The AI Hype Cycle, Cloud LLM Economics, and the Knowledge Graph Exit.

Why agentic AI is still cresting the Peak of Inflated Expectations while core cloud LLMs are already in the Trough of Disillusionment — and why enterprise context, not more tokens, is the missing layer.

95%
GenAI pilots with no measurable P&L impact

40%+
Agentic projects forecast cancelled by end of 2027

17% vs 60%+
Deployed agents today vs. intent within two years

80–97%
Token reduction range in published GraphRAG work

Inside this brief

Prepared as a working brief for enterprise architecture, data, and AI leadership.

Print the infographic page at 17 × 11 in. · Remaining pages are US Letter.

The AI Hype Cycle vs. the Cloud LLM Economy

Why slide-deck magic is stalling inside real enterprise data architectures — and why context, not more tokens, is the missing layer.

95% GenAI pilots with no measurable P&L

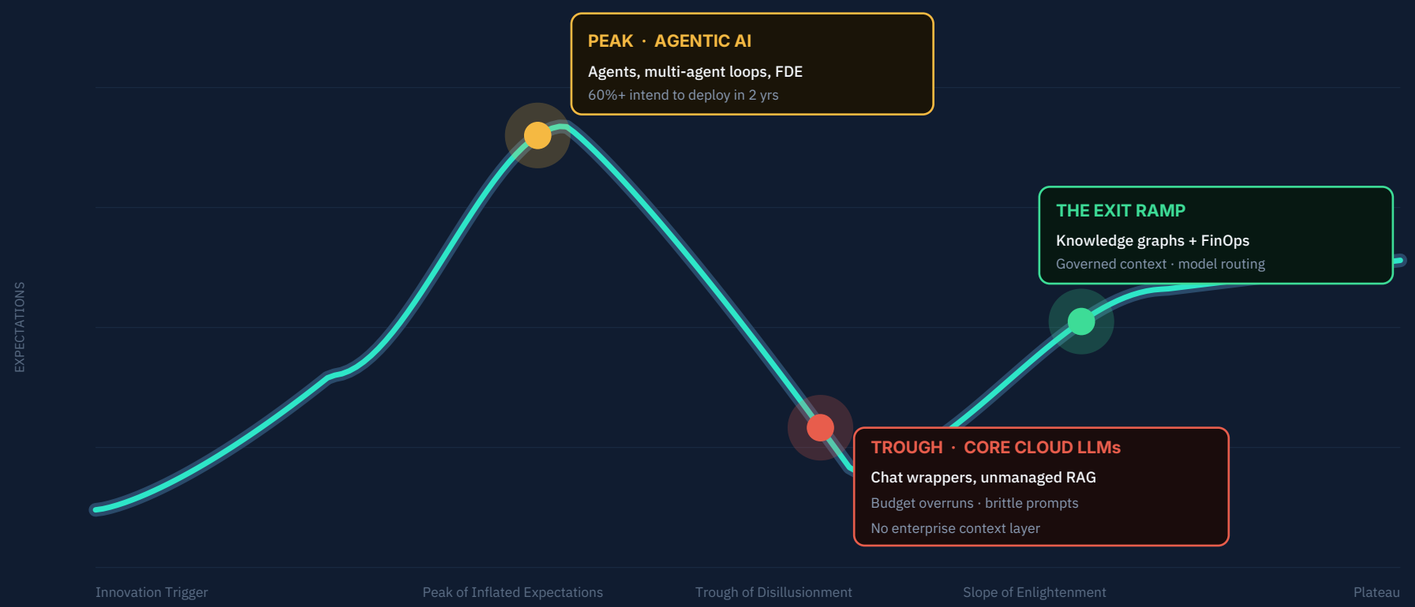
40%+ agentic projects cancelled by 2027

17% enterprises have deployed agents

93% orgs exceed their AI budgets

ADAPTED GARTNER HYPE CYCLE · GENAI + AGENTIC AI · MID-2026

Two technologies, two different points on the same curve



Core foundational models are sliding into the trough while agentic workflows are still cresting the peak. The collision is happening inside the enterprise stack.

THE HYPE PEAK

Agentic AI & custom wrappers

- **Autonomous agents**
  Promised end-to-end workflow automation with multi-agent loops that plan, tool-call, and act.
- **Forward-deployed squads**
  FDE teams wrapping flagship cloud models around each process — fast demos, fragile production.
- **Agent washing**
  Gartner: only ~130 of thousands of "agentic" vendors are the real thing. The rest is rebranded RPA.
- **Intention ≠ installation**
  17% deployed vs. 60%+ intending to. The installed base is years behind the narrative.

THE REALITY CHECK

Governance, cost, and data foundations

- **Token economics broke the business case**
  Agents resend long-lived context every step. Agentic tasks can burn ~1,000× the tokens of a chat turn. Price-per-token fell; total spend tripled.
- **No enterprise context layer**
  Cloud LLMs can read rows. They do not know how your company books revenue, who owns a policy, or which version is still valid.
- **Tech debt from bolted-on wrappers**
  Gartner: ≥50% of GenAI projects overrun budget from poor architecture. Most custom-model efforts get abandoned.
- **Governance is the hidden work**
  93% of enterprises hit permission/access issues. Teams spend ~77% of AI engineering hours on governance workarounds.

WHY THE CLOUD LLM ECONOMY IS NOT DELIVERING

The bill is a context bill — not a model bill

01 Stateless models, stateful business
Every turn re-ships history. Context rot and "lost in the middle" make long windows both expensive and inaccurate.

02 Vector RAG ≠ business understanding
Similarity search has no notion of ownership, authority, time, or permission. Demos work. Production does not.

03 Silos + over-sharing
Disconnected CRM, ERP, tickets, and docs. AI inherits years of access sprawl and invents confident answers across the gaps.

04 Wrong unit of value
Orgs bought tokens and copilots. Value only appears when workflows, cost-to-serve, and decisions actually change.

THE MISSING LAYER

Enterprise context via knowledge graph

A knowledge graph encodes entities, relationships, ownership, policy, and time — the map the LLM never had. Hybrid GraphRAG + governed retrieval is how slide-deck agents survive contact with a real architecture.

- **Grounded reasoning**
  Multi-hop answers over how the business actually works, not nearest-neighbor chunks.
- **Token discipline**
  Graph retrieval has cut token use 80–97% vs. naive RAG in published GraphRAG work.
- **Permissions & provenance**
  Who owns it, who can see it, which version is valid — encoded, not hoped for.
- **FinOps + routing**
  Smaller, cheaper models on a rich context layer beat flagship models on a pile of PDFs.

Sources: Gartner Hype Cycle for Agentic AI / GenAI 2026; Gartner cancellation forecast (40%+ agentic projects by EOY 2027); MIT Project NANDA (~95% of GenAI pilots, no P&L impact); McKinsey agentic economics (93% over AI budget); GraphRAG / enterprise KG literature. Figures are directional, not a substitute for primary research.

THE PATTERN

Are we seeing that friction? Yes — and it has a pattern.

This brief does not sit in anyone's steering committee. It does sit on the public evidence and on the shape of enterprise modernization programs in 2026. Those two sources line up tightly with the claim that we are straddling two points on the same hype cycle at once.

17%
Orgs that have actually deployed agents

60%+
Intend to deploy agents within 2 years

40%+
Agentic projects cancelled by EOY 2027

93%
Organizations exceeding AI budgets

Two clocks are running at once

Gartner's 2026 Hype Cycle for Agentic AI still places agentic AI at the Peak of Inflated Expectations. Only about 17% of organizations have actually deployed agents, while more than 60% say they will within two years. That gap *is* the inflated peak: intention running years ahead of the installed base. Gartner also notes that only around 130 of the thousands of vendors claiming "agentic AI" are the real thing. The rest is agent washing — rebranded automation and chatbots.

Meanwhile, GenAI-as-chat-wrapper and unmanaged RAG have already burned enough budget to enter the Trough of Disillusionment. Core foundational models are no longer the slide-deck magic. They are an operating cost that finance can see. The collision is not happening on vendor keynotes. It is happening inside enterprise data architecture — where token meters, permission systems, and brittle prompt layers meet production workflows.

The cancellation wave is priced in

Gartner's working forecast is that more than 40% of agentic AI projects will be cancelled by the end of 2027 — on costs, unclear business value, and weak risk controls. MIT's Project NANDA is the directional P&L number everyone quotes: on the order of 95% of enterprise generative-AI pilots showing no measurable profit-and-loss impact. Treat the exact percentage as directional. The qualitative finding is robust across McKinsey, Domino, Deloitte, and Gartner: capability is up; attributable returns are not.

Gartner's own Generative AI cycle is blunt about the middle layer. At least half of GenAI projects will overrun budget because of poor architectural choices and lack of operational know-how. Most organizations that try to build custom models will abandon the effort on cost, complexity, and technical debt. None of the 30 GenAI technologies Gartner mapped in 2026 had reached the Plateau of Productivity.

The spend curve did not follow the price curve

Token unit prices collapsed. Enterprise LLM *spend* still tripled. McKinsey reports that roughly 93% of organizations exceed their AI budgets, and about one-fifth have already constrained use because of operating cost. Agentic workloads are the amplifier. Stateless models resend long-lived context at every step. Agentic tasks can consume on the order of 1,000× the tokens of a single-turn chat. That is why FinOps for AI, model routing, and eval harnesses showed up in the conversation so fast.

THE BILL IS A CONTEXT BILL

Tokens are not value. Tokens are the invoice. Every extra noisy chunk retrieved by naive RAG is both a cost and a hallucination risk. Until context is treated as an operating asset — governed, typed, permissioned, and reusable across agents — the cloud LLM meter will keep outrunning the business case.

Three stacked failures — not a model failure

The friction inside modernization programs is not “the models don’t work.” Flagship models reason fluently. The failures stack underneath them.

01
ARCHITECTURE
Wrappers on systems never built for loops

02
CONTEXT
Access without meaning. No as-of. No owner.

03
ECONOMICS
Chatbot case. Always-on invoice.

1. Architecture debt. Wrappers bolted onto systems never designed for tool-calling loops. Multi-agent demos assume clean APIs, idempotent actions, and a human who will catch the miss. Production systems have none of those by default. Forward-deployed engineering squads can make a process look automated in six weeks. They also leave behind a prompt-and-glue estate that finance will later call technical debt.

2. Data and context debt. The model can read the warehouse and still not know how *this* company books revenue, which policy supersedes which, or who is allowed to see the answer. Connectors provide access. Models provide reasoning. Neither provides an understanding of how the business works unless that meaning is made explicit.

3. Economic debt. No unit economics that survive contact with production query volume. Proof-of-concept token forecasts assume short prompts and a handful of users. Production agents carry conversation state, tool traces, retrieved chunks, and retries. The original business case was written for a chatbot. The invoice is written for an always-on worker.

THE ECONOMY

Why the cloud LLM economy is not delivering

This is the part enterprises keep misdiagnosing as a model problem. Buying a more capable foundation model does not fix a missing semantic contract. In several controlled enterprise tests, the most expensive model configuration has been the worst end-to-end system — because extra reasoning was spent cleaning messy context instead of completing the job.

Cloud LLMs sell reasoning. Enterprises need meaning.

Pointing a GPT-class model at a lakehouse answers "what do these rows look like." It does not answer "how does *our* operating margin work in EMEA under *our* recognition rules." Pattern completion sounds confident either way. That is an expensive way to be wrong. Durable business context is no longer optional. It is what keeps enterprise AI affordable and its answers trustworthy.

Vector RAG was the demo layer, not the production layer

Similarity search has no native concept of ownership, authority, temporal validity, or permission. A vector index will retrieve a 2021 memo, a 2023 draft, and a 2026 standard with similar scores if they share vocabulary. That is context collapse. It is also why 93% of enterprises hit governance or access problems in AI programs, and why teams report spending on the order of 77% of AI engineering hours on governance workarounds and access-control repair rather than features. Flat retrieval cannot represent "who owns this, who can see it, and which version is still in force."

Agentic workflows make retrieval worse, not better

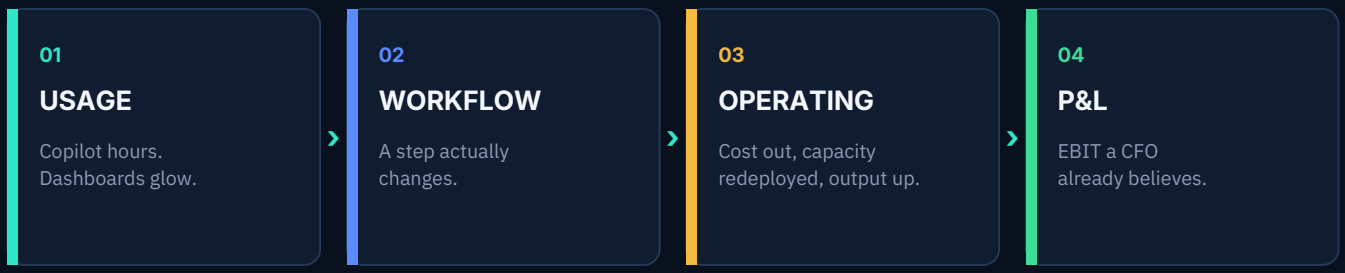
Agents issue orders of magnitude more data requests than a human analyst. Pipelines built for one-shot RAG choke on that volume. Every extra noisy chunk is both a token and a quality defect — the lost-in-the-middle problem is well documented. Context architecture, a live governed map the agent can query at runtime, is replacing "stuff the window." Teams still optimizing chunk size as their primary strategy are solving last year's problem.

Cloud cost models assumed SaaS traffic, not inference chains

Public-cloud LLM meters were designed for experimentation: unpredictable but modest chat traffic, linear-ish cost curves, and data that could live wherever the region happened to be. Agentic production breaks all three. Token spikes do not look like SaaS MAU. Vector-store bills never appeared in the original business case. Data-sovereignty rules constrain where context can sit. Latency kills multi-step agent loops. That combination is why a large share of enterprises are evaluating or executing AI workload repatriation — not because the model is bad, but because the economic and control model of "everything through the public API" does not survive proprietary process knowledge.

Value was defined as usage

Hours “saved” in a copilot dashboard do not become EBIT unless cost is removed, capacity is redeployed, or output rises against a steady cost base. Between “people are using AI” and “the business is measurably better off” sit four links — usage, workflow effect, operating effect, and financial effect — plus attribution. Most programs instrument the first link only. That last-mile measurement gap is why so many organizations cannot prove ROI even when adoption charts look healthy. Enterprise AI spend has grown into the tens of billions; the share of organizations that can attribute EBIT impact remains a minority, and most of those impacts are small.



Most programs instrument only 01. Attribution lives in 04. The gap is why adoption charts look healthy and the P&L does not.

WHAT THIS MEANS FOR A MODERNIZATION PROGRAM

Stop asking “which model?” first. Ask “what meaning must be true for this agent to be allowed to act?” If that meaning lives only in people’s heads, SharePoint, and conflicting dashboards, no cloud LLM will invent it cheaply or safely.

THE EXIT RAMP

Why a knowledge graph is the actual exit ramp

A knowledge graph is not a prettier index. It is a semantic contract: the map the LLM never had. It encodes the objects the business already has names for, the relationships RAG cannot infer reliably, the time and validity rules that stop 2021 from impersonating 2026, and the permissions that make retrieval safe enough for an agent to act rather than merely chat.

The contract, in four clauses

01 ENTITIES

Account, policy, SKU, control, vendor, legal entity, region. If teams already argue the definition, it belongs here.

02 RELATIONSHIPS

Owns, supersedes, constrains, approved-by, booked-to. Multi-hop paths RAG treats as coincidences of vocabulary.

03 TIME & VALIDITY

Effective-from, as-of, superseded-by. A retired memo cannot impersonate the live standard across a dead edge.

04 PERMISSIONS

Who owns the node, who may see it, which system attested it. Retrieval becomes a query, not a similarity lottery.

That is why GraphRAG and context-graph work keep showing large token reductions — published figures in the 80–97% range versus naive RAG on some tasks — and better multi-hop accuracy. You stop shipping the whole haystack into the context window and start handing the model the subgraph that actually answers the question. Smaller routed models on a rich graph routinely beat flagship models on a pile of PDFs. That is the FinOps story hiding inside the architecture story.

THE HONEST CAVEAT

Graphs are not free. Ontology work, entity resolution, curation, and freshness are real cost. GraphRAG implementations often cost 3–5× baseline RAG to stand up. Gartner still places GraphRAG years from mainstream. Pure document Q&A does not need a graph. Relationship-heavy, permission-sensitive, multi-system enterprise work does. Hybrid — graph first for structure, vectors for evidence, memory for session continuity — is what is surviving production. Adoption has been flatter than the hype; the constraint is operating model, not theory.

SOLUTION

A focused Enterprise Knowledge Graph + GraphRAG design

Do not buy “a graph database” and declare context solved. An enterprise knowledge graph is semantic infrastructure. GraphRAG is one consumer of that infrastructure. Agents are another. If you invert the order — agents first, graph later — you will rebuild the same brittle wrapper estate with Cypher in the comments.

Five-layer reference architecture

Treat the platform as five layers with clear owners. Layers 1, 3, and 5 are mostly known enterprise technology. Layer 2 is where programs stall. Layer 4 is where the LLM finally earns its keep. The product sequence is the same four movements the platform is built around.

01	02	03	04
CAPTURE	**CONNECT**	**ACTIVATE**	**TRUST**
Wrangle the mess	Build the graph	Ground the agents	Prove the answers

01 Source & quality
CDC and connectors from systems of record. Quality rules, classification, PII tags, and lineage before anything is allowed to become a node.

02 Ontology & contract
A bounded domain schema: allowed entity types, relation types, attributes, and validity. The schema is the control surface for extraction, retrieval, and review.

03 Graph & evidence store
Property or RDF graph for entities and edges. Vector index attached to evidence spans. Identity resolution. Versioned named graphs per domain or business unit.

04 GraphRAG runtime
Planner chooses graph traversal, vector evidence, or both. Retriever returns a typed subgraph plus citations. Validator checks grounding, policy, and schema types before the answer leaves.

05 Consumption
Copilots, agents, MCP tools, and APIs share one semantic foundation so “operating margin” means the same thing in finance, ops, and the agent.

RUNTIME

How a production query should actually run

Naive RAG: embed the question, fetch nearest chunks, stuff the window, hope. Enterprise GraphRAG is a short, inspectable loop. Bind the asker first. Walk the graph second. Cite and authorize before anything acts.

1. **Bind identity and purpose**
   Who is asking, from which system, with which entitlements, against which as-of date? This is not a preface. It constrains every hop that follows.

2. **Classify intent against the schema**
   Map the question onto entity and relation types the ontology already knows. “EMEA operating margin last quarter” is Region, Ledger, Period, RecognitionPolicy — not text that mentions margin.

3. **Traverse first, retrieve second**
   Walk the live edges: Region → LegalEntity → BookedTo → Account → GovernedBy → Policy (valid in period). Only then pull evidence spans. The subgraph is the context. The documents are the footnotes.

4. **Plan tools, don’t dump tokens**
   A planner chooses graph query, vector search, metric service, or human escalation per sub-question. Schema-guided retrieval is where published systems show accuracy gains and large token cuts.

5. **Validate before act**
   Check citations, policy, and type constraints. If the agent may act, the graph must supply the authority edge — not a similar paragraph.

NAIVE RAG

Embed question → nearest chunks →
stuff the window → hope.
No identity. No as-of. No authority edge.

PRODUCTION LOOP

Bind → classify → traverse → plan → validate.
Subgraph is the context. Documents are footnotes.
Act only if the graph supplies the edge.

WHAT TO BUILD

What belongs on the graph — and what does not

The fastest way to recreate the trough is to graph everything. Start with the objects that already cause arguments in operating meetings. Those arguments are the ontology.

Graph these first

+ **Canonical business objects**
  Customer / legal entity, product and SKU, policy and control, contract and obligation, vendor, asset, employee role, region, book, and cost center. One ID per object, many source keys mapped to it.

+ **Governing relationships**
  Owns, reports-to, approved-by, supersedes, constrains, billed-to, served-from, in-force-for. These are the paths agents get wrong when they only have chunks.

+ **Validity and authority**
  Effective dating, status, classification, system of record, and the human or role accountable for the node. Without this, GraphRAG is just RAG with extra hops.

+ **Permission attributes on the node**
  Enforce at query time — RBAC is not enough once agents traverse. Attribute-level rules on the graph boundary are what scale in regulated environments.

Do not graph these — yet

– **Raw document corpora as the graph**
  Minutes, tickets, and PDFs are evidence attached to nodes, not nodes themselves. Community summaries can sit above the graph. They should not replace typed entities.

– **Every edge an LLM can imagine**
  Open extraction without a seed schema produces ontology sprawl and 73–94% entity-resolution accuracy depending on industry. That is not a production control plane. Allow proposed types; do not auto-commit them.

– **A second copy of the warehouse**
  Virtual graphs and federation exist so you do not ETL the universe. Materialize the semantic spine. Leave high-volume facts where they already live, and join at query time.

– **An agent platform as the source of truth**
  Memory stores and prompt caches are session state. They go stale. The graph is the shared, governed meaning. Agents consume it. They do not own it.

Hybrid is the production pattern

Graph for structure, multi-hop paths, permissions, and validity. **Vectors** for unstructured evidence attached to those nodes. **Memory** for the working set of a session or an agent run. **Metrics services** for numbers that already have a semantic layer in BI. The failure mode is almost never “we picked the wrong one of these.” It is that the data flowing into all of them is ungoverned, stale, or semantically sloppy.

HOW TO LAND IT

Operating model and a 90-day path

Technology is roughly 20% of the value. The rest is how work, ownership, and measurement change. A GraphRAG program that has no curator, no freshness SLO, and no kill criteria is another pilot.

Who owns what

Domain owners approve entity definitions and in-force relationships. **Knowledge engineering** maintains the schema, resolution rules, and review queue. **Data platform** runs connectors, quality, and the graph/vector stores. **AI platform** runs the planner, routing, evals, and spend controls. **Risk / security** owns the policy attributes on the graph boundary. If those five seats are one hero engineer, the graph will rot the week they go on leave.

90 days, one high-value slice — not a corporate ontology

Days 1–15 · Bind the question

Pick one workflow where a wrong answer has a cost: invoice exception, policy eligibility, customer entitlement, control mapping. Write ten gold questions the graph must answer. Name the 15–30 entity and relation types those questions require. Cap the schema on purpose.

Days 16–40 · Stand up the spine

Resolve identity for those objects from two or three systems of record. Attach evidence spans. Stamp classification and owner. Turn on query-time permission filters before any LLM sees a node. Measure entity-resolution precision weekly; do not proceed below an agreed bar.

Days 41–65 · GraphRAG against gold questions

Planner + graph-first retrieval + vector footnotes + validator. Compare to the incumbent RAG or copilot on accuracy, citations, token cost, and latency. Route cheap models at the graph; reserve flagship models for the residual hard reasoning. This is where the FinOps case appears.

Days 66–90 · Constrain one action

Let the agent do one real thing the graph can authorize — draft, ticket, or recommend — with a human on the authority edge. Instrument usage → workflow change → operating change. Kill the slice if it cannot move a metric the CFO already believes. Only then clone the pattern to a second domain.

The modernization sequence, restated

If you are advising a program right now, the useful order is not “pick an agent platform.”

1. **Kill or cap unmanaged token firehoses**
   Routing, caching, spend caps, eval harnesses. This is FinOps, not MLOps theater.

2. **Name the context objects the business already uses**
   Before you buy another wrapper. If two departments disagree on the object, that disagreement is the work.

3. **Build a governed context layer**
   Graph + permissions + freshness that every copilot and agent shares, so “operating margin” means the same thing in finance, ops, and the agent.

4. **Only then attach agentic loops**
   And only to workflows where the graph can constrain action. Autonomy without an authority edge is a demo.

5. **Instrument the last mile**
   Usage → workflow change → operating change → P&L. Or you will be back in the trough with a nicer demo.

> **THE TEST THAT MATTERS**
>
> Can two agents, two copilots, and a human analyst give the same answer to the same as-of question, with the same citations and the same permission cut? If they cannot, you do not have enterprise context. You have a collection of prompts.

SOURCES & READING NOTES

Figures in this brief are directional

The statistics used here come from publicly reported analyst and research summaries current as of August 2026. They are included to show the shape of the market, not as a substitute for primary Gartner, MIT, or McKinsey documents behind a client paywall. Where a study has been criticized for methodology — notably MIT NANDA's narrow definition of success — the brief treats the number as directional and leans on the cross-source pattern.

- Gartner Hype Cycle for Agentic AI, 2026 — agentic AI at the Peak of Inflated Expectations; ~17% deployed, 60%+ intending within two years; governance, security, and cost profiles now appear across the cycle.
- Gartner Hype Cycle for Generative AI, 2026 — ≥50% of GenAI projects forecast to overrun budget from architecture and operating gaps; most custom-model efforts abandoned; no profile yet on the Plateau of Productivity.
- Gartner cancellation forecast — more than 40% of agentic AI projects cancelled by the end of 2027 on cost, unclear value, and weak risk controls; "agent washing" concentrated among undifferentiated vendors.
- MIT Project NANDA — widely cited finding that on the order of 95% of enterprise GenAI pilots produced no measurable P&L impact; use as a directional signal.
- McKinsey QuantumBlack on agentic economics — ~93% of organizations exceeding AI budgets; agentic tasks can consume ~1,000× the tokens of single-turn chat because of long-lived context.
- Enterprise retrieval and GraphRAG literature — permission and governance issues in the majority of AI programs; published GraphRAG and schema-guided retrieval results showing large token reductions and multi-hop accuracy gains versus naive RAG; ontology and entity-resolution effort remains the binding constraint.
- Cloud / infrastructure surveys, 2026 — growing evaluation of AI workload repatriation driven by cost unpredictability, sovereignty, and latency of agent loops.

The infographic page in this file is the 17 × 11 in. poster. Print it at 100% on tabloid / A3+ or keep it as a single-slide visual. The Letter pages are the working argument and the GraphRAG design that the poster cannot hold.