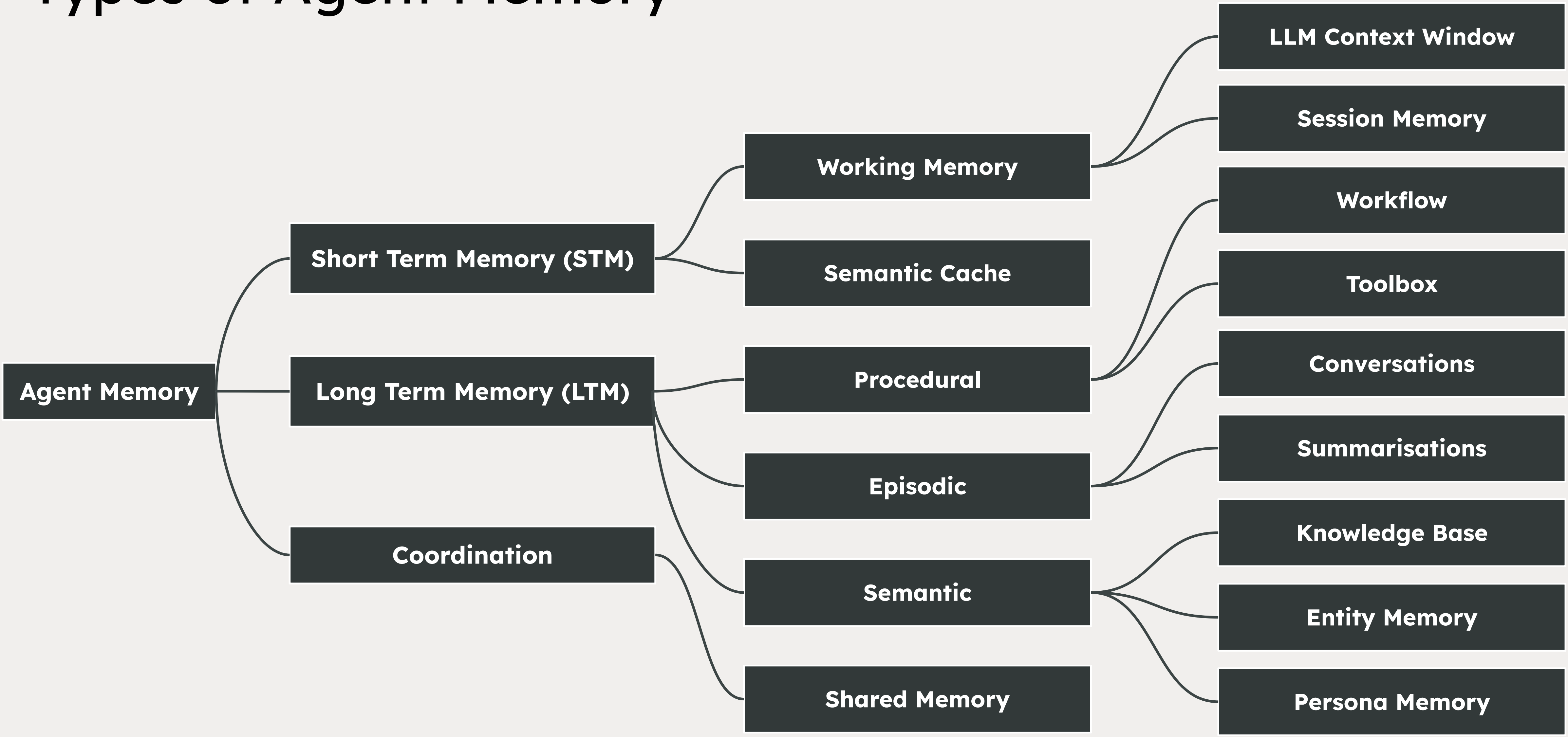
Remembering your wedding day

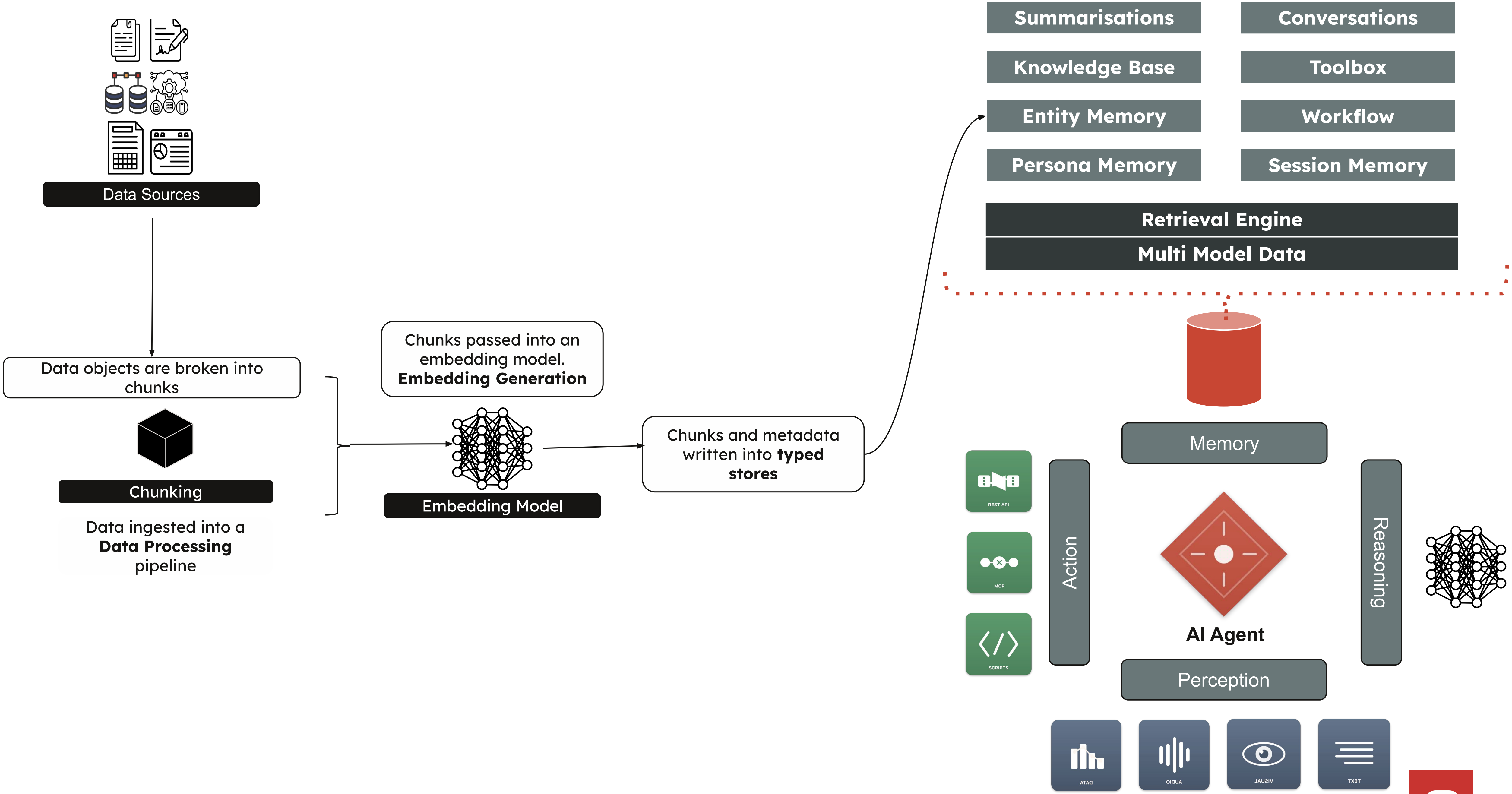
Riding a bike without thinking about balance

Your first day of school

Knowing that dogs are animals

# Types of Agent Memory





Data Sources

Data objects are broken into chunks

Chunking

Data ingested into a **Data Processing** pipeline

Chunks passed into an embedding model. **Embedding Generation**

Embedding Model

Chunks and metadata written into **typed stores**

Summarisations

Conversations

Knowledge Base

Toolbox

Entity Memory

Workflow

Persona Memory

Session Memory

Retrieval Engine

Multi Model Data

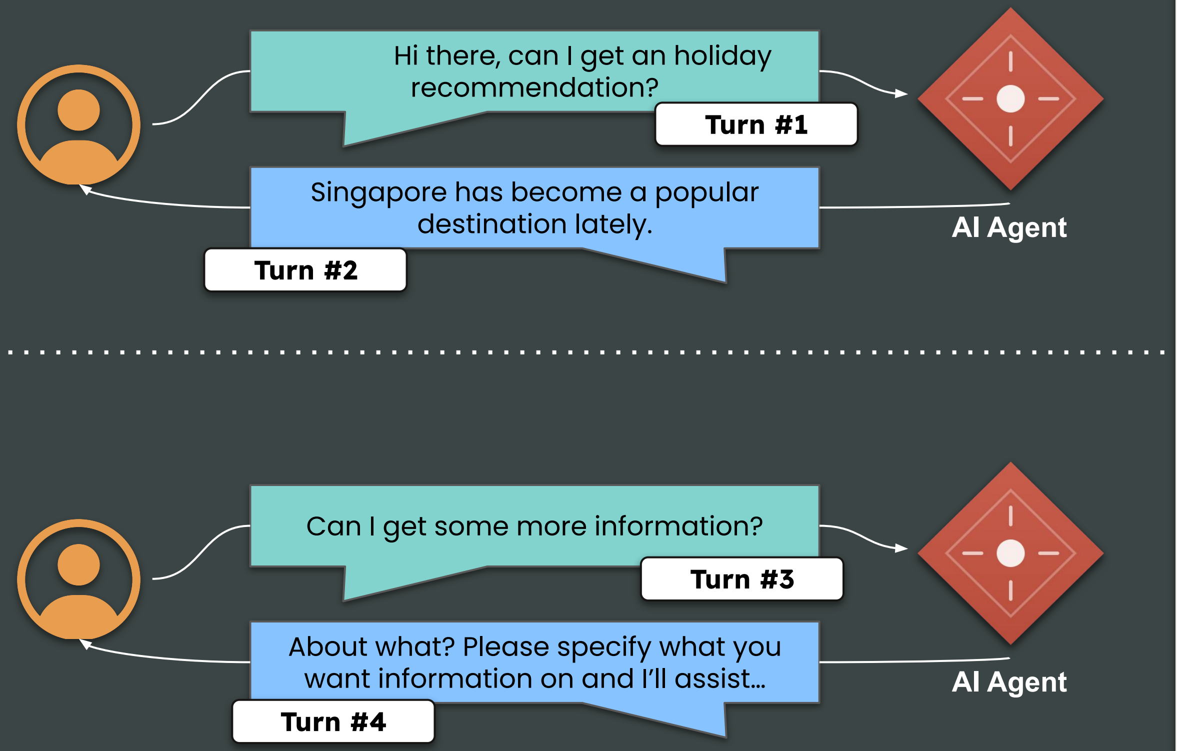
Memory

Action

Reasoning

AI Agent

Perception



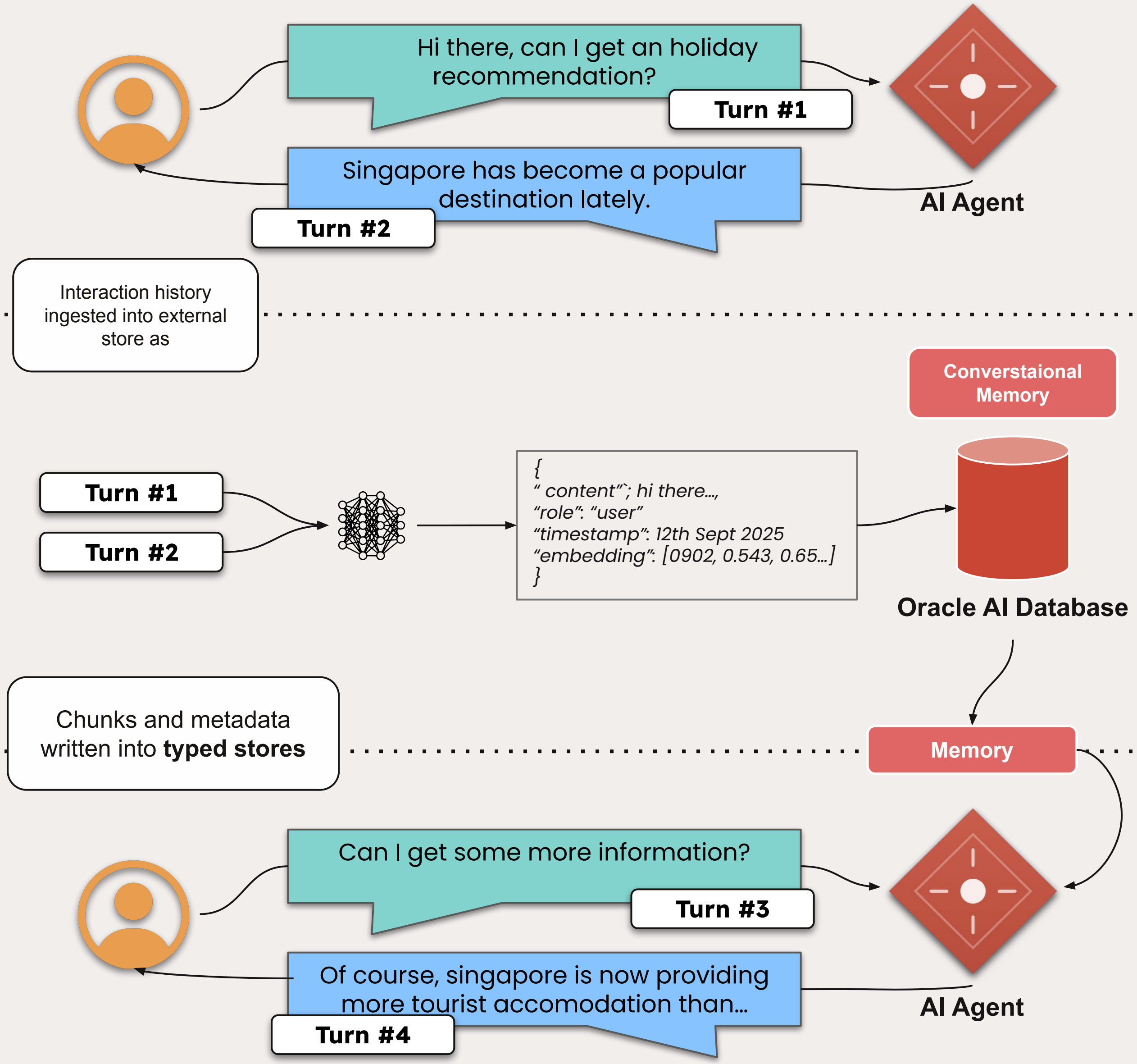
Hi there, can I get an holiday recommendation?

Singapore has become a popular destination lately.

Can I get some more information?

About what? Please specify what you want information on and I'll assist...

**Stateless Agent**



Hi there, can I get an holiday recommendation?

Singapore has become a popular destination lately.

Interaction history ingested into external store as

```
{
" content"; hi there...,
"role": "user"
"timestamp": 12th Sept 2025
"embedding": [0902, 0.543, 0.65...]
}
```

Chunks and metadata written into **typed stores**

Can I get some more information?

Of course, singapore is now providing more tourist accomodation than...

**Memory Augmented Agent**

# Agent Memory

A computational exocortex for AI agents made up of a dynamic, systematic process that integrates an agent's LLM memory (context window and parametric weights) with a persistent external memory management system.

This process relies on three core components working in concert: the LLM for reasoning and synthesis, an embedding model for encoding information into vector representations that enable semantic retrieval, and a database for persistent storage and retrieval. Together, these enable an agent to accumulate knowledge, maintain continuity across interactions, and adapt its behaviour based on historical patterns.

# Agent Memory Signals

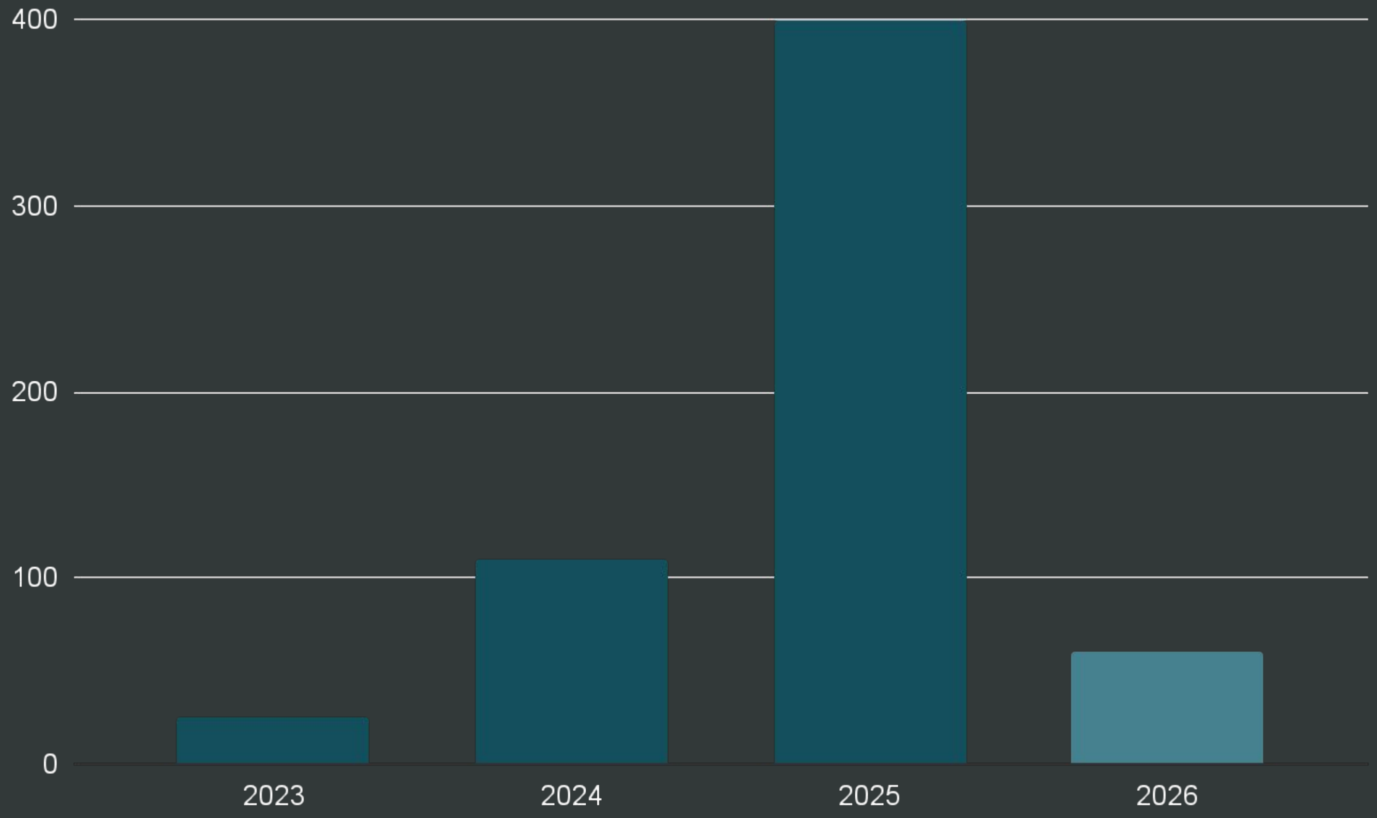
Research

Open Source

Industry

Research Volume Signal

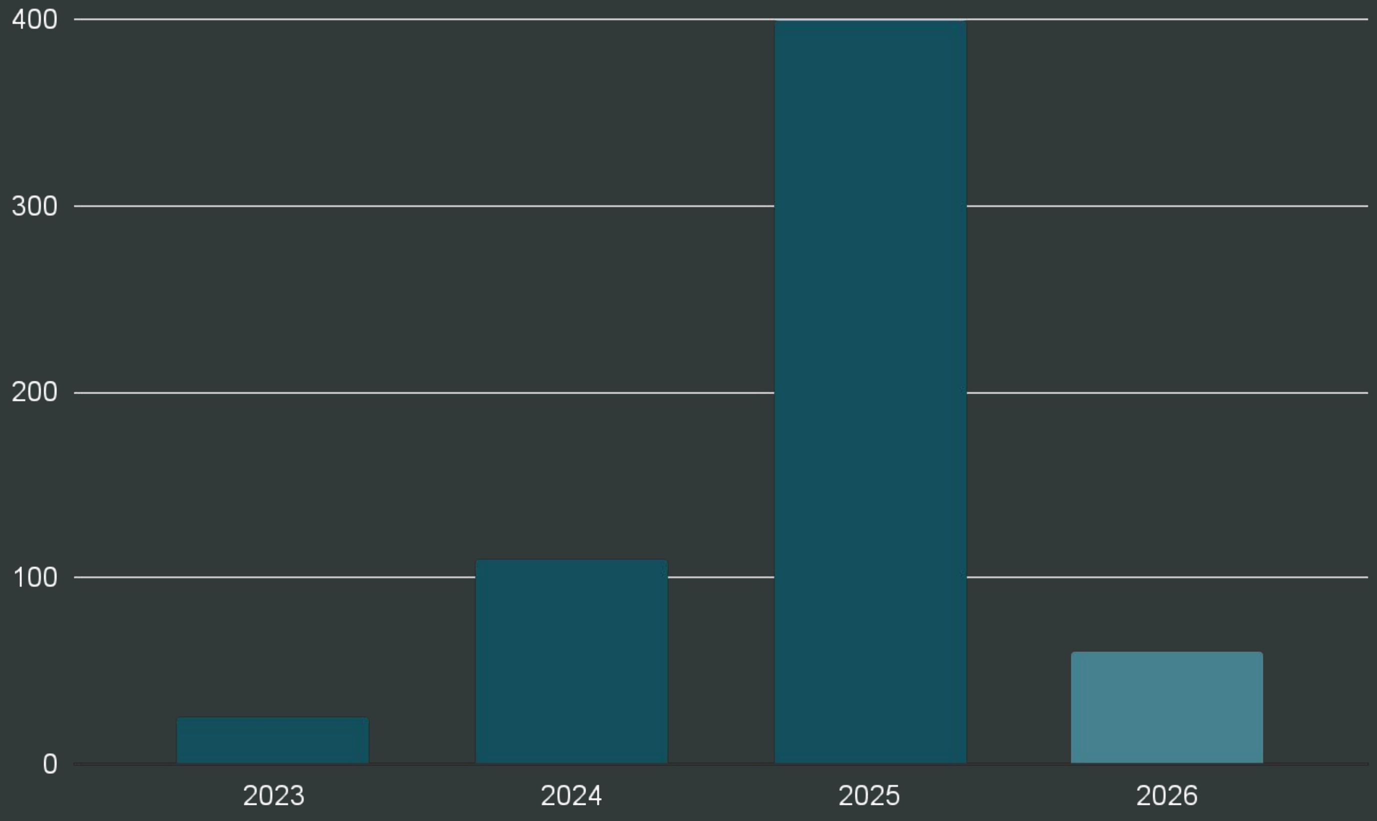
# Papers on Agent Memory
# Went From Niche to Critical



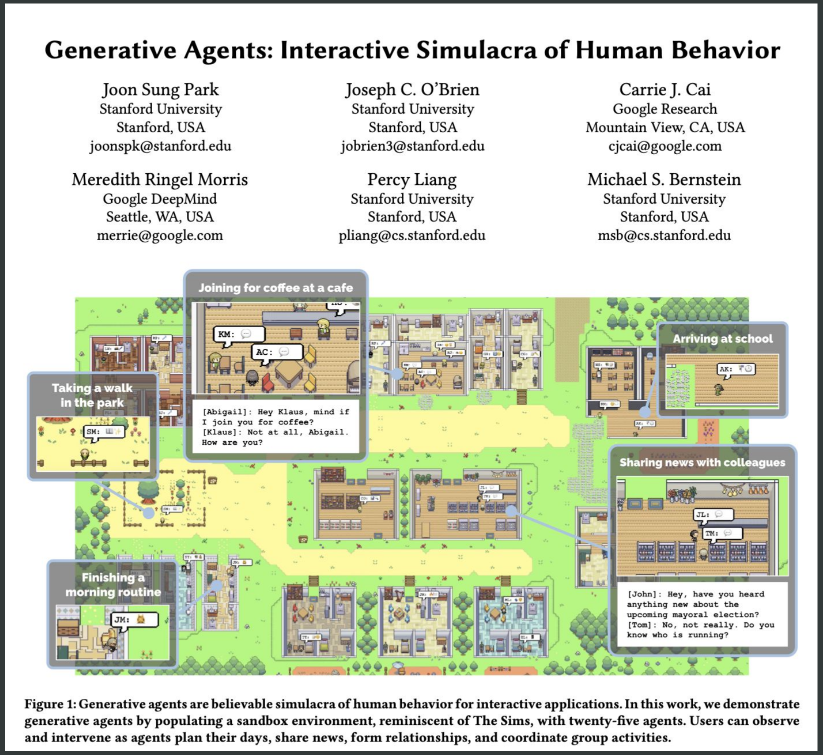
15x growth in 2 years

Research Volume Signal

# Papers on Agent Memory
# Went From Niche to Critical



16x growth in 2 years



## Generative Agents: Interactive Simulacra of Human Behavior

Joon Sung Park
Stanford University
Stanford, USA
joonspk@stanford.edu

Joseph C. O'Brien
Stanford University
Stanford, USA
jobrien3@stanford.edu

Carrie J. Cai
Google Research
Mountain View, CA, USA
cjcai@google.com

Meredith Ringel Morris
Google DeepMind
Seattle, WA, USA
merrie@google.com

Percy Liang
Stanford University
Stanford, USA
pliang@cs.stanford.edu

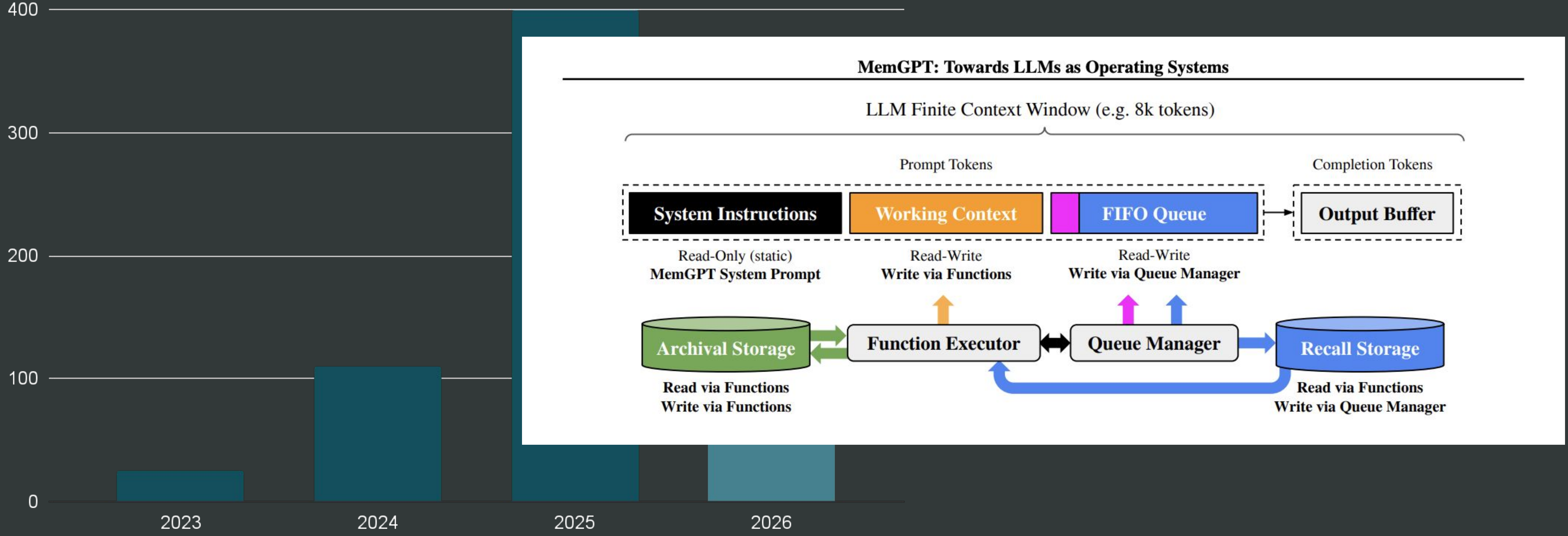
Michael S. Bernstein
Stanford University
Stanford, USA
msb@cs.stanford.edu

Joining for coffee at a cafe
KM:
AC:
[Abigail]: Hey Klaus, mind if I join you for coffee?
[Klaus]: Not at all, Abigail. How are you?
Taking a walk in the park
SM:
Arriving at school
AK:
Sharing news with colleagues
JL:
TM:
[John]: Hey, have you heard anything new about the upcoming mayoral election?
[Tom]: No, not really. Do you know who is running?
Finishing a morning routine
JM:

Figure 1: Generative agents are believable simulacra of human behavior for interactive applications. In this work, we demonstrate generative agents by populating a sandbox environment, reminiscent of The Sims, with twenty-five agents. Users can observe and intervene as agents plan their days, share news, form relationships, and coordinate group activities.

Research Volume Signal

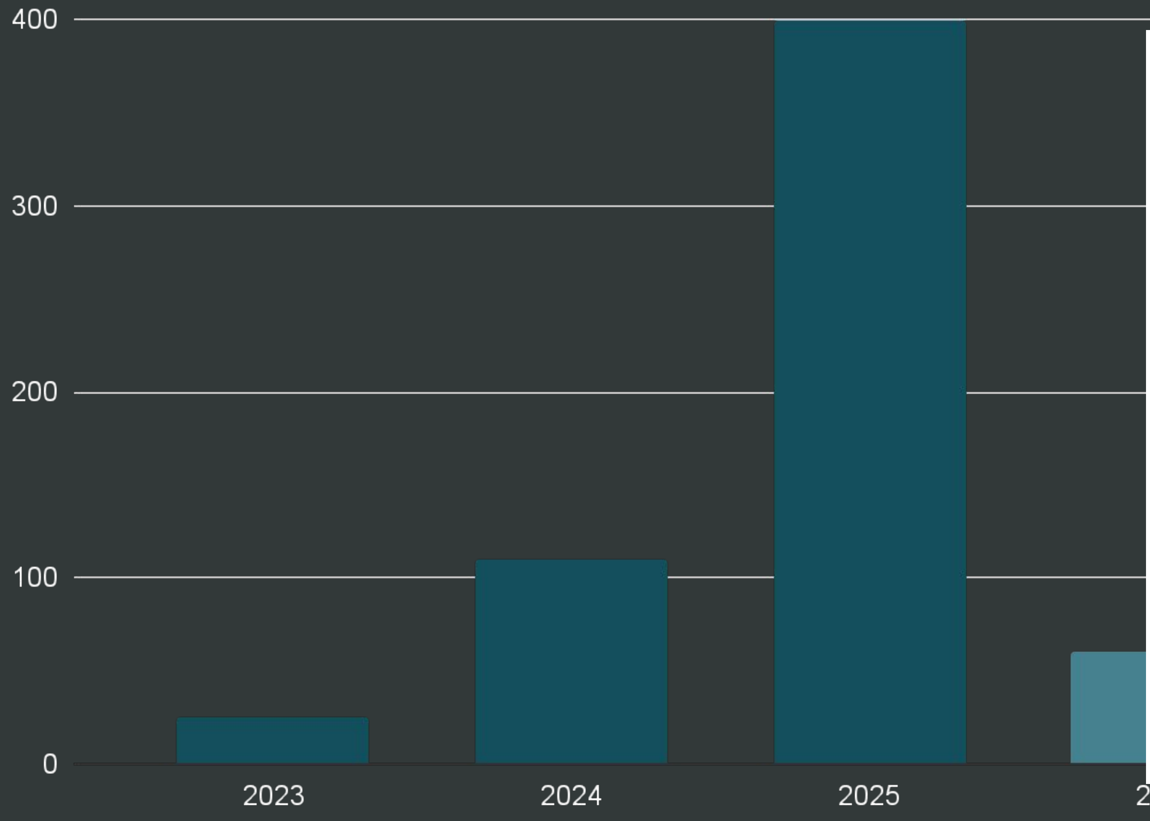
# Papers on Agent Memory
# Went From Niche to Critical



16x growth in 2 years

Research Volume Signal

# Papers on Agent Memory
# Went From Niche to Critical



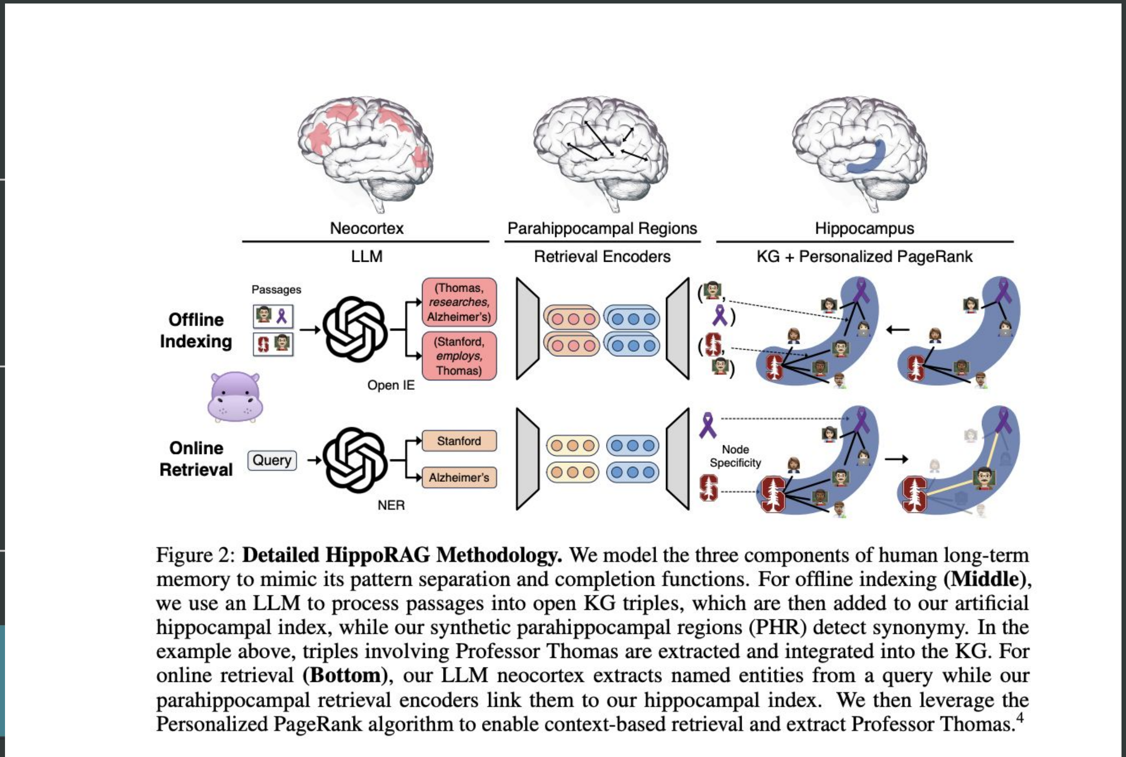
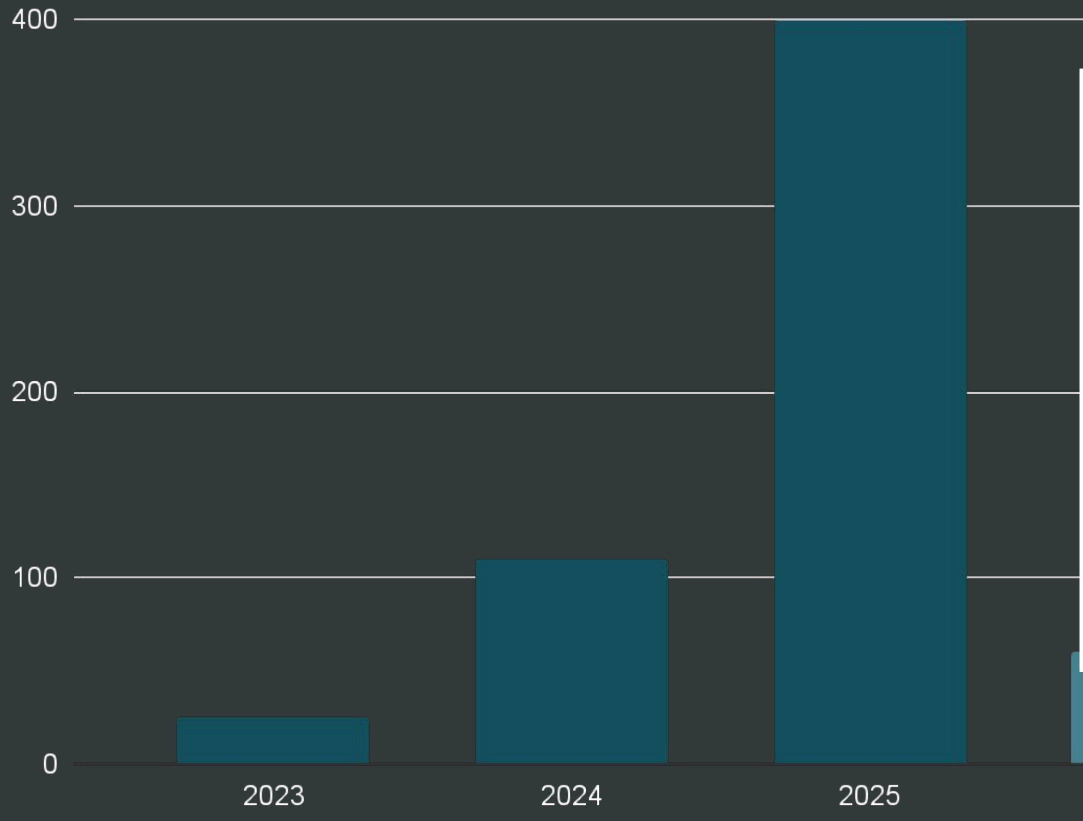


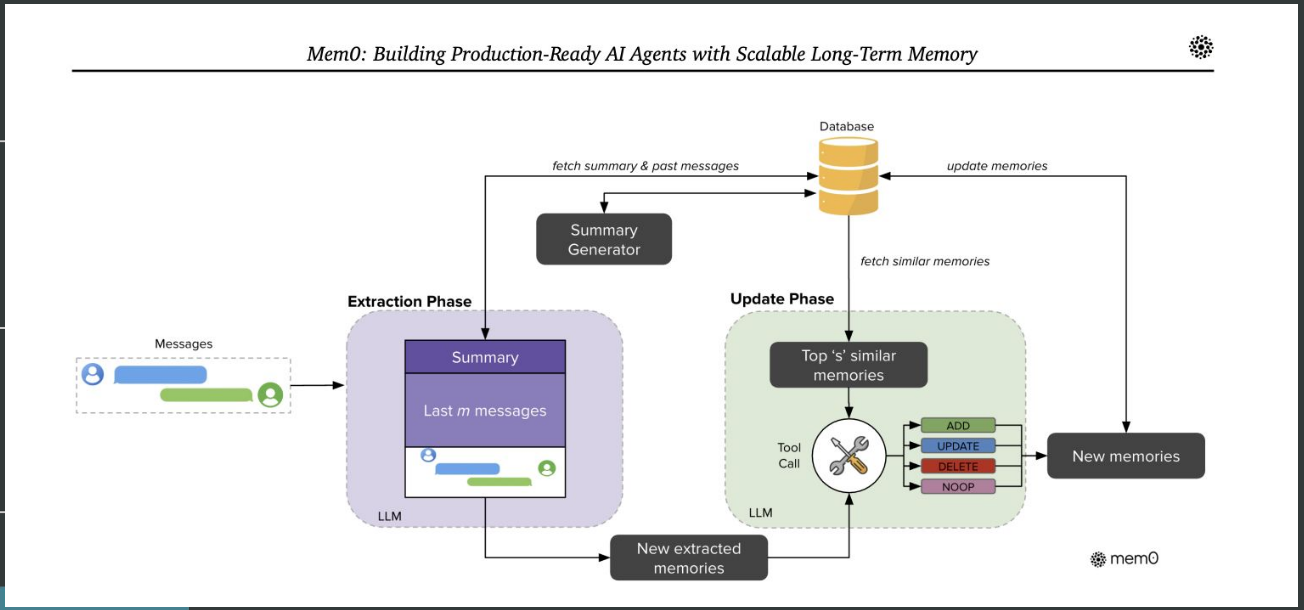
Figure 2: **Detailed HippoRAG Methodology.** We model the three components of human long-term memory to mimic its pattern separation and completion functions. For offline indexing **(Middle)**, we use an LLM to process passages into open KG triples, which are then added to our artificial hippocampal index, while our synthetic parahippocampal regions (PHR) detect synonymy. In the example above, triples involving Professor Thomas are extracted and integrated into the KG. For online retrieval **(Bottom)**, our LLM neocortex extracts named entities from a query while our parahippocampal retrieval encoders link them to our hippocampal index. We then leverage the Personalized PageRank algorithm to enable context-based retrieval and extract Professor Thomas.[4]

16x growth in 2 years

Research Volume Signal

# Papers on Agent Memory
# Went From Niche to Critical




*Mem0: Building Production-Ready AI Agents with Scalable Long-Term Memory*

16x growth in 2 years

# Keywords and Terminology Growth

## 2023

- Long-term
- Short-term
- Episodic
- Semantic
- Procedural
- Working
- RAG
- Context Window
- Vector Stores
- + more

## 2024

- Agent Memory
- Semantic Cache
- Memory management
- Shared Memory
- Associative Mem
- Memory Retrieval
- Memory Aug
- LLM Memory
- + more

## 2025

- Exocortex
- Memory Eng
- Memory Blocks
- Memory Units
- Token Space
- Memory Lifecycle
- Forgetting mech
- + more

OPEN SOURCE SIGNAL

# The community has voted with stars.

### Mem0

Universal memory layer

**29K+**

GitHub stars · May 2025

### Letta

Formerly MemGPT · UC Berkeley

**16K+**

GitHub stars · May 2025

### The Ecosystem

Dedicated memory projects in 2025

**20+**

Active OSS repos · growing weekly

INCLUDES · Zep · LangMem · MemOS · OpenMemory · Graphiti · Memary · Cognee · Second Me

**In 2023 there were 2–3 memory libraries.**

**In 2025 there are 20+ dedicated projects.**

In 2 years.

INFRASTRUCTURE SIGNAL

# Memory is no longer a feature. It's a layer.

FRAMEWORKS SHIPPING MEMORY NATIVELY

**LangChain**
Shipped LangMem as native LangGraph component

**CrewAI**
Integrated Mem0 natively for persistent user memory

**Microsoft AutoGen**
Uses Mem0 as a first-class memory provider

**MCP (Model Context Protocol)**
Memory tools now ship as MCP servers by default

THE BOLDEST SIGNAL

## Memory OS

The community isn't just building memory *libraries*.

**They're building an operating system for memory.**

v2.0 (Dec 2025):

- Multi-modal memory · Tool memory
- MCP integration · Lifecycle management
- Persistent storage · Memory consolidation

INDUSTRY SIGNAL 01 / 02

PLATFORM SIGNAL

# Every major AI platform shipped memory in 12 months.

**OpenAI**
ChatGPT Memory
**Sep 2024**
All tiers · saved
+ chat history

**Microsoft**
Copilot Memory
**2025**
Persistent memory
across M365

**Google**
Vertex AI Memory Bank
**Jul 2025**
Agent-native · Gemini
powered extraction

**Anthropic**
Claude 4 Memory Files
**2025**
Opus 4: "skilled at
maintaining memory files"

**xAI**
Grok Memory
**2025**
Cross-session
persistent context

MARKET SCALE

**$3.8B**
Raised by AI agent startups in 2024

**$13B**
Enterprise AI agent ARR by end 2025

**42%**
Agent startups at commercial scale

> "Without memory, agents treat each interaction as the first, asking repetitive questions and failing to recall preferences."
> *— Google Vertex AI Documentation, 2025*

INDUSTRY SIGNAL 02 / 02

VALIDATION SIGNAL

# Memory is the make-or-break variable.

**MIT NANDA · THE GENAI DIVIDE · JULY 2025**

**95%**

of $30–40B in enterprise AI investment delivers zero measurable P&L impact

THE ROOT CAUSE, PER MIT:

**"Most GenAI systems do not retain feedback, adapt to context, or improve."**

63% of executives demand context retention
66% want systems that learn from feedback
based on 300+ initiative reviews, 52 interviews, 153 senior leaders

**PALO ALTO NETWORKS · MOLTBOT REPORT · JAN 2026**

Moltbot: 85K GitHub stars in 1 week
viral because of one capability:

**Persistent Memory**

PALO ALTO ANALYSIS:

"Moltbot succeeded by solving what users cited as the #1 friction point in AI: having to re-explain context **every single time."**

The most starred agent project in GitHub history
Memory was the differentiator

## Prompt Engineering

## Context Engineering

## Memory Engineering ?

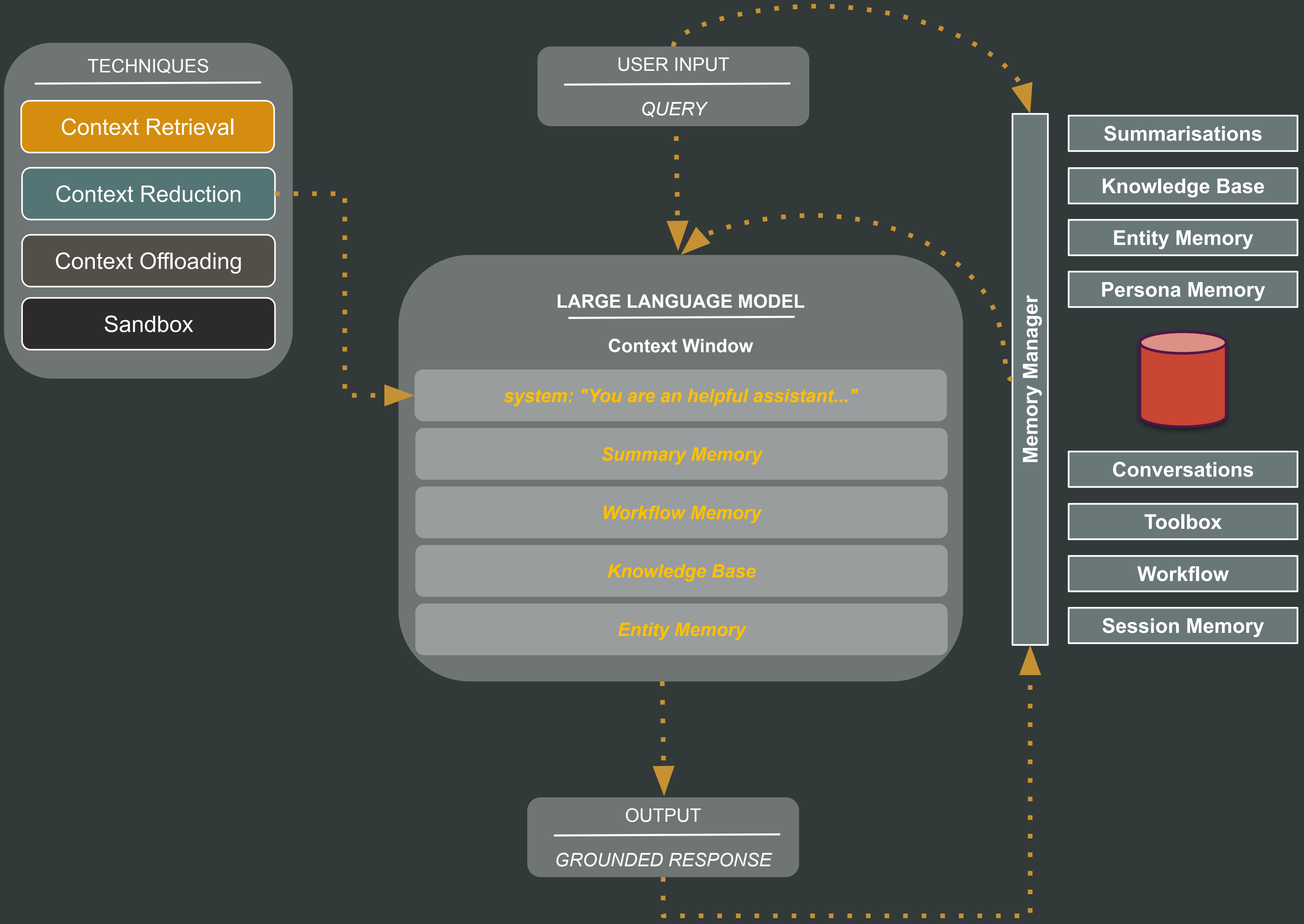
# Memory Engineering

The discipline of creating systems that systematically process, store, retrieve, and update information; transforming raw data into persistent memory components, then dynamically assembling the most relevant of those components to populate an LLM's context window for each interaction, session, or operational round.

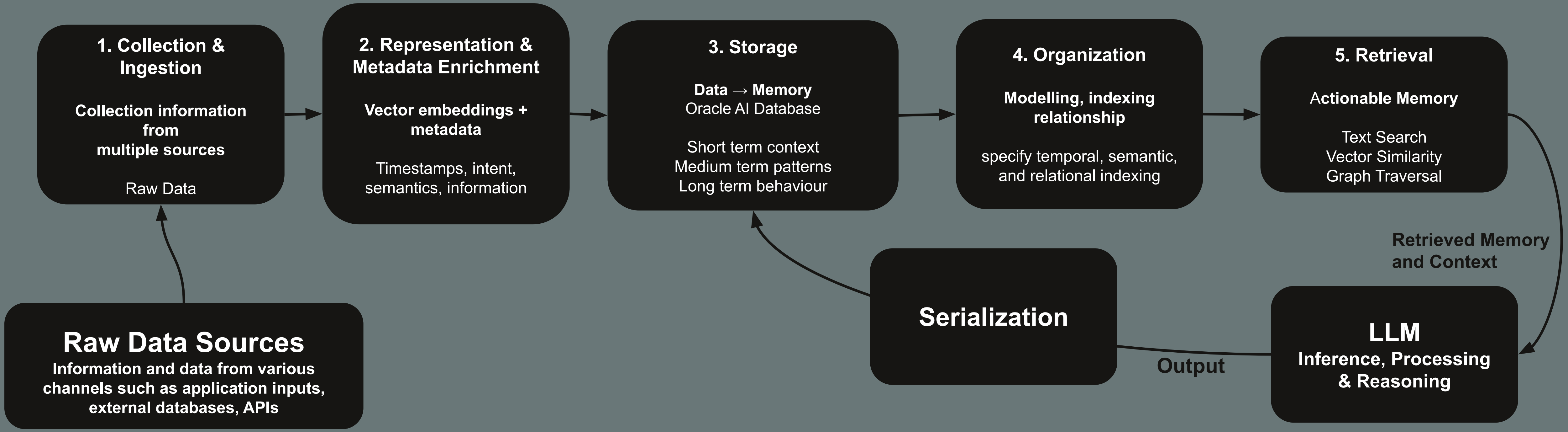
The goal is to build AI agents that don't just respond intelligently in the moment, but continuously adapt to new information, growing more reliable, believable, and capable over time.

# Memory Engineering

The engineering discipline focused on designing, building, and maintaining memory systems for AI agents. It encompasses the storage, retrieval, classification, and lifecycle management of agent memory.



# AI Memory Lifecycle



**1. Collection & Ingestion**

**Collection information from multiple sources**

Raw Data

**2. Representation & Metadata Enrichment**

**Vector embeddings + metadata**

Timestamps, intent, semantics, information

**3. Storage**

**Data → Memory**
Oracle AI Database

Short term context
Medium term patterns
Long term behaviour

**4. Organization**

**Modelling, indexing relationship**

specify temporal, semantic, and relational indexing

**5. Retrieval**

**Actionable Memory**

Text Search
Vector Similarity
Graph Traversal

**Retrieved Memory and Context**

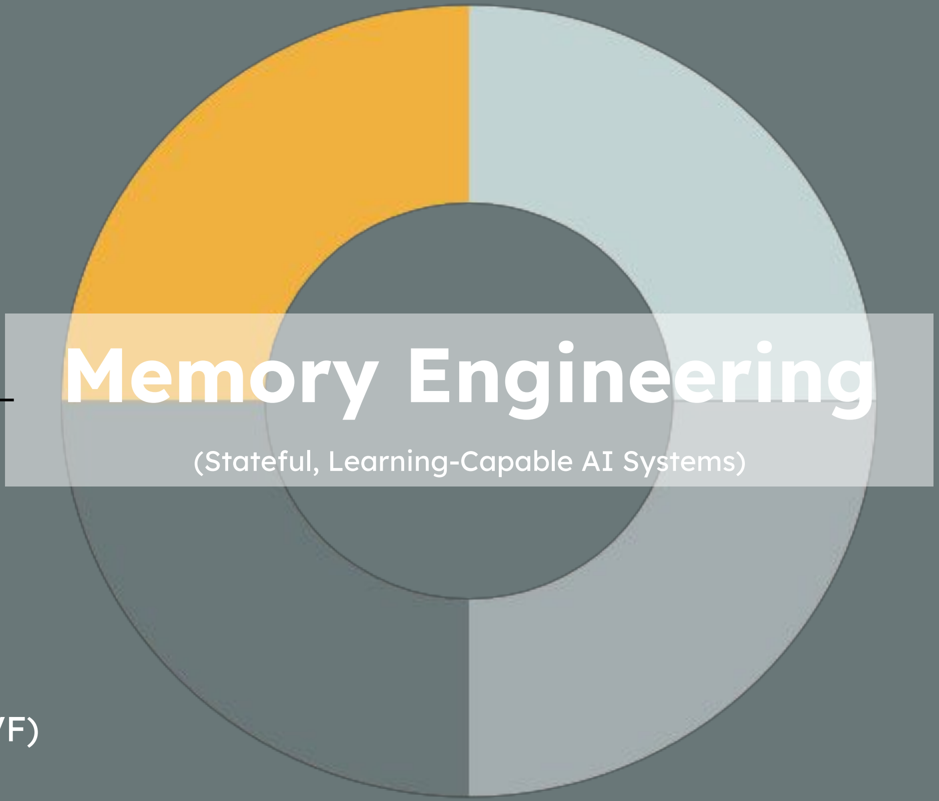
**LLM**
**Inference, Processing & Reasoning**

**Output**

**Serialization**

**Raw Data Sources**
**Information and data from various channels such as application inputs, external databases, APIs**

**Continuous Learning Cycle**
Memory feeds LLM reasoning and outcomes become new memory components



# Memory Engineering

(Stateful, Learning-Capable AI Systems)

## Database Engineering

Persistent Storage
Typed Schemas
ACID Transactions
Multi-Store Architecture
Versioning & Backup

## Agent Engineering

Memory Lifecycle
Write-Back Loops
Memory Extraction
Autonomous Consolidation
Context-Aware Routing

## Information Retrieval

Hybrid Search
Vector Indexes (HNSW, IVF)
Relevance Ranking
Context Assembly
Query Optimization

## Machine Learning Engineering

Embedding Models
Fine-Tuning (SLMs)
Model Versioning
Reranking Pipelines
Continual Learning

# Memory Engineering

(Stateful, Learning-Capable AI Systems)

## Database Engineering

Provides the persistent storage layer for agent memory systems. Manages typed memory schemas (episodic, semantic, procedural, associative), ensuring ACID transactions for memory writes and consistent state across multi-store architectures. Handles backup, versioning, and recovery strategies to maintain memory integrity over time. Designs sharding, replication, and caching strategies that enable memory systems to scale across distributed environments while maintaining performance and reliability.

## Agent Engineering

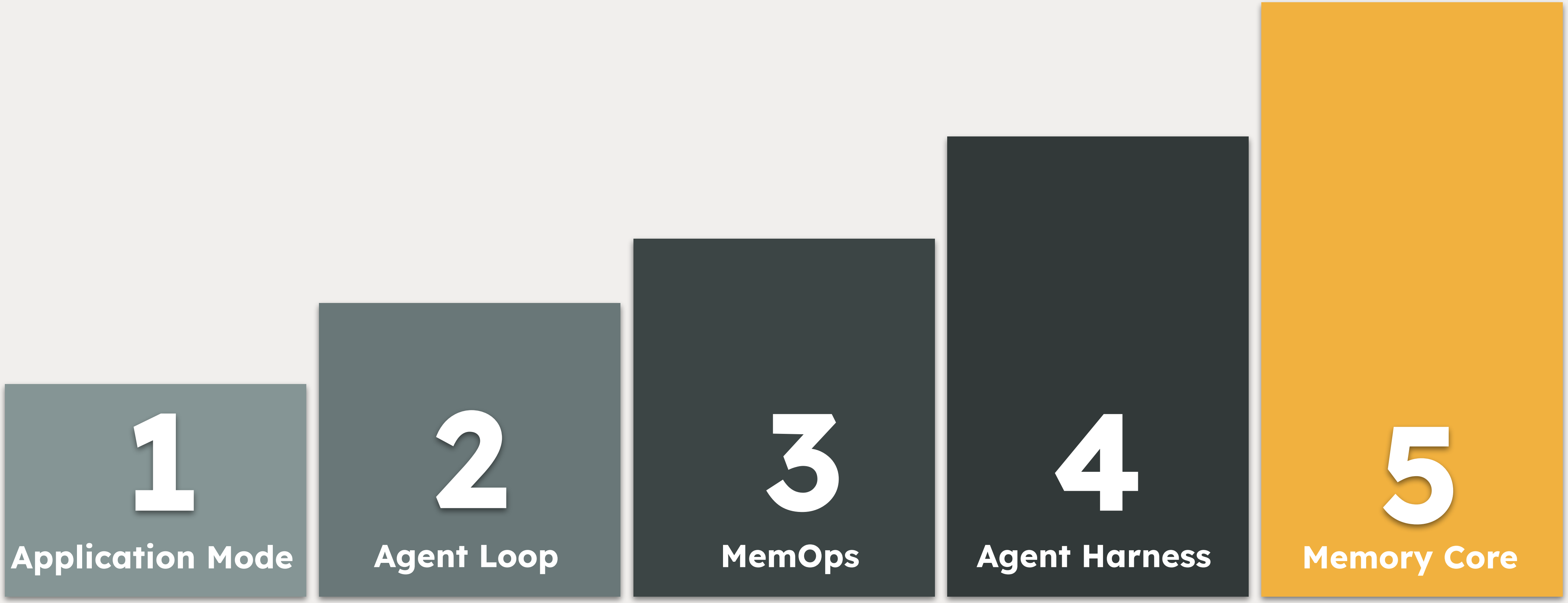
Defines the memory lifecycle: creation, consolidation, decay, and retrieval. Designs write-back loops that extract learnings from agent outputs, classify memory types, and route updates to appropriate stores. Enables autonomous memory maintenance where agents self-manage their own knowledge without manual intervention.

## Information Retrieval

Optimizes how memories are found and ranked through hybrid search strategies combining dense vector similarity with sparse keyword matching. Tunes index selection (HNSW, IVF) and approximate nearest neighbor algorithms to balance retrieval speed with accuracy. Ranks results by relevance, recency, and importance, then assembles context from multiple stores to fit within LLM token limits while preserving critical information. Manages the full retrieval pipeline from query to context delivery.

## Machine Learning Engineering

Fine-tunes embedding models and small language models for domain-specific memory needs, optimizing semantic representations for enterprise contexts. Manages model versioning, deployment pipelines, and performance monitoring across the memory stack. As the system matures, ML Engineering becomes the arbiter of model selection, determining when to use foundation models versus fine-tuned variants, and balancing accuracy, latency, and cost across the entire memory architecture.



1 Application Mode

2 Agent Loop

3 MemOps

4 Agent Harness

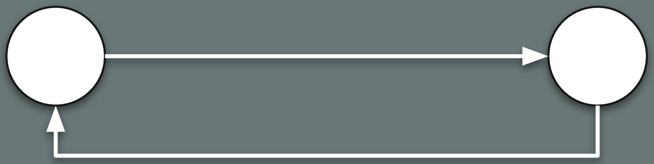
5 Memory Core

# Common Application Modes in AI Agents

## Assistant

Conversational, reactive, relationship-driven agents.

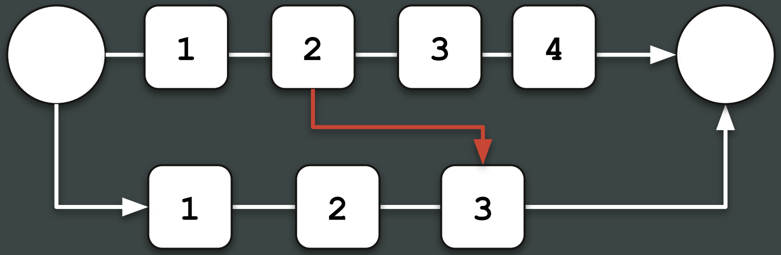
Assistant agents typically operates in an interactive, turn-by-turn conversation loop, responding to user instructions, maintaining short-term context, and retrieving long-term preferences or past interactions when needed.



## Workflow

Multi-step, goal-oriented, stateful process execution.

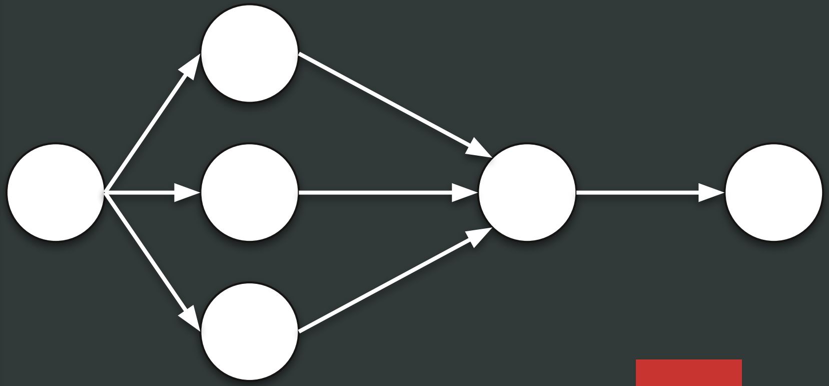
Workflow Mode agents follow structured, multi-step procedures, often involving tool use, planning, and state tracking. They act like autonomous workers executing a task from start to finish.



## Deep Research

Long-horizon reasoning, synthesis, and multi-source investigation.

Agents perform extended, multi-turn investigations, pulling information from diverse sources, iterating, reflecting, and consolidating knowledge across time.

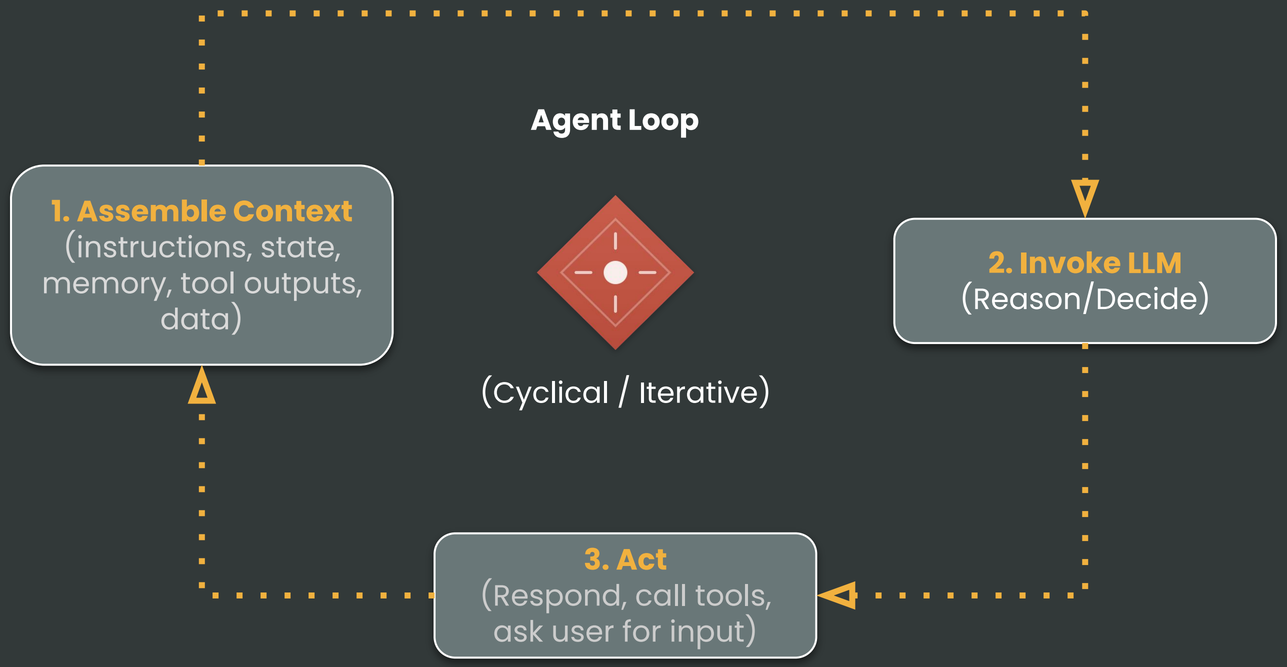


# Agent Loop

A cyclical, iterative execution pattern inside a single agent run/turn where an agent repeatedly:

1. **assembles** context (instructions, conversation state, retrieved memory, tool outputs, relevant data),
2. **invokes** an LLM to reason/decide, and then
3. **acts** (responds, calls tools, writes memory/state, or updates the plan),

until a stop condition is met, e.g., a final answer is produced, a goal is completed, an error/timeout occurs, or the agent explicitly decides to exit.

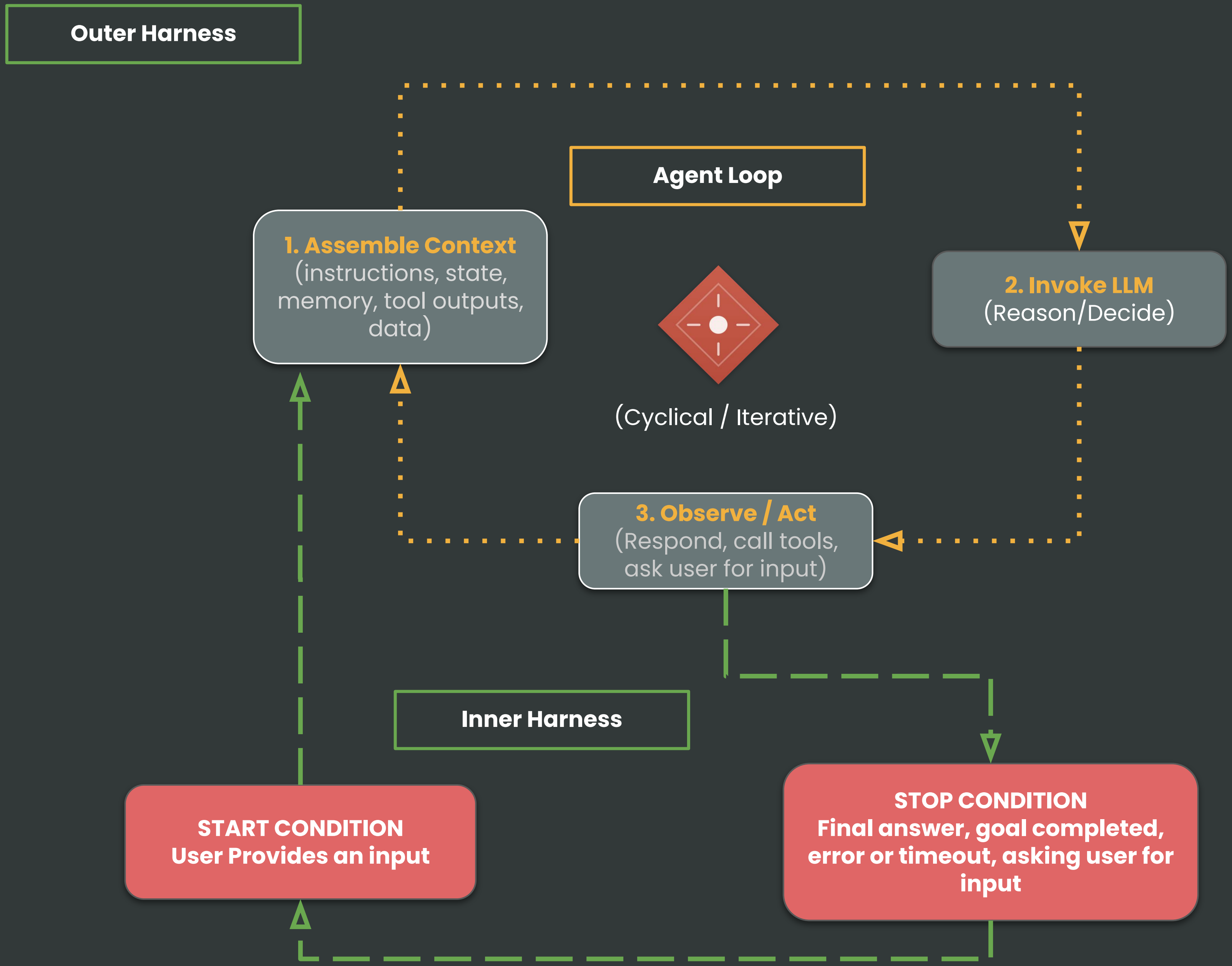

Agent Loop
1. Assemble Context
(instructions, state, memory, tool outputs, data)
(Cyclical / Iterative)
2. Invoke LLM
(Reason/Decide)
3. Act
(Respond, call tools, ask user for input)

# Agent Harness and MemOps

The infrastructure and scaffolding surrounding an LLM-enabled agent designed to ensure reliable, desirable outcomes.

Components include memory operations, control systems, environment information, coding principles, and task-specific tooling.

**Memory operations are particularly critical, enabling agents to adapt reliably and extend their capabilities over time.**



Outer Harness

Agent Loop

**1. Assemble Context** (instructions, state, memory, tool outputs, data)

**2. Invoke LLM** (Reason/Decide)

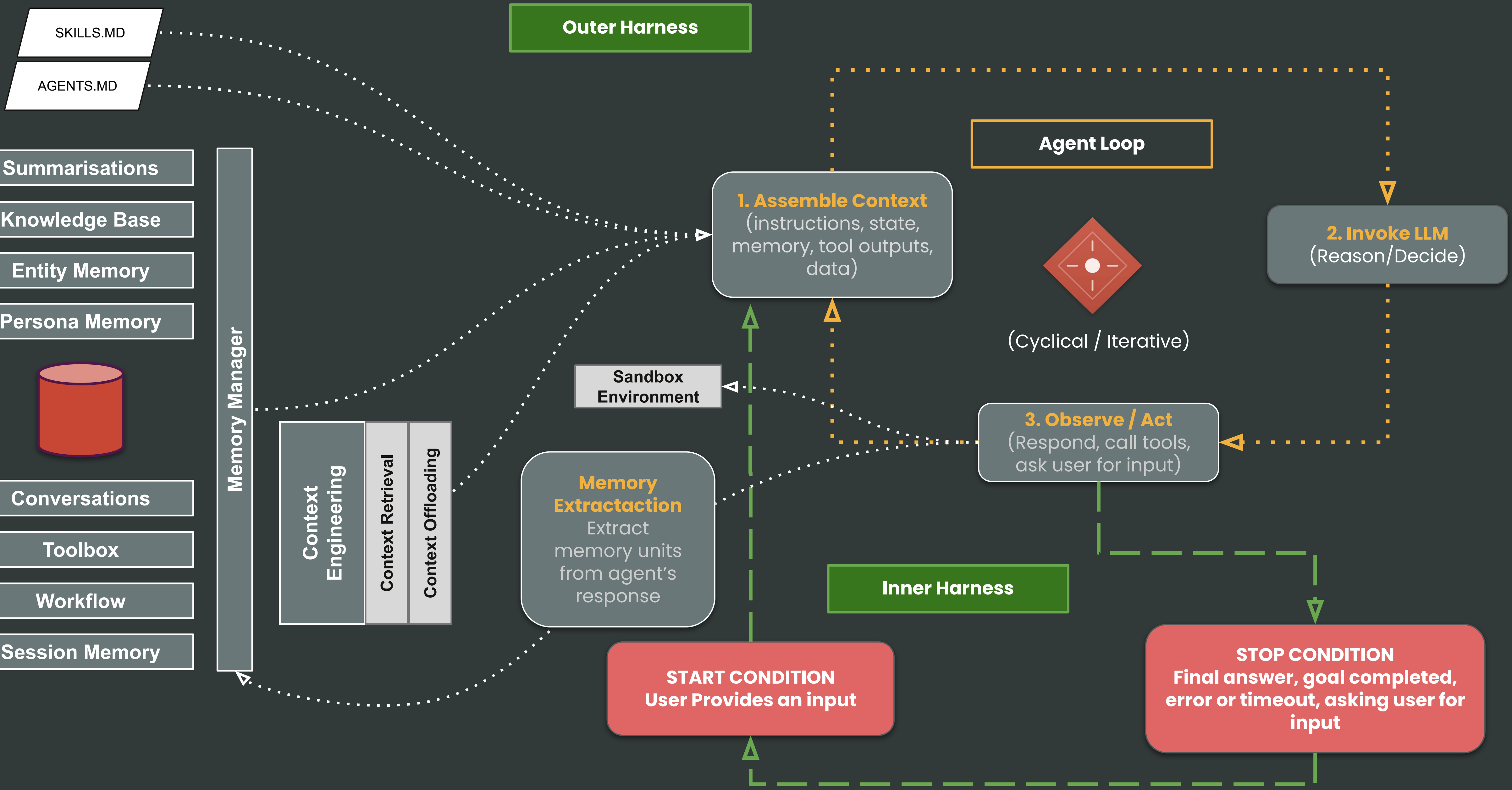
**3. Observe / Act** (Respond, call tools, ask user for input)

(Cyclical / Iterative)

Inner Harness

**START CONDITION**
**User Provides an input**

**STOP CONDITION**
**Final answer, goal completed, error or timeout, asking user for input**



Outer Harness

SKILLS.MD

AGENTS.MD

Summarisations

Knowledge Base

Entity Memory

Persona Memory

Conversations

Toolbox

Workflow

Session Memory

Memory Manager

Context Engineering

Context Retrieval

Context Offloading

Agent Loop

1. Assemble Context (instructions, state, memory, tool outputs, data)

2. Invoke LLM (Reason/Decide)

(Cyclical / Iterative)

Sandbox Environment

3. Observe / Act (Respond, call tools, ask user for input)

Memory Extractaction Extract memory units from agent's response

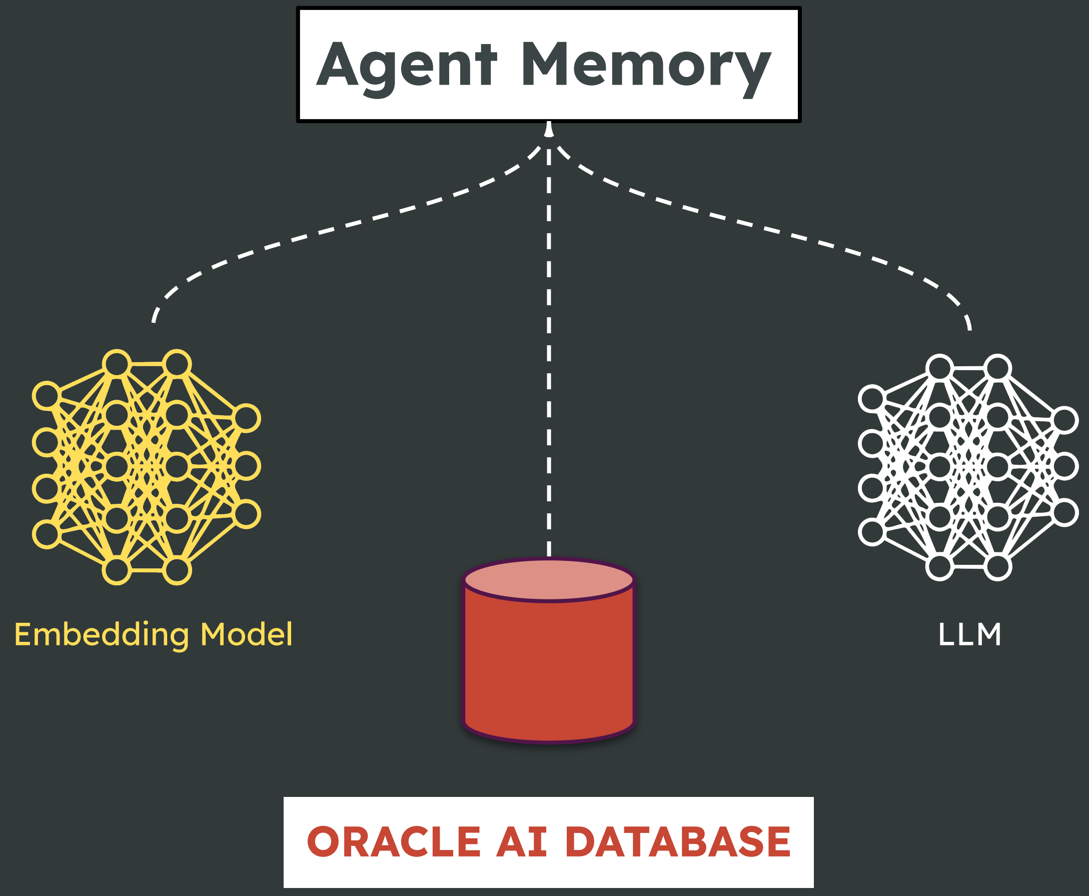
Inner Harness

START CONDITION
User Provides an input

STOP CONDITION
Final answer, goal completed, error or timeout, asking user for input

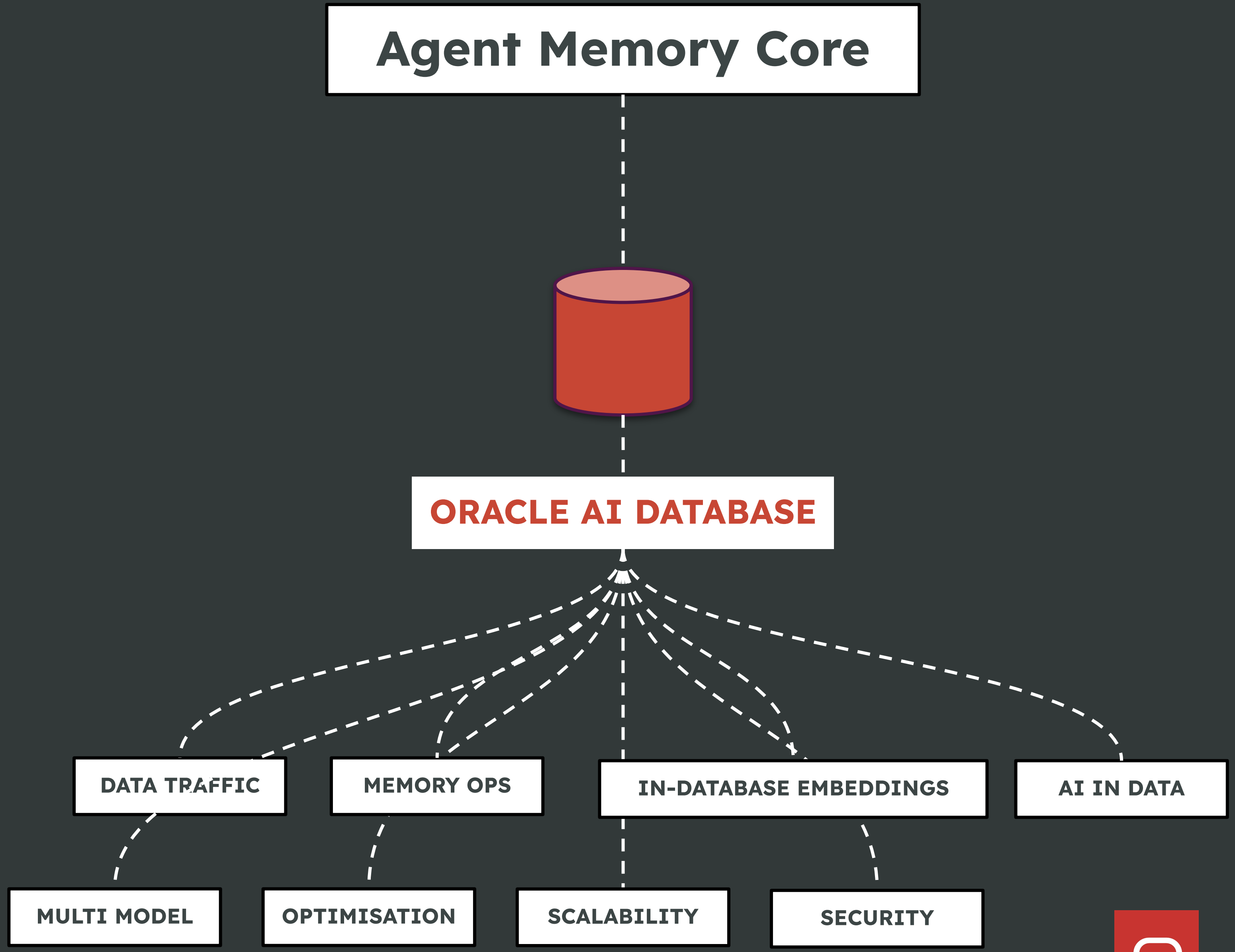
# Agent Memory Core

The primary data infrastructure component of an agent system, responsible for managing the complete lifecycle of agent memory. This database layer handles persistent storage, efficient retrieval, and memory operations that enable agents to adapt to new information, learn from interactions, and maintain consistent performance across sessions.



# Agent Memory Core

The primary data infrastructure component of an agent system, responsible for managing the complete lifecycle of agent memory. This database layer handles persistent storage, efficient retrieval, and memory operations that enable agents to adapt to new information, learn from interactions, and maintain consistent performance across sessions.



***Database technology will be at the very center***

**Insert Target Here** is DEAD

**SaaS** is DEAD

RAG is DEAD

# SWE is DEAD

IDE is DEAD

*the internet is not going to end the database market. It will drastically expand it.*

# The Agent Stack

| | |
|---|---|
| Application | Applications, UIs, and enterprise systems that embed, trigger, or consume agent outputs. This is where users or automated systems initiate agent tasks. |
| Gateway and Connectivity | Communication pathways, APIs, chat channels, connectors, and distributed messaging. Handles external calls into the agent and internal asynchronous communication between services/agents. |
| Orchestration | The agent's control logic: planning, routing, decomposing tasks, enforcing policies, and managing the plan → act → observe loop. |
| Reasoning | The cognitive engine. Performs LLM reasoning, interpretation, tool-call generation, synthesis, and context understanding. |
| Memory Managers | Responsible for embedding generation, query vectorisation, reranking, memory-type selection, summarisation, consolidation, and deciding what and how the agent retrieves before filling Working Memory. |
| Tooling | The agent's action execution system. Tools/APIs the agent can call to perform real operations: queries, integrations, scripts, functions, pipelines. |
| Governance and Reliability | Ensures safe, observable, compliant, and reliable agent behaviour: guardrails, IAM, logging, monitoring, auditing, anomaly detection. |
| Memory Core | Durable memory storage + the engine that executes vector search, JSON retrieval, relational and graph queries, and hybrid search. |
| Infrastructure | Runtime foundation for agents: compute, containers, networking, scaling, observability, service mesh. Ensures operational reliability. |

# The Agent Stack

| | |
|---|---|
| Application | Oracle APEX, Oracle Fusion Apps |
| Gateway and Connectivity | API Gateway, Integration Cloud, Digital Assistants, Oracle Streaming Service, Service Bus, OCI Notifications, Open Agent Specification |
| Orchestration | OCI Generative AI Agents, Wayflow, OCI Functions, AgentSpec |
| Reasoning | OCI Generative AI (LLMs), OCI Data Science Model Deployment |
| Memory Managers | OCI Language Models and Embeddings Models, Oracle AI Integration Ecosystem |
| Tooling | Oracle Integration Cloud, OCI Functions |
| Governance and Reliability | Cloud Guard, OCI IAM, OCI Logging, OCI Monitoring, OCI Audit, Data Safe |
| Memory Core | Oracle AI Database, Autonomous Database, FreeSQL, Oracle Feature Store |
| Infrastructure | OCI Compute, OKE, OCI Functions, Load Balancer, Oracle Data Science |

Our mission is to help people see data in new ways, discover insights, unlock endless possibilities.

# JOURNEY