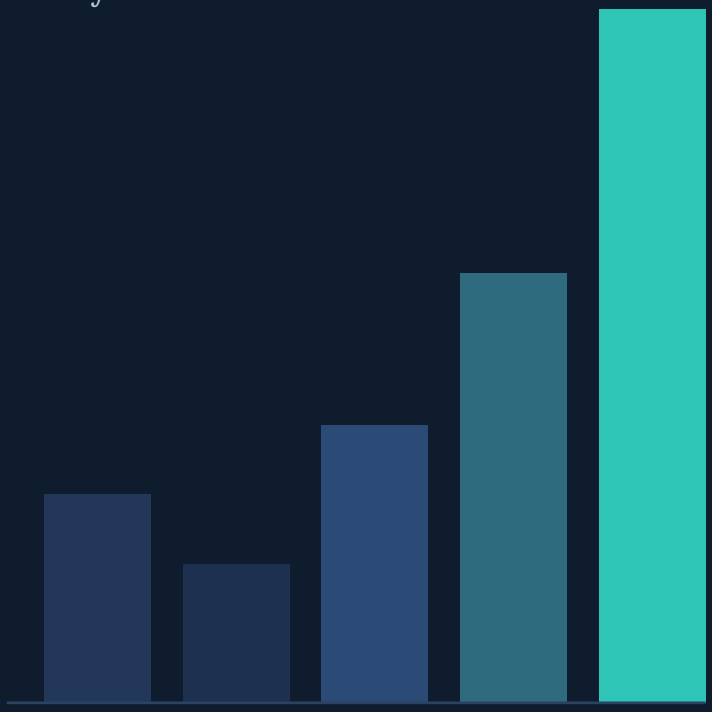
WHITEPAPER

You have AI.
It isn't in your P&L.

Why enterprise AI stalls before it reaches the P&L — the limits of retrieval, what graph memory fixes, and why the interface decides whether any of it pays.



XAMUN TECHNOLOGIES

Position paper · v1.0 · August 2026 · xamun.ai

Research-validated — all sources cited in §10

SUMMARY OF THE ARGUMENT

Enterprise AI adoption is nearly universal; enterprise AI impact is rare. The evidence — MIT, McKinsey, Gartner, and payroll-linked field studies — converges on one diagnosis: the constraint is not model quality but the failure to change how work is done. Retrieval-augmented chat, the default architecture, fails in documented and structural ways, and even graph-enhanced retrieval leaves the burden of knowing what to ask on the user. This paper argues that value arrives only when governed intelligence — a deterministic rule engine paired with a bounded language model over one graph memory — is delivered invisibly inside purpose-built applications, on workflows re-engineered around what AI now makes possible. We call this architecture Two Minds.

IN THIS PAPER

SECTION 1

The scoreboard: adoption without impact

Three independent research programmes, using different methods and samples, arrive at the same place. MIT's Project NANDA reviewed more than 300 enterprise deployments and found that 95% of generative-AI pilots produced no measurable P&L impact [1]. McKinsey's State of AI survey — roughly 1,993 organisations across some 105 countries — found only about 6% of firms attribute more than 5% of EBIT to AI [2]. Gartner forecasts that over 40% of agentic-AI projects will be cancelled by the end of 2027, citing escalating cost, unclear business value and inadequate risk controls [3].

95%
of pilots: no P&L impact
MIT NANDA, 2025 [1]

~6%
attribute real EBIT to AI
McKinsey, 2025 [2]

40%+
agentic projects to be cancelled
Gartner, by end-2027 [3]

An estimated US$30–40 billion has been spent establishing this. The 95% figure has been criticised for a narrow success definition and a small interview base; the criticism targets precision, not direction — McKinsey's far larger sample lands within a point of it. The core finding is not that the models are weak. MIT names the cause the learning gap: tools that retain no organisational context, do not adapt to the workflow they were bought to change, and do not improve with feedback [1].

SECTION 2

You bought a tool, not a change

The modal enterprise AI programme installs a chat assistant beside an unchanged process. The process — which is what actually consumes the cost, the cycle time and the risk — stands exactly where it was, while five new jobs quietly move onto staff: learn to prompt; judge whether the answer is right; find the source and verify it; re-key the result into the real system; absorb the blame when it is wrong. Every one of these

was previously carried by software.

The burden did not disappear. It changed owners.

This is why individual enthusiasm coexists with enterprise disappointment. The same professionals who use consumer chatbots daily describe them as unreliable inside enterprise systems [1] — not because the model changed, but because the stakes did, and the verification tax that is tolerable for a personal draft is intolerable for a regulated decision.

SECTION 3

The limits of retrieval

Retrieval-augmented generation (RAG) — fetch document chunks by vector similarity, hand them to a language model — is the default architecture for referencing internal knowledge. Its failure modes are documented, not mysterious. Barnett et al. studied three production systems and catalogued seven failure points: missing content; missed top-ranked documents; content lost in consolidation; content present but not extracted; wrong output format; wrong specificity; and incomplete answers — the most expensive failure, because a partially right answer passes review [4].

The handoff from retriever to model compounds these. Chroma Research tested eighteen frontier models and found every one degrades as input length grows — on tasks as simple as retrieval and copying, long before the context window is full [5]. Liu et al. showed accuracy follows a U-curve against position: when the relevant passage sits mid-context, performance drops by more than 30%, replicated across six model families [6]. Semantically similar but irrelevant chunks actively mislead, and degradation is sharpest exactly where real business questions live — low similarity between question and evidence [5]. A bigger context window is not a knowledge strategy.

Nor does grounding guarantee truth. Stanford's RegLab and HAI ran the first preregistered evaluation of commercial RAG-based legal research tools — products marketed as "hallucination-free". Lexis+ AI and Westlaw AI-Assisted Research hallucinated on 17% and roughly 33% of queries respectively; the best performer answered 65% of queries accurately [7]. Retrieval reduces hallucination against a bare model. It does not remove it.

In practice, the enterprise RAG assistant often lands below direct use of a frontier model. Retrieval noise actively misleads [5]; a third tool in the Stanford study refused or gave incomplete answers on more than 60% of queries [7]; the corporate deployment is frequently pinned to a weaker model than the public one; and MIT records the behavioural verdict — the same professionals who use consumer chatbots daily describe the enterprise versions as unreliable, and quietly route around them [1]. Many organisations paid to move from level 1 to level 2 of the ladder in Section 6, and went backwards.

18 / 18
models degrade as input grows
Chroma Research [5]

>30%
accuracy loss mid-context
Liu et al., six model families [6]

17–33%
hallucination in commercial legal RAG
Stanford RegLab / HAI [7]

SECTION 4

What graph memory fixes — and where it does not

Chunking severs relationships: an obligation, a precedence between versions, an exposure across counterparties cannot be represented as a similarity score. Structured, graph-based memory restores them, and the evidence for it is specific. On GraphRAG-Bench, graph-structured retrieval beats text chunks by roughly ten points on complex reasoning (53.4 vs 42.9) and thirteen on contextual summarisation (64.4 vs 51.3) — while simple fact lookup is a statistical tie (60.1 vs 60.9) [8]. Across multi-hop benchmarks, graph-based retrieval outperforms dense retrieval by an average of 27 points [9]. In production, LinkedIn's knowledge-graph-enhanced customer-service system improved retrieval accuracy by 77.6% (MRR) and cut median issue-resolution time by 28.6% — from seven hours to five — in a six-month A/B deployment [10].

+77.6%
retrieval accuracy (MRR)
LinkedIn, in production [10]

−28.6%
median resolution time, 7h → 5h
six-month A/B deployment [10]

+27
points on multi-hop QA
graph vs dense retrieval [9]

Honesty requires the other half. A defensible domain graph takes weeks to months of modelling against days for a vector pipeline; on single-hop lookups the structure is overhead the query never needed; and Douwe Kiela, a co-inventor of RAG, argues that GraphRAG as naively built amounts to data augmentation [8]. The graph earns its keep precisely on the relationship questions enterprises actually ask — and it is a memory substrate, not a product. On its own it still leaves a person in front of a text box, deciding what to ask.

SECTION 5

Memory is necessary. The interface decides whether it pays.

Nielsen Norman Group named the obstacle in 2023: the articulation barrier. To get value from a prompt-driven product, a user must know what the system can do, decide what they want from it, and express that want precisely in prose — and missing any one means they type nothing at all [11]. Jakob Nielsen estimates roughly half the population of wealthy countries cannot reliably clear the third hurdle. In NN/g's study of Amazon's Rufus assistant, not one test participant noticed the AI feature despite its placement in global navigation — the existing mental model won [11].

The field data shows what this costs. Humlum and Vestergaard matched AI-adoption surveys for 25,000 workers across 7,000 Danish workplaces to actual payroll records: average time savings of 2.8% of work hours, and "no significant impact on earnings or recorded hours in any occupation", with only 3–7% of the productivity gain reaching pay [12]. METR's randomised controlled trial adds the perception gap that keeps such programmes funded: experienced developers using AI tools took 19% longer on real tasks while believing they had been 20% faster; METR's 2026 follow-up finds genuine speed-ups only after the same developers had absorbed a year of learning the tools [13].

2.8%
of work hours saved
Denmark, payroll-linked [12]

0
effect on earnings or hours
in any occupation studied [12]

19%
slower — while feeling 20% faster
METR RCT, early-2025 tools [13]

Minutes saved by individuals are not enterprise value. Without a process redesigned to collect them, they evaporate.

SECTION 6

The maturity ladder

Five levels describe how organisations put internal knowledge to work with AI. The first three change how knowledge is stored — and the field evidence above applies to all of them alike. The step change comes when intelligence enters the workflow (level 4), and again when the workflow itself is re-engineered around what AI makes possible (level 5). Level 2 is drawn below level 1 deliberately: bolting noisy retrieval onto chat frequently degrades the experience of simply using a frontier model directly, and staff respond by routing around the corporate assistant (Section 3). The gap between levels 3 and 4 is the divide MIT's research names. Level 5 is drawn far above level 4 deliberately: level 4's gains are capped by the shape of the old process, while level 5's are bounded only by how far the work can be redesigned — which is why workflow redesign is the strongest EBIT correlate McKinsey found [2]. Bar heights remain qualitative: the shape is the claim, not the scale.

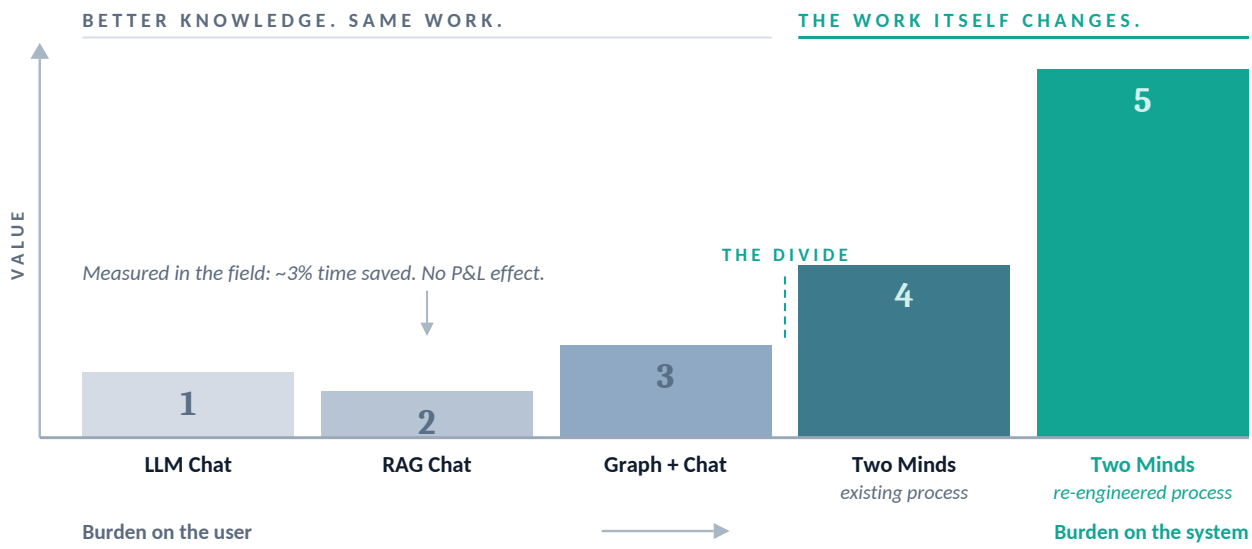


Figure 1 — The Xamun maturity ladder. Levels 1–3 plateau — and level 2 frequently lands below level 1 in practice. Level 4 embeds intelligence in the existing process; level 5, where the process itself is redesigned, is where the value concentrates.

SECTION 7

Two Minds: place generation, don't patch it

Hallucination is a feature. Just never in your ledger.

Generation is what a language model is for: the mechanism that drafts a good clause is the same mechanism that invents a case citation. The Stanford results show what happens when vendors try to suppress it and claim success [7]. Our architecture takes the opposite position: the generative behaviour is not patched, it is placed — confined to the work only a language model can do, and structurally unable to touch anything that must be provably right.

YOUR APPLICATION — the only thing anyone touches. No prompt. No chat window.

Deterministic Mind
Rules · entitlements · calculations · compliance · audit
Holds the ledger. Never a language model.

Governed LLM
Language · extraction · drafting · explanation
Imagines where imagination helps. Never the last word.

GRAPH MEMORY One model of the business — entities, obligations, precedence, lineage — read and written by both minds.

Figure 2 — The Two Minds architecture. The deterministic mind runs first and has the final say; the governed model is never the last word on a decision.

The user touches neither mind directly. Intelligence arrives through the screens, queues, exceptions, approvals and alerts of a normal business application — no prompt, no chat window, no prompt-engineering skill required. The system raises the exception before anyone asks; the recommendation surfaces in context with its evidence chain; behaviour is specified once, in the build, not typed daily by whoever happens to be articulate.

SECTION 8

What actually correlates with EBIT

McKinsey tested roughly 25 organisational attributes against reported EBIT impact from generative AI. Fundamental workflow redesign ranked first — ahead of talent, tooling, budget and governance structure [2]. Yet only 21% of adopters have fundamentally redesigned any workflow: nearly four in five are layering AI onto a process they have not changed. High performers are about three times more likely to have redesigned workflows, and 3.6 times more likely to pursue transformation rather than efficiency alone [2].

#1
workflow redesign, of ~25 tested
strongest EBIT correlate [2]

21%
of adopters redesigned any workflow
nearly 4 in 5 have not [2]

3.6×
more likely: transformation intent
high performers, vs efficiency [2]

This is the empirical case for level 5. Bolting a chatbot onto an unchanged process and teaching the staff to prompt adds a skill requirement, a verification tax and a failure surface while the process stands still — and then waits for efficiency. On the evidence, this is the modal enterprise AI programme. It is also the modal failure.

SECTION 9

Where to start

Pick one workflow — the one costing the most time and risk — and map it as it runs today. Measure the baseline in week one, before anything is built, on the three numbers a board already tracks: cycle time (per

case, end to end, not per task), error rate (counted against the governing rules, not against opinion), and throughput (same team, with no added headcount in the comparison). A baseline measured before the build is the only defence against the perception gap Section 5 documents.

Then re-engineer rather than decorate. If the process map is unchanged after deployment, the return case is fiction; the workflow itself — its triggers, its handoffs, its decision points — must be re-shaped around what governed intelligence makes possible. We publish no projected figures in this paper deliberately: quoting a number before a baseline exists is precisely the behaviour Sections 1–5 document. What moves, and how far, is a property of the workflow — and it is measured against that workflow's own baseline, not against a vendor's slide.

SECTION 10

References

[1] MIT Project NANDA. The GenAI Divide: State of AI in Business 2025. 300+ deployments, 52 executive interviews, 153 survey responses. 2025.

[2] McKinsey & Company. The State of AI: agents, innovation and transformation. n ≈ 1,993 across ~105 countries; with the March 2025 EBIT correlation analysis (Johnson's Relative Weights, $R^2 = 0.20$). 2025.

[3] Gartner, Inc. Press release: over 40% of agentic AI projects will be cancelled by end-2027. Poll of 3,412 respondents, January 2025. June 2025.

[4] Barnett, S., Kurniawan, S., Thudumu, S., Brannelly, Z., Abdelrazek, M. Seven Failure Points When Engineering a Retrieval Augmented Generation System. Deakin University / ICSE-AI 2024. arXiv:2401.05856.

[5] Hong, K., Troynikov, A., Huber, J. Context Rot: How Increasing Input Tokens Impacts LLM Performance. Chroma Research, 2025. 18 frontier models.

[6] Liu, N. F. et al. Lost in the Middle: How Language Models Use Long Contexts. TACL, 2024.

[7] Magesh, V. et al. Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. Stanford RegLab / HAI. arXiv:2405.20362.

[8] GraphRAG-Bench. Accuracy by task type: where graph structure provides measurable benefit — and where it does not. ICLR 2026.

[9] Do We Still Need GraphRAG? Benchmarking RAG and GraphRAG for Agentic Search Systems. 2026. arXiv:2604.09666.

[10] Xu, Z. et al. Retrieval-Augmented Generation with Knowledge Graphs for Customer Service Question Answering. LinkedIn / SIGIR 2024. arXiv:2404.17723.

[11] Nielsen Norman Group. The Articulation Barrier: Prompt-Driven AI UX Hurts Usability (2023); hybrid-interface research and the Amazon Rufus discoverability study (2023–24).

[12] Humlum, A., Vestergaard, E. Large Language Models, Small Labor Market Effects. NBER Working Paper 33777, 2025.

[13] METR. Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity. RCT, 16 developers, 246 tasks; 2025, revised February 2026.

ABOUT XAMUN TECHNOLOGIES

Xamun builds vertical AI operating systems on the Two Minds architecture: a deterministic rule engine paired with a governed language model over one graph memory, delivered inside purpose-built applications on re-engineered workflows. The placement mechanics — how the deterministic mind intercepts, binds and audits the model's output — are the subject of Xamun's implementation work and outside this paper's scope. Xamun operates from Dubai and Manila. Enquiries: xamun.ai.

Figures reflect the studies as published; where findings have been revised or contested, this paper says so in the body rather than in a footnote. Correspondence: xamun.ai. *Version 1.0, August 2026.*