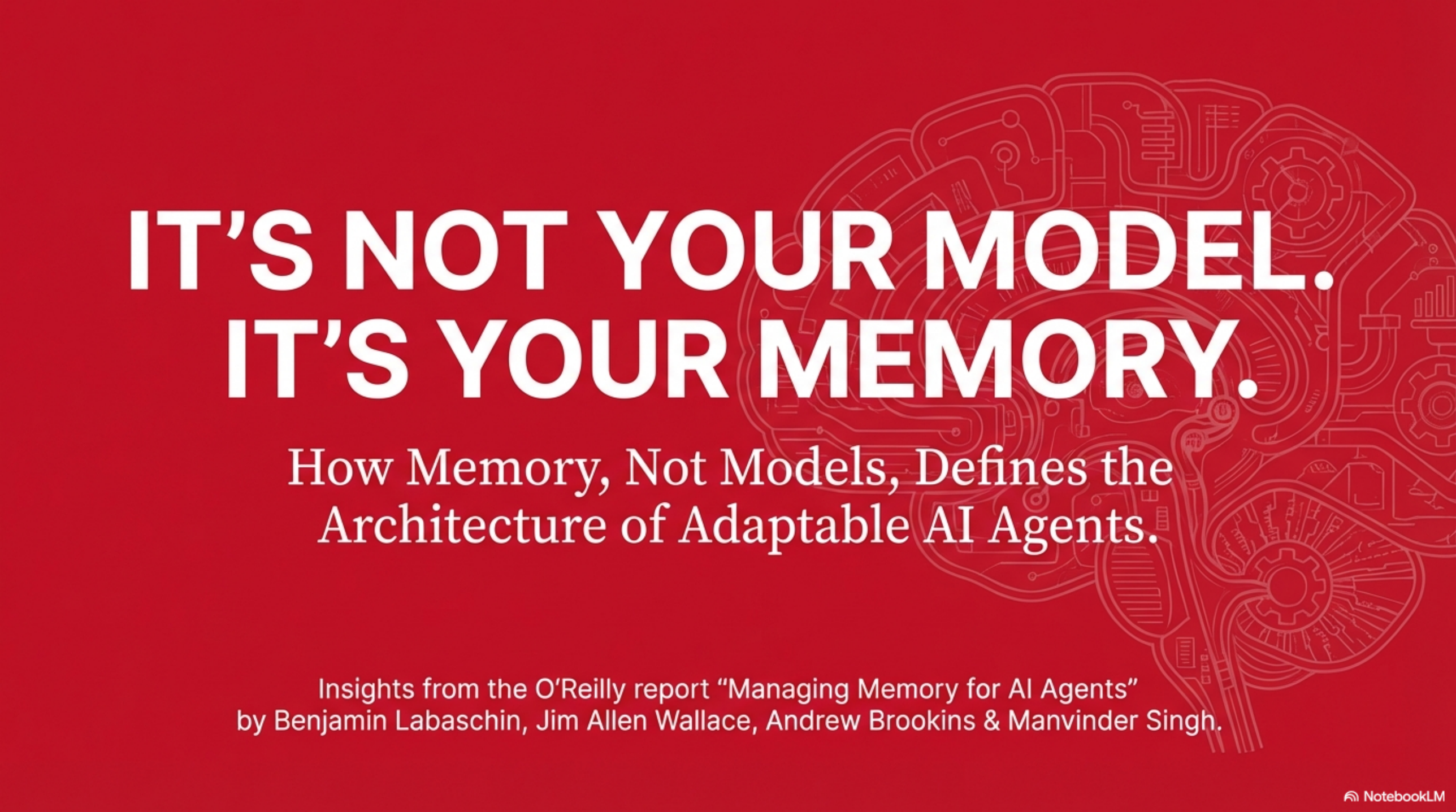


IT'S NOT YOUR MODEL. IT'S YOUR MEMORY.

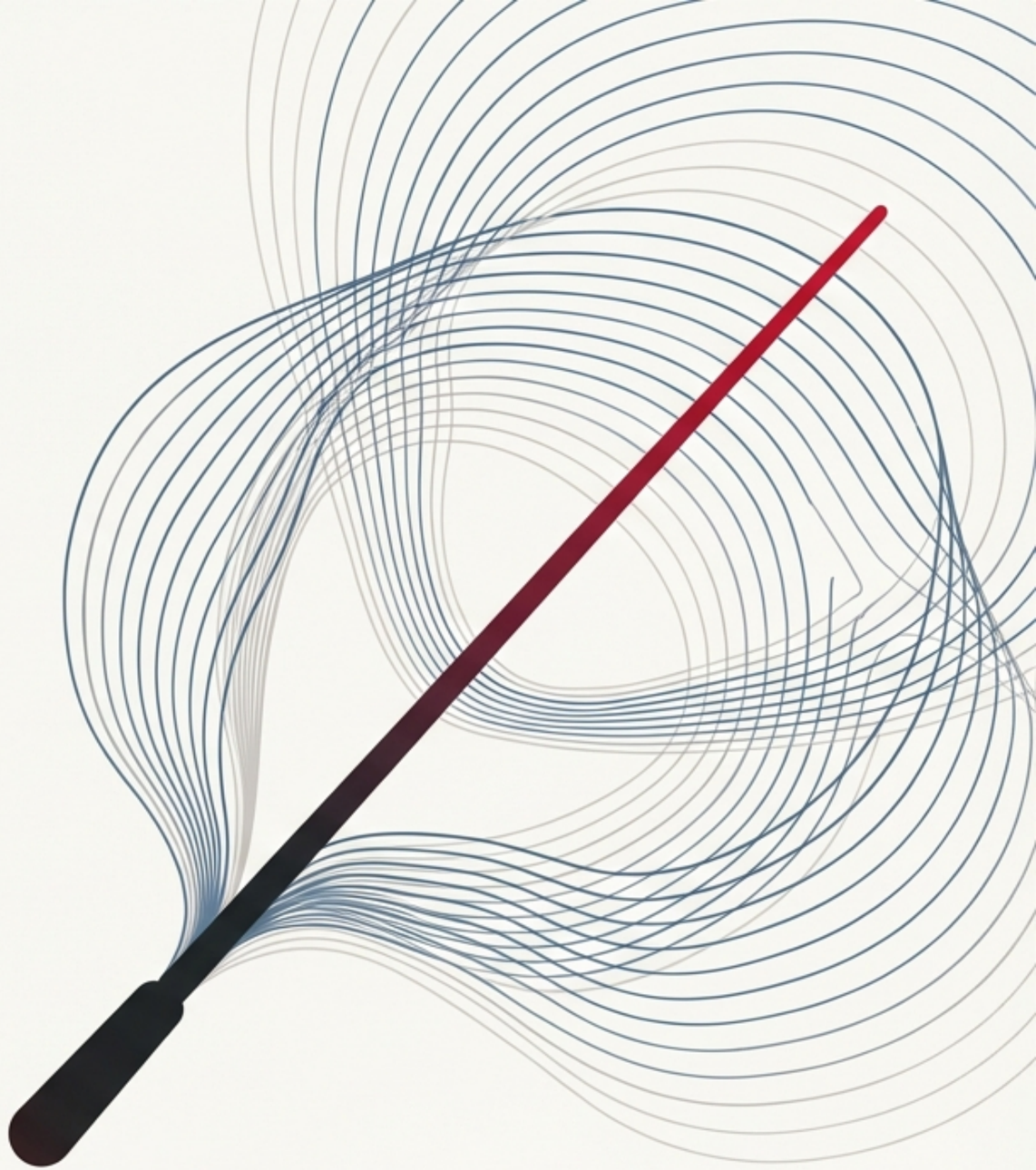
How Memory, Not Models, Defines the Architecture of Adaptable AI Agents.

Insights from the O'Reilly report "Managing Memory for AI Agents"
by Benjamin Labaschin, Jim Allen Wallace, Andrew Brookins & Manvinder Singh.

The Most Important Agent Will Always Be the Human Agent.

We are the conductors guiding these powerful new orchestras. The systems we build—vector databases, semantic caches, collective memory platforms—are instruments in our hands. They can store vast amounts of data and retrieve it in clever ways, but it is we who decide what constitutes success, what memories matter most, and how these systems should serve our goals.

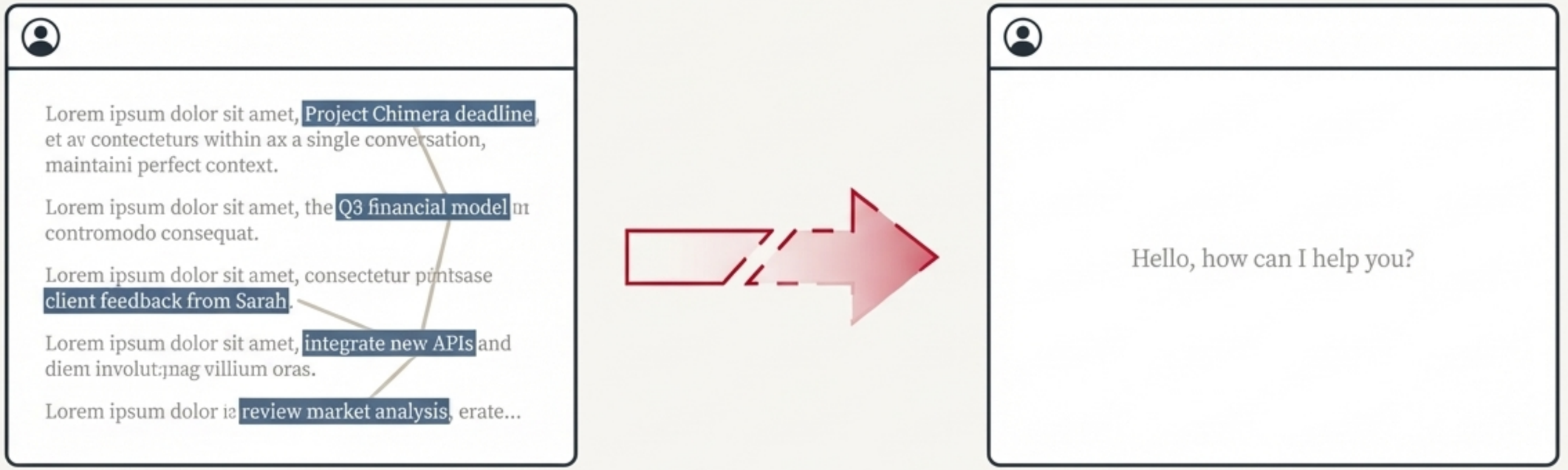
This presentation explores how to master the most critical instrument: Memory.



The Problem: Agent Amnesia.

We've all experienced it. An agent that's brilliant within a single conversation, maintaining perfect context. But tomorrow, or even an hour later, it's met with the digital equivalent of a blank stare.

This isn't an inconvenience; it's a fundamental limitation. The agent's amnesia forces us into a loop of recontextualisation, turning what should be an ongoing collaboration into a series of disconnected encounters.



Previous Session

New Session

It’s Data Management, But for a Nondeterministic World.

While agent memory is about data, storage, and retrieval, how agents ***use*** that data is fundamentally different from any system we’ve built before.

Deterministic & Precise

- SELECT statements return exact records.
- Retrieval is based on perfect matches.
- Data importance is static.

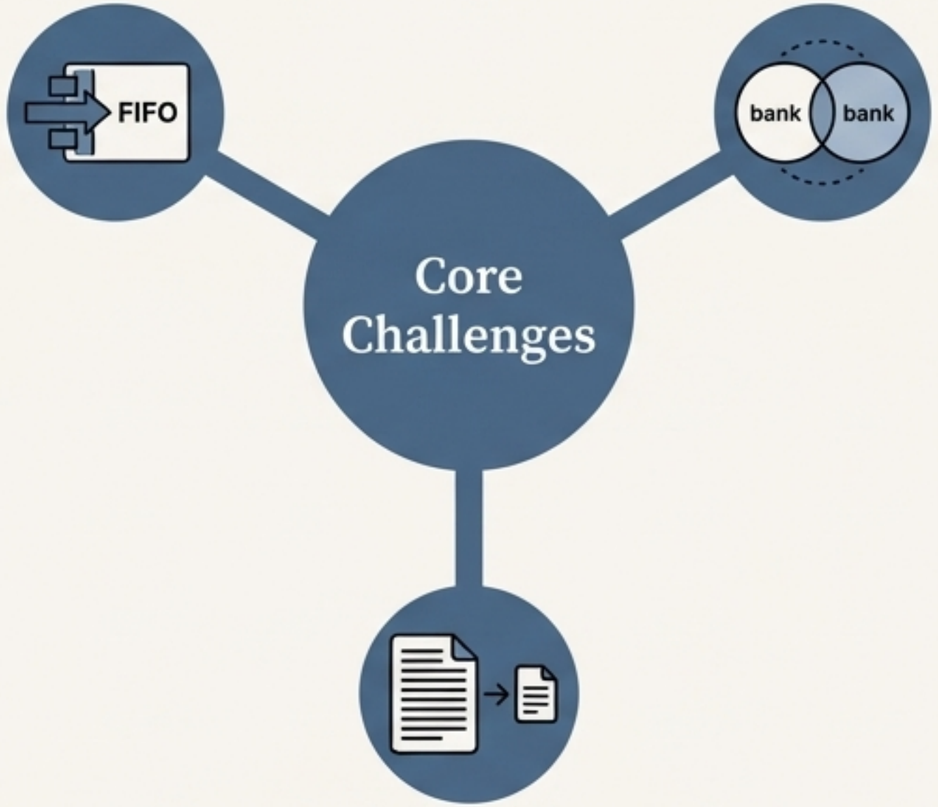
Nondeterministic & Fuzzy

- Retrieval is a fuzzy search through semantic space.
- Relevance is calculated, not guaranteed.
- Agents must dynamically decide what to retain, compress, or forget.
- The same query may yield different results.

The Technical Anatomy of the Memory Challenge.

Context Window Limits

The architecture of transformers requires **quadratically** more processing as context increases. Simple strategies like **FIFO** (first in, first out) mean early context is lost.



The Imprecision of Retrieval

Retrieval is based on semantic similarity, which is inherently fuzzy.

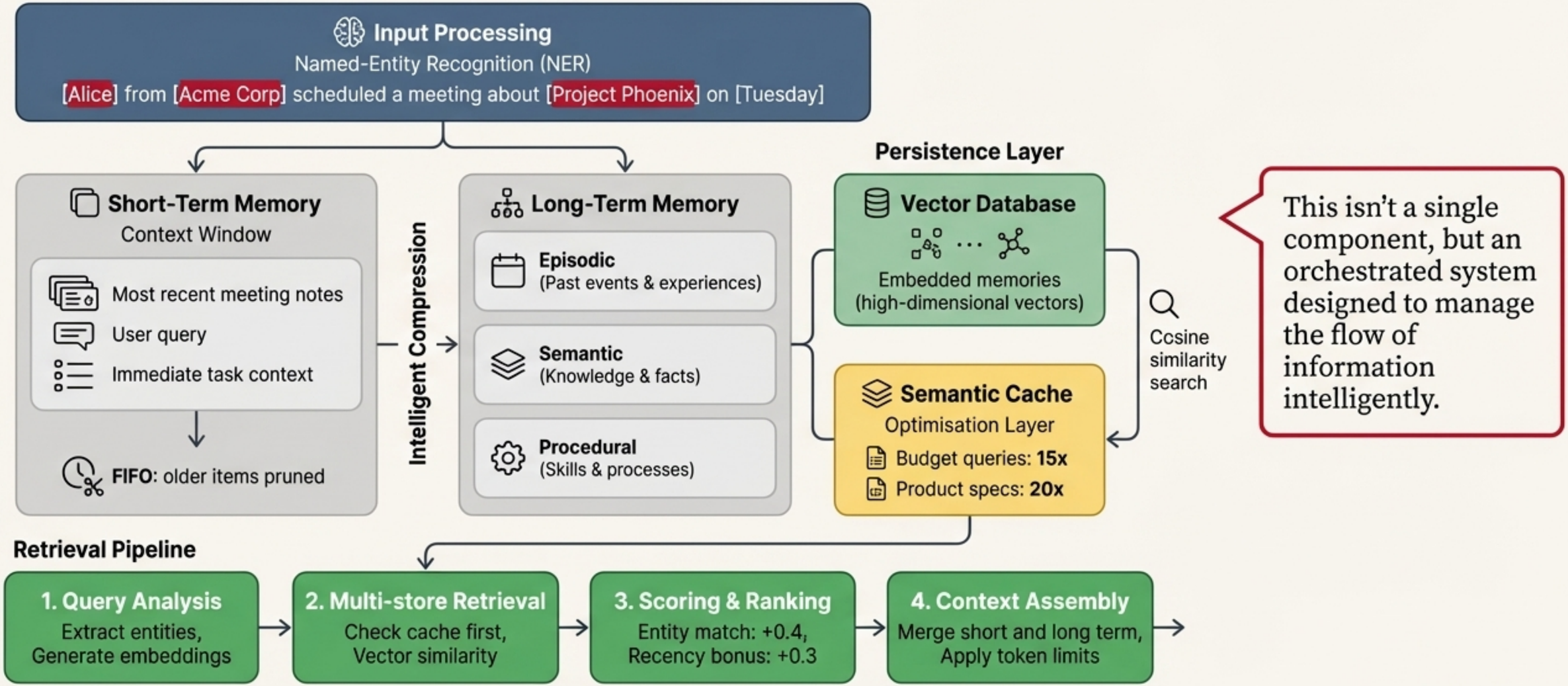
The classic example: `bank` (financial) and `bank` (riverside) live in completely different neighbourhoods of meaning.

Model choice and embeddings dictate which algorithm to use (cosine similarity, Euclidean distance, etc.).

The Peril of Compression

Summarisation by definition means losing detail. For a legal text, you might get the broad strokes but lose a critical negation or case reference that completely changes the meaning.

A Blueprint for a Solution: The Modern Memory Pipeline.



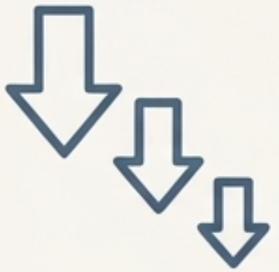
This isn't a single component, but an orchestrated system designed to manage the flow of information intelligently.

The Solution Toolkit: Core Memory Strategies

To overcome the memory challenge, engineers employ several key strategies to decide what to store, what to promote, and how to retrieve it.

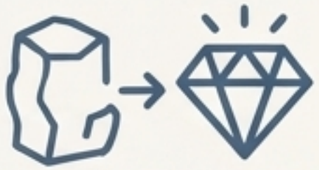
Importance Scoring

Calculating memory importance based on recency, frequency of reference, user engagement, and keyword relevance.



Cascading Memory

Allowing the agent itself to choose what to promote to long-term storage and what to retrieve from it.



Intelligent Compression

Using specialised models to condense conversation history into key details, events, and decisions, rather than crude summarisation.

Checkpointing

Periodically saving an agent's internal state (its memory) to persist information across sessions, often using fast, real-time systems like Redis.

Enhancing Precision with Structured Data

Named-Entity Recognition (NER) is essential for **accurate, retrievable memory.** It transforms the fuzzy world of natural language into structured precision that agents can **reliably work** with.

Before

Unstructured Query

What did John say about the Q4 budget during our last meeting in Paris?

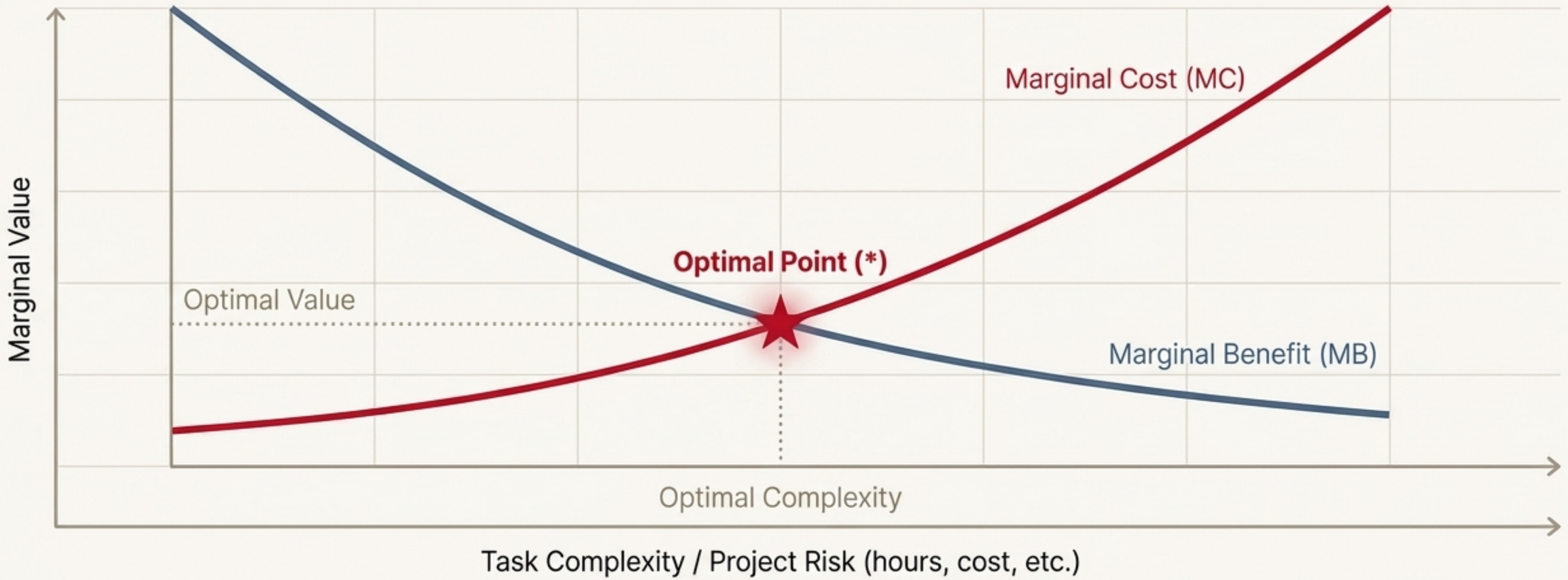
After

Structured with NER

What did **John** [PERSON] say about the **Q4 budget** [TOPIC] during our last meeting in **Paris** [LOCATION] ?

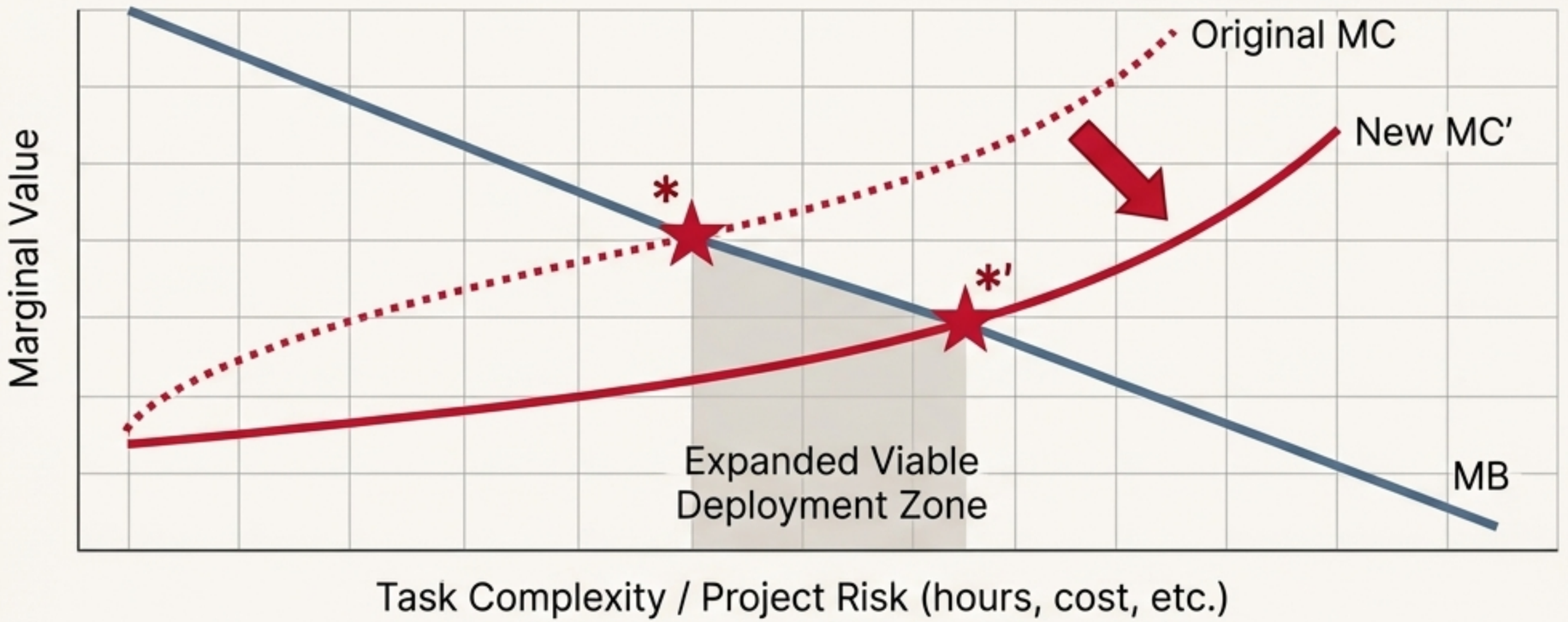
This allows an agent to filter its memory for the entity (**John**) and the topic (**budget**) instead of performing a broad, imprecise keyword search.

The Economic Frontier of Agent Deployment



As task complexity rises, the additional value an AI model adds (MB) falls, while the additional cost or risk (MC) climbs. The intersection marks the optimal point. Deploy AI for tasks to the left of this point, where benefit still outweighs cost (MB > MC).

How New Models Expand the Deployment Frontier.



When a breakthrough model or capability arrives, the marginal cost of tackling complexity decreases. This shifts the MC curve, moving the optimal point to the right. The zone where MB > MC expands, making more complex tasks economically viable for AI automation. Your 'deployment frontier' widens.

The Architect's Dilemma: Build vs. Framework vs. Hosted

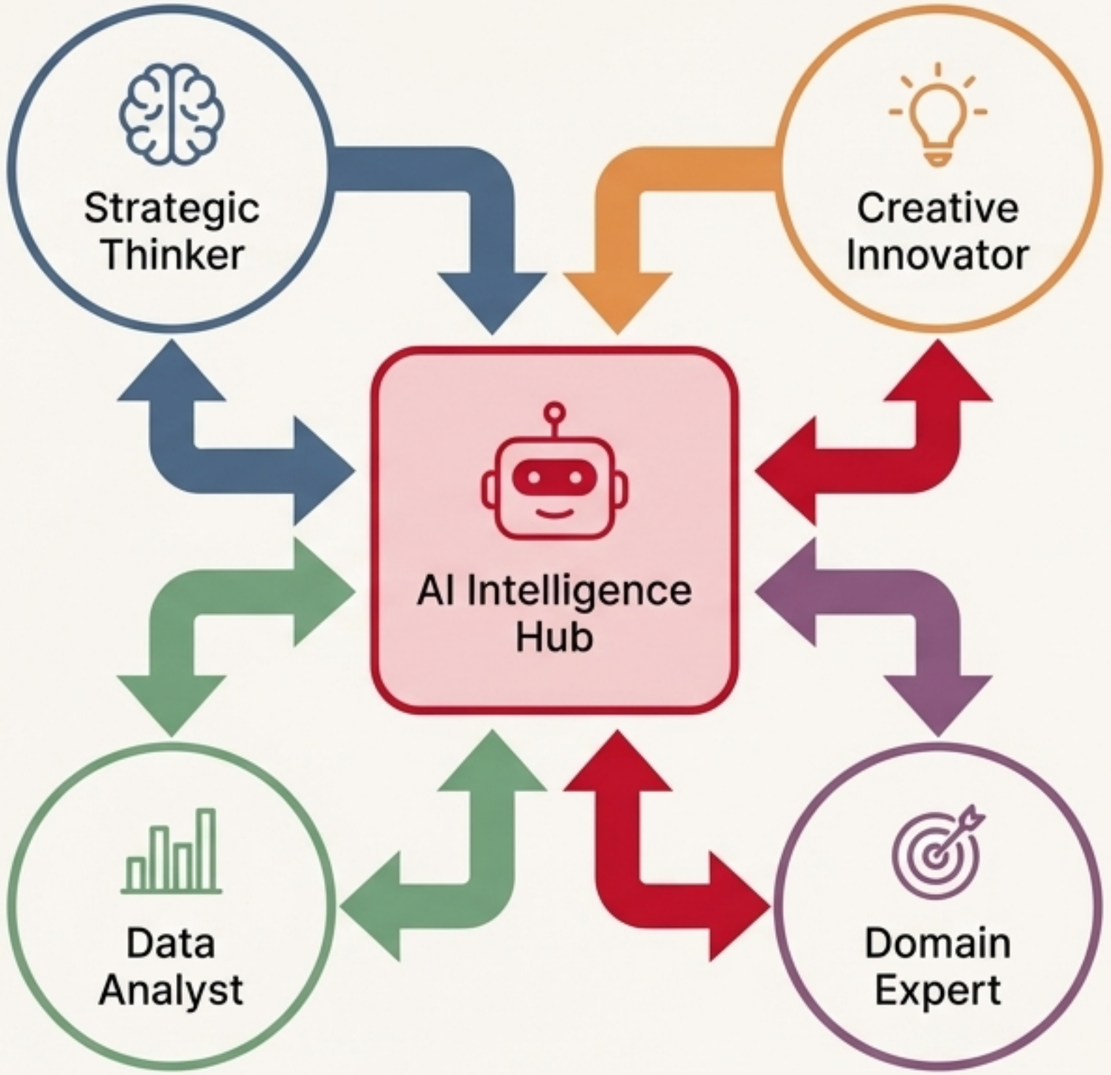
> "If it's a core business function—do it yourself no matter what."
> – *Joel Spolsky, adapted for the AI era*

	Custom Build	Framework	Hosted Solution
Internal Productivity Tool	⚠ Learning opportunity	💡 Consider	✓ Recommended
Consumer-Facing Assistant	💡 High Control & Cost	✓ Balanced Approach	⚠ Limited Customization
Rapid Prototype	⚠ Slowest Time to Market	✓ Speed & Flexibility	✓ Fastest Deployment
Enterprise Solution with Compliance	✓ Full Governance Control	💡 Compliance Complexities	⚠ Data Privacy Concerns

From Individual Agents to Collective Intelligence.

Why limit agent memory to a single user? The next frontier is creating shared memory systems that preserve organisational knowledge beyond any individual's tenure.

Psychologists call this a **Transactive Memory System (TMS):** a group-level system for encoding, storing, and retrieving information. It's about "knowing what other team members know" and being able to access it on demand.



The Power of Shared Knowledge: A Case Study

A 2023 study of a call centre where AI assistants were deployed to support customer interactions demonstrated a dramatic effect. The AI effectively captured and disseminated the expertise of top performers throughout the organisation.

+34%

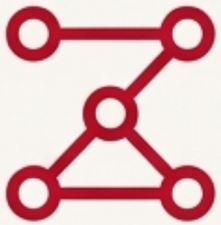
Productivity improvement for novice workers

Minimal gains for experienced workers

Collective memory doesn't just preserve information—it democratises expertise and accelerates learning for the entire team.

Architecting Organisational Memory: Platforms & Protocols.

Several platforms and protocols are emerging, each taking a different approach to building shared, persistent knowledge systems.



Temporal Knowledge Graphs

Creates memory systems that track how team interactions and business data change over time.

Unified Knowledge Search

Connects enterprise applications (Slack, Google Drive, etc.) to create a unified, secure search layer and AI assistant system.

A Decentralised Protocol

Standardises how applications provide context to LLMs, allowing teams to construct custom, evolving knowledge graphs from multiple sources.

The Conductor’s Vision.

As these systems evolve, we must remember a crucial truth.
Like an orchetra, the music may be known, the notes memorised,
but it’s the conductor who has the overarching vision.

“...it will be us, the humans, who will decide on what we want, who will guide the agent not only on what success is but ***also what we want success to be.”***