

The memory layer our AI investment is missing

The first step to automating all of our projects — using organizational, project, team, agent and personal memory.

One shared memory layer. Every AI model our teams already use. Every project, every agent, every person — learning together, and saving a great deal of computing cost and time.

DESIGNED TO WORK WITH

Claude

GPT

Gemini

Grok

DeepSeek

Llama

Mistral

Qwen

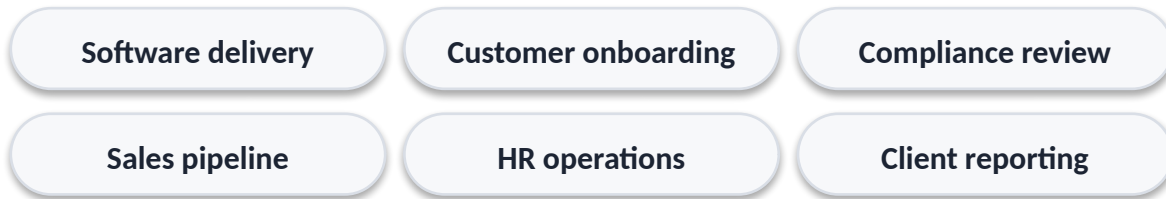
Cohere

Local models

The first step to automating all our projects



The same pattern runs through every project we run



Today vs. with shared memory

Today: every step needs a person to set context, hand work over and repeat explanations.

With shared memory: every step starts with what the organization already knows, so the chain runs with minimal human effort.

The point for management: this is not a tool purchase. It is the base layer that makes every future automation project cheaper, faster and smarter.

Where the cost savings come from

COMPUTING

Fewer tokens per request

Only the relevant memories are retrieved for each request — not the whole history. Model bills go down and answers come back faster.

TIME

No repeated onboarding

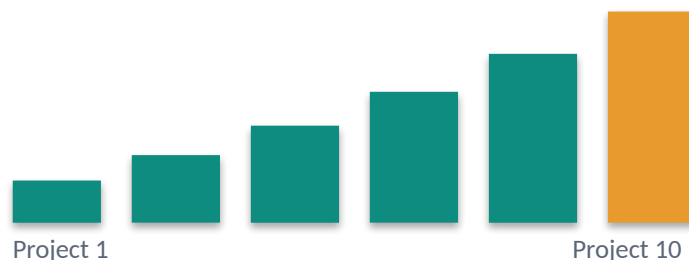
People and agents inherit context automatically. No re-explaining, no re-teaching, no cold starts for new hires or new agents.

REWORK

No duplicated effort

What one team learns, every team can use. A mistake made once is recorded, so it is not repeated in the next project.

Compound interest for AI



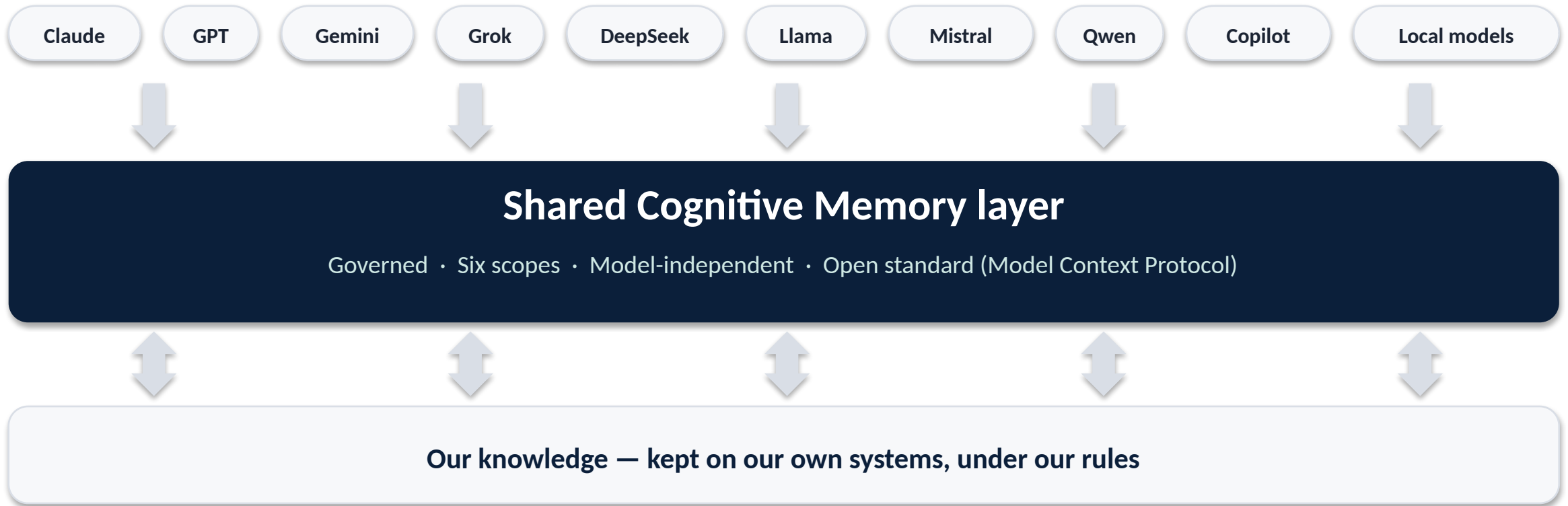
The first project saves some cost. The tenth saves dramatically more, because it inherits everything the first nine learned. Savings grow with every project we run. (Illustration only — the pilot will measure the real numbers.)

Why memory is the foundation for automation

Automation layer	What it needs from memory	Comes from
Workflow automation	The organization's rules and the project's context	Organizational + Project memory
Task automation	The team's conventions and past decisions	Team memory
Agent autonomy	What worked, what failed, what to try next	Agent memory
Personal assistance	Each person's preferences and history	User + Personal memory

Without this layer, every automation project starts from scratch. With it, every automation project starts with everything the organization already knows.

One memory layer. Every AI model our teams already use.



Think of it like this: employees are not forced to use one brand of phone or one email client. Our AI memory should not be locked to one model either.

Why model freedom matters



No vendor lock-in

The memory belongs to us, not to any AI vendor.



Best model for each task

Claude for one job, GPT for another, a local model for sensitive data — same memory underneath.



Easy model switching

When a better model arrives, we swap it in. No migration, no re-teaching.



Our data stays ours

Stored on our own systems with our governance rules. Models are interchangeable users of it.



Lower total cost

With the right context, lower-cost models can do more of the work while quality stays high.



Future-proof

New models arrive every quarter. The memory layer keeps working with all of them.

The uncomfortable truth

We have already invested in AI. Our teams use it daily. Yet most of that intelligence evaporates every time a session ends.

The AI forgets what it learned.

The next session starts from zero.

The next team member repeats the same onboarding.

The next agent makes the same mistake.

We are paying for intelligence and getting amnesia.

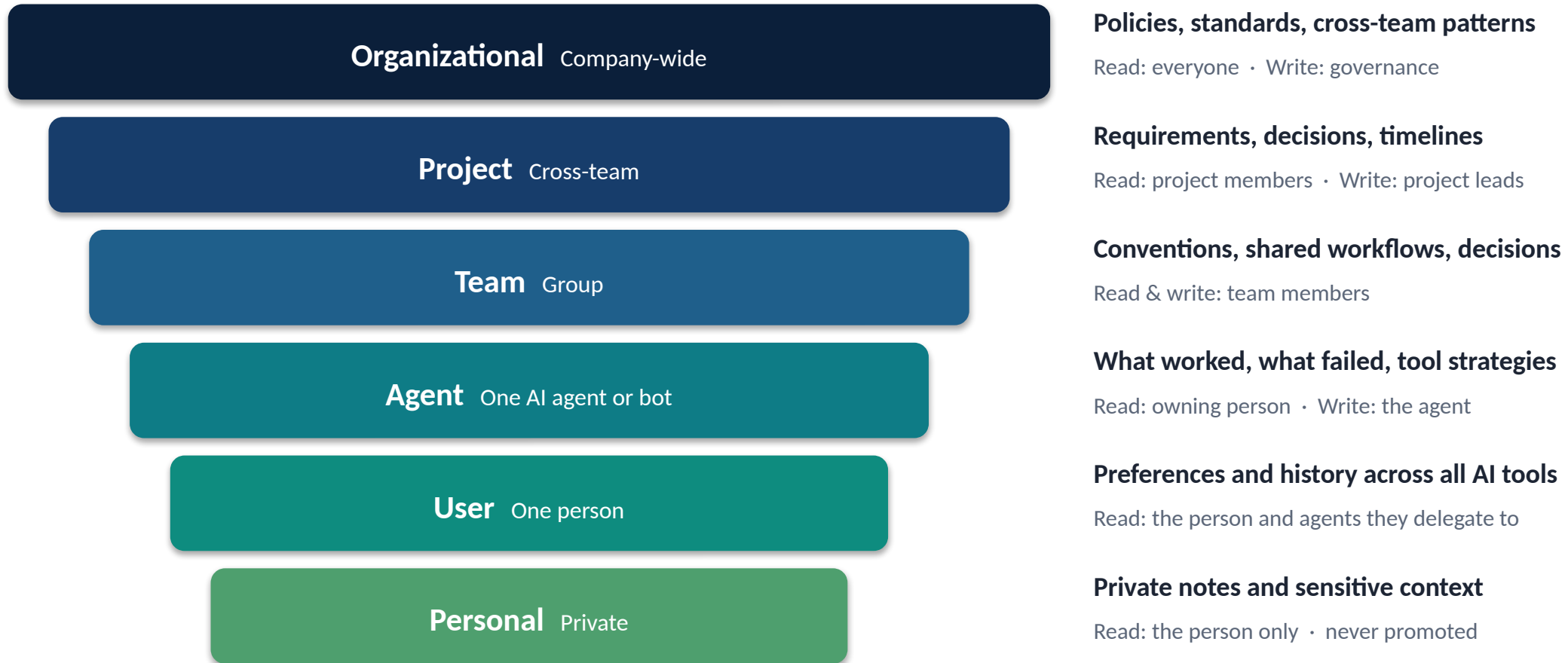
The question is not whether our teams use AI. The question is whether our organization is getting smarter because of it.

What we are actually losing today

What happens now	What it costs us
Each AI session starts cold	Wasted tokens, wasted time, repeated context-setting
Knowledge stays with individuals	No organizational learning; when people leave, their knowledge leaves too
Agents cannot coordinate	Duplicate work, conflicting outputs, inconsistent answers
No record of what the AI knew	Compliance exposure in regulated work
Personal and team data mixed	Privacy risk, data leakage, loss of trust

This is a value-leakage problem, not a technology problem. Our AI models are capable. The missing infrastructure is the bottleneck.

Six memory scopes, one governed flow



Two more levels complete the picture: client memory (walled off per client) and accounts that can manage several organizations.

The real innovation: memory that flows



UPWARD

Promotion

Validated insights move up. A team lead approves each one, and it is made anonymous first. A pattern shared by several people becomes a team convention, then a company standard.



DOWNWARD

Inheritance

Higher scopes shape lower ones automatically. A company policy guides every project; a team convention becomes each member's default. New members inherit context instantly.



SIDEWAYS

Synthesis

Several team memories combine into project memory; several projects combine into organizational memory. Patterns emerge that no single team could see.

Without flow	With flow
Each person re-teaches their AI from scratch	Preferences follow the person to every AI tool
Team lessons never leave the team	Approved patterns become company standards
Compliance has no visibility	Every promotion is logged, anonymous and auditable

How a single insight would travel

A developer discovers an efficient deployment pattern



What this delivers: an insight that used to live and die in one chat window compounds across the company — without leaking private context and without anyone re-teaching the system.

Safety and control, designed in from day one



Private stays private

Personal memory is never promoted or shared.



A person approves

Nothing becomes team or company knowledge without a lead's approval.



Client walls

What we know about one client is walled off from every other client.



Full audit trail

Every read, write and approval is recorded: who, what, when, on whose behalf.



Company sign-in

Company accounts; administrators also use a one-time code.



Sensitive data in-house

Real company data is processed only by AI models on our own servers.



Four-eyes for connections

Linking Jira, databases or mailboxes needs a second person to approve.



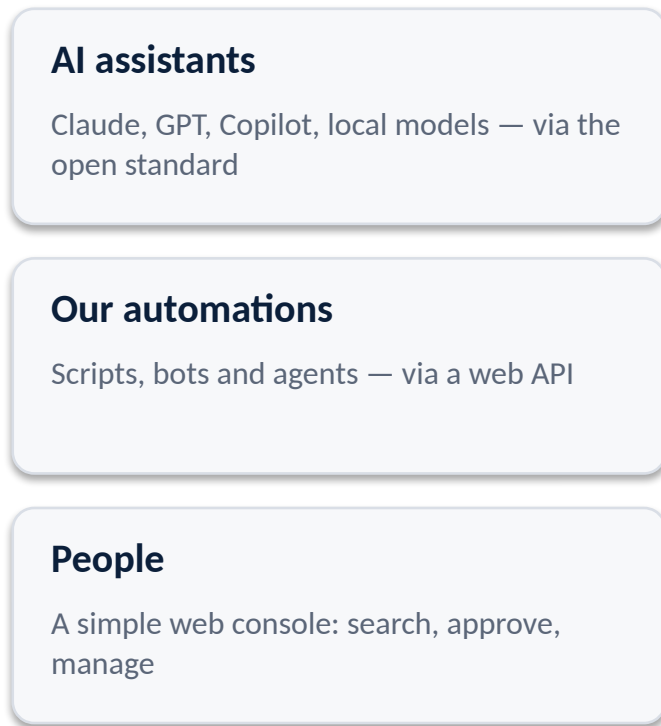
Encrypted and backed up

Secrets encrypted at rest; encrypted daily backups with restore tests.

The pilot runs on our own company's data first; selling to other companies would follow only after a full security and compliance phase.

One layer between our AI tools and our knowledge

WHO USES IT



Shared Cognitive Memory

- 1 Check permission on every request
- 2 Search the right scopes (meaning + keywords)
- 3 Apply flow rules: promote, inherit, combine
- 4 Ask a person to approve where needed
- 5 Record everything in the audit trail

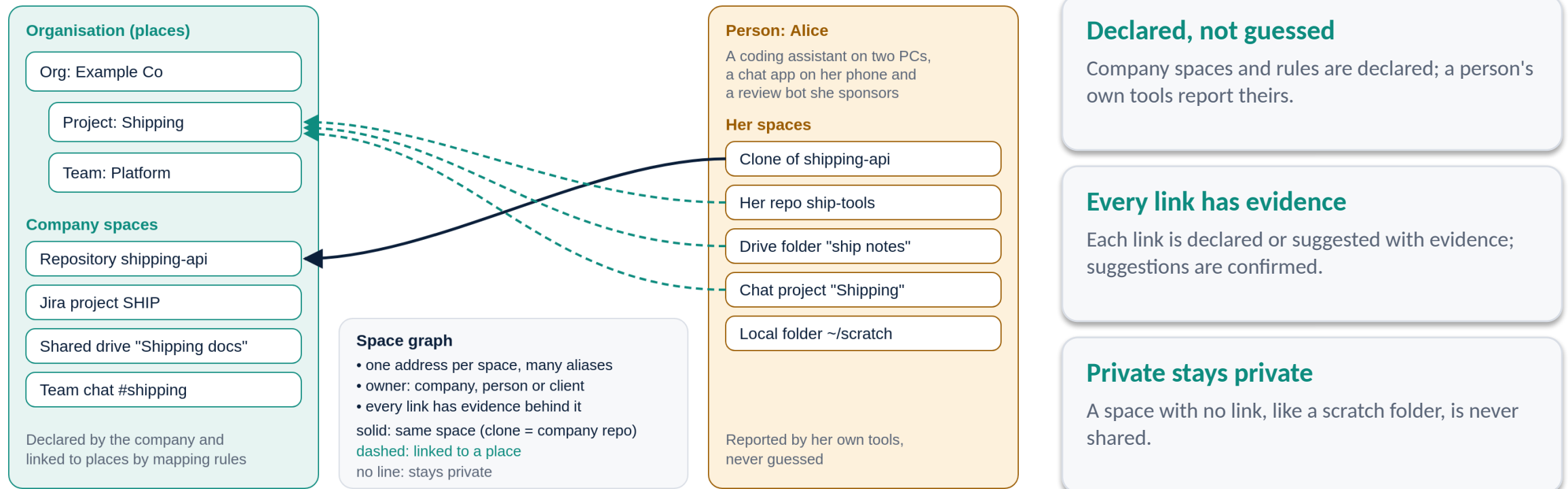
WHAT IT LEARNS FROM



Each source is connected only after a second person approves it.

Built from proven open-source parts — our own servers, our own database, our own sign-in. No dependence on one AI vendor.

The right memory loads wherever a person works



Any AI app, on any device, knows which project, team or client it is working for: her clone is the company repository, and her own repo, drive folder and chat project belong to Shipping.

Illustrative example · spaces registry and mapping: built and in testing on the pilot server

One governed memory instead of many connectors

EVERYONE'S OWN CONNECTORS

- ✗ Raw mail and files reach the AI with no scopes, client walls, approvals or audit
- ✗ Approved knowledge and unapproved raw data disagree, and the AI cannot tell which is which
- ✗ Dozens of tool definitions and large results crowd the AI's working memory
- ✗ Private data, untrusted inbound mail and a 'send' tool end up in one agent

COMPANY SOURCES THROUGH THE MEMORY LAYER

- ✓ Read on our own servers; facts extracted by a small model we run ourselves
- ✓ Shared knowledge only after a person approves it, with client walls and an audit trail
- ✓ The AI gets one context pack within a token budget we set (500 to 100,000 tokens), labelled as data
- ✓ A small, stable set of memory tools, read-only towards outside systems

+3,139 tokens

added to every AI request by six idle mail, drive, calendar and document connectors

~23,400 tokens

when one mail connector's 30 tools load; up to 25,000 more for a single result

-66% • -\$504/month

tokens and list-price cost when connectors are switched on only per task, team of 20 (simulated)

Claude Code test, 13 Oct 2026, plus a simulation • source: "What connectors cost", it-labs.in research

What management would actually get



1 Compounding productivity

Every interaction makes the system smarter. New people and new agents start with what we already know.



2 Lower token and computing cost

Only relevant memory is sent to the model. Less waste, lower bills, faster answers.



3 Compliance and audit readiness

Every memory is logged and attributable. We can show what the AI knew when it acted.



4 Governed AI behavior

Personal stays personal. Team knowledge stays in the team. Clients see only their own.



5 Model freedom

Use any model — commercial or local — on the same governed memory.



6 Competitive advantage

Institutional memory compounds, and competitors cannot copy it overnight.

Why now

01

AI is already here

Our people already use AI, managed or not. The question is whether that use is governed and cumulative, or fragmented and wasted.

02

Regulation is tightening

The EU AI Act, data-protection enforcement and sector rules are moving from guidance to enforcement.

03

The technology is ready

Open standards now let any model connect to the same memory. The storage and security building blocks are mature.

The window to be an early mover is open. It will not stay open.

Is this a real need? The market says yes

2026

All big AI vendors added it

Anthropic (April), then OpenAI, Google and Microsoft. Memory is becoming a basic expectation for enterprise AI.

ANALYSTS

The next battleground

Shared memory is called the layer that decides enterprise AI, and a likely standard requirement in corporate tenders.

EARLY RESULTS

Fewer errors, lower cost

Rakuten reports 97 % fewer first-pass errors at 27 % lower cost once agents remember past mistakes (vendor-reported).

Strong fit

- Many people and AI tools on shared projects
- Many clients that must be kept apart
- Regulated or export-controlled work
- Long programmes and staff turnover

Weak fit

- Very small teams
- Little AI use yet
- Fully tied to one vendor's suite
- Mostly one-off work with nothing to reuse

What exists today — and the gap we would fill

Group	Examples	What is missing for a company-wide memory
Developer memory kits	Mem0, Zep, Letta	No company governance: no scopes, approvals or client walls
Memory inside one AI vendor	Claude, Microsoft Copilot, ChatGPT, Gemini	Locked to that vendor; mostly cloud only; little governed sharing
Company search	Glean, Atlassian Rovo	Finds what is written; does not learn from AI work
Governed context layers	Writer Enterprise Brain, Tabnine	Marketing content or code only; early stage

Our position: any model · governed flow across six scopes · client walls · runs on our own servers · full audit trail.

Risk: big vendors bundle memory for free. Our edge has to be independence and governance — the window is probably 12–18 months.

The cost of waiting

Our AI spending keeps delivering session-level value, not organizational value.

Our teams keep solving the same problems independently.

Our compliance position becomes harder to defend.

Competitors who build shared memory onboard faster, serve customers better and operate cheaper.

Every new AI model becomes another silo instead of another contributor.

The cost of inaction is not zero. It is the compounding advantage we are not building.

A phased build with a go / no-go gate at every step

Step 0	Design and rules	Agree scopes, access and retention rules with a governance partner	done
Phase 1a	Personal, user and team memory	Pilot team uses shared memory across their AI tools	built
Phase 1b-c	Project and agent memory; Jira	Agents improve across sessions; the first connector	built
Phase 1d-f	Client walls, approvals, web console	Lead-approved promotion; knowledge compounds across teams	built
Phase 1g-t	Automatic saving, spaces, digital workers	Memory lands in the right place without people filing it	built
Phase 2	Security, compliance, product	SSO, customer-held keys, pen test, SOC 2 — ready for other companies	after Phase 1

◆ Go / no-go gate: measure time saved and answer quality, then decide. Status 11 October 2026: Phase 1 built and tested on a pilot server; next, its own hardware with a GPU for the local AI model.

Three ways other companies could run it (Phase 2)

A Hosted service

We run it for them

GOOD

Fastest start; we handle updates

TRADE-OFF

Their data sits with us; heavier compliance for us

Start here

B In their cloud

We run it in their cloud account

GOOD

Data stays in their environment; we still operate it

TRADE-OFF

Some setup per customer

C Self-hosted

They run it on their servers

GOOD

Full control; suits strict or offline sites

TRADE-OFF

They need their own operations team

One product, three ways to run it. This decision is for Phase 2; Phase 1 is for our own company.

What success would look like

“ *Our AI remembers our project context without us re-explaining it.*

“ *New team members get up to speed in days, not weeks.*

“ *We have not repeated the same mistake twice.*

“ *We can see exactly what the AI knew and when.*

“ *We switched models last month — and nothing was lost.*

WHAT WE WILL MEASURE

Time for a new member to be productive

Questions asked more than once

Tokens and cost per task

Answer quality, rated by users

Approvals and audit completeness

The goal: the organization is smarter than it was last quarter because of how we use AI.

What we are asking for today

1

Executive sponsorship

Treat memory as a core layer of our AI infrastructure — not a feature of any single tool or vendor.

2

A pilot team

The first team to work with persistent, shared memory across several AI models.

3

A governance partner

Someone from legal, compliance or security to set access and retention rules from day one.

4

A phased budget

Funding for the phased build, with a clear go / no-go gate at every milestone.

Not a leap of faith: a structured experiment with measurable outcomes.

CLOSING

Memory is not a feature. It is the foundation.

We have the models, the teams and the use cases. What we do not yet have is the layer that turns individual AI use into organizational intelligence — one that works with every model we use today and every model that arrives tomorrow.

Learn faster

Automate faster

Switch models freely

Keep what we know

That is the opportunity. That is what we are asking to build.

One-page leave-behind

THE PROBLEM

AI forgets everything between sessions. Our organization is not learning from its own AI use.

THE SOLUTION

A shared memory layer with six scopes and three directions of flow — promotion, inheritance, synthesis — under clear access rules.

MODEL FREEDOM

Works with Claude, GPT, Gemini, Llama, Copilot and local models through an open standard. No vendor lock-in.

THE PAYOFF

The first step to automating all our projects; every automation starts with what we already know.

THE OUTCOME

Compounding productivity, lower compute cost, audit readiness, governed AI behavior, and one governed feed instead of everyone's own connectors.

THE PATH

A phased build — our own company first — with a go / no-go gate at every stage.

THE ASK

A sponsor, a pilot team, a governance partner and a phased budget.

THE RISK OF WAITING

Competitors will build this. We would pay for the same lessons twice.