The Hallucination Tax.

A field guide to defensible enterprise AI

Contents

Executive summary

The industry is telling enterprises that hallucinations are mostly solved. The data agrees — for the workflows nobody runs in production.

Three findings shape this paper. First, the aggregate hallucination rate on isolated benchmarks is falling. That data is real, and it is the data being cited. Second, in the conditions enterprises actually deploy under — agentic workflows, reasoning over broad retrieval, high-stakes domain queries — hallucinations are not falling. They are rising sharply. Third, every hallucinated output is paid for at full token rates. The agentic shift has amplified token consumption by orders of magnitude, and the enterprise AI bill is rising even as unit prices fall.

Three implications follow. First, model selection is not the lever enterprises have been told it is. Better models are coming; they will not solve this. Second, retrieval and source architecture is the lever, and most enterprise AI stacks lack the architecture to operate it. Third, the metric CFOs need to ask for does not yet exist on most invoices: **Cost Per Defensible Output** — the fraction of AI spend that produced auditable, source-grounded answers.

> In a per-token economy, the hallucinated answer costs the same as the correct one. Often more, because hallucinated reasoning chains run longer before they land.

One recommendation. Enterprises that intend to scale AI past pilot purgatory in 2026 will adopt source-controlled architecture as a procurement criterion. They will require their AI stack to see the sources behind every model output, score the influence of each source, and refine which sources their models are allowed to reason from — in production, without retraining. We call this **sourced AI**. It is what SeekrFlow is built to deliver.

Numbers anchoring the paper

33%

OpenAI o3 hallucination rate on PersonQA, per OpenAI's own April 2025 system card. Predecessor o1: 16%.[1]

86%

GPT-5.5 hallucination rate on the independent AA-Omniscience benchmark when relying on its own weights, April 2026. OpenAI marketed a 60% reduction; the system card actually reports 3% at the response level.[2]

69–88%

Stanford RegLab's measured hallucination range on specific legal queries. 75%+ on questions about a court's core ruling.[3]

100×

Growth in enterprise AI token consumption over two years, while unit prices fell 10×. Jevons paradox in inference.[4]

5.5%

Share of enterprises seeing significant financial value from AI, per McKinsey's 2025 State of AI. The other 94.5% are in pilot purgatory.[5]

CHAPTER 01

The aggregate is real. Under load, the picture flips.

In May 2025, Anthropic's CEO Dario Amodei told reporters that today's frontier models "probably hallucinate less than humans."[6] In April 2026, OpenAI launched GPT-5.5 with launch-day coverage citing a 60% reduction in hallucinations. Vectara's HHEM leaderboard now shows top models hallucinating under 1% on grounded summarization tasks.[7] The headlines are clear. The hallucination problem is on its way out.

Pull the AI usage logs of any large enterprise running these models in production. The story doesn't match.

1.1 What the aggregate measures

The Vectara number is honest. The Hughes Hallucination Evaluation Model measures one specific task: given a document, summarize it. The model is handed grounding material and asked to stay close to it. In that condition, the best models are remarkably good.[7]

That is not what enterprises do.

Enterprise AI workflows are multi-step. They retrieve from broad, often unverified corpora. They run reasoning chains across multiple model calls. They invoke tools whose outputs become inputs to the next step. The grounding material is not handed to the model; the model goes looking for it, and what it returns is paid for at the same rate whether it was the right material or the wrong one.

Vectara is measuring a single condition that has been mostly solved.

The aggregate is real. The conditional is the opposite.

1.2 The OpenAI evidence

OpenAI shipped o3 and o4-mini in spring 2025. The system card disclosed something the industry has not absorbed: on PersonQA — OpenAI's internal hallucination benchmark — o3 hallucinated on **33%** of prompts. o4-mini hallucinated on **48%**. The predecessor o1 hallucinated on **16%**. o3-mini hallucinated on **14.8%**.[1]

> The newer reasoning models doubled
> and tripled the rate of their predecessors.

OpenAI's technical report said only that "more research is needed."

Twelve months later, the gap between marketing and documentation has widened. When OpenAI launched GPT-5.5 on April 23, 2026, the launch coverage converged on a "60% reduction in hallucinations." That number is not in OpenAI's system card. The deployment safety hub reports a 23% reduction at the claim level and a **3% reduction at the response level** — evaluated on conversations users had previously flagged as containing factual errors on prior models.[2]

Independent evaluation pushes the picture further. On the AA-Omniscience benchmark, GPT-5.5 hallucinates 86% of the time when it has to rely on its own weights — more than double Anthropic's Claude Opus 4.7.[2] The marketing keeps outrunning the data.

1.3 Why this is structural, not a tuning issue

An ICLR 2026 paper called "The Reasoning Trap" demonstrated that training models to reason harder amplifies hallucination instead of reducing it. The paper introduced the SimpleToolHalluBench benchmark and showed the relationship is causal: deeper reasoning trains models to chain claims, and longer claim chains contain more wrong claims AND more right ones.[8]

Chain-of-thought compounds confidence, not accuracy. A reasoning model produces more output per query, so it produces more correct content AND more fabrication. On benchmarks where the ratio is what matters, the curve bends up.

This is before agents.

> Stanford RegLab measured 69–88% hallucination on specific legal queries. On questions about a court's core ruling — the questions lawyers most need answered correctly — hallucination rates exceeded 75%.

1.4 Where it matters most

Stanford RegLab's legal hallucination study measured hallucination rates between 69% and 88% on specific legal queries. On questions about a court's core ruling — the questions lawyers most need answered correctly — hallucination rates exceeded 75%.[3]

The downstream effects are now visible in the docket. In February 2025, U.S. District Judge Kelly Rankin sanctioned three Morgan & Morgan attorneys $5,000 in *Wadsworth v. Walmart* for submitting a motion citing eight AI-generated fake cases. The citations came from Morgan & Morgan's in-house AI platform, not ChatGPT. Other sanction cases followed across 2025 and into 2026.[9] In Q1 2026, U.S. courts issued **$145,000 in sanctions** against attorneys who filed AI-generated false citations — the highest quarterly total in legal history.

Clinical case summaries hit **64.1%** hallucination without structured mitigation prompts in a 2025 MedRxiv study. Even with mitigation, the best-performing model (GPT-4o) hallucinated 23% of the time.[10] The downside in regulated medical decision support is asymmetric and the underlying rate is not improving at the speed of model releases.

In each of these domains, the same dynamic holds: aggregate benchmarks paint a picture that bears little resemblance to the workflows where decisions actually get made.

CHAPTER 02

The agentic amplifier

The shift toward agentic AI is the single biggest force expanding enterprise AI spend in 2026. It is also the single biggest force expanding the hallucination tax.

2.1 What agentic adds

McKinsey's 2025 State of AI reports that 23% of enterprises are scaling agentic AI in at least one business function, with another 39% experimenting.[5] The technology has moved from a research preview into mainstream enterprise planning in roughly 18 months.

An agentic workflow differs from a chat workflow in three ways that matter for both accuracy and cost:

- **Multi-step decomposition.** A single user request triggers multiple model calls. Each call's output becomes the next call's input.
- **Tool invocation.** The model calls external systems — search, retrieval, code execution, third-party APIs. Tool outputs feed back into the reasoning chain.
- **Context compounding.** The system maintains state across the chain. Each call carries forward the prompt, the prior outputs, and the accumulating context. Token consumption grows non-linearly with steps.

All three amplify hallucination. All three amplify cost.

2.2 The propagation research

Microsoft Research published VeriTrail in January 2026, a system for detecting hallucination and tracing provenance in multi-step AI workflows. The paper's central finding: hallucination detection at the workflow level requires per-step provenance because errors propagate through chains. An upstream hallucination at step three becomes downstream evidence at step seven, and by the final output the chain has produced confident-sounding nonsense from a single contaminated input.[11]

The AgentHallu benchmark formalized the dynamic. Planning hallucinations — wrong decisions made early in a workflow — propagate into downstream tool calls and final outputs. The compounding effect means agents fail more often than the sum of their per-step failure rates would suggest.[12]

> Hallucination detection at the workflow level requires per-step provenance because errors propagate through chains.

Put plainly: if step one is wrong by 10% and step two is wrong by 10%, step two's output is not 90% reliable. It is closer to 80%, because step two is reasoning over step one's output. A five-step chain with 90% per-step reliability produces a final output reliable at 59%, not 90%.

This is not a curve enterprises can outrun by buying a better step-one model. The architecture is the problem.

2.3 What this means in production

A recent example shows the dynamic at consulting-firm scale.

In October 2025, Deloitte Australia refunded part of a AU$440,000 contract to the Australian Department of Employment and Workplace Relations after a 237-page government report was found to contain fabricated academic citations, references to nonexistent papers, and an invented quote from a Federal Court judgment. The report had been produced with Azure OpenAI GPT-4o. The errors were caught on public release by a Sydney University law professor.[13]

The Deloitte case is instructive. The model was not the fundamental issue. The architecture was. The workflow had no provenance check between research output and final document — no system to verify that the citations the model produced corresponded to documents that actually existed.

> Deloitte's own subsequent CFO advisory on AI economics, published three months later, was widely cited and well-researched. The same firm, two adjacent dynamics: world-class analysis when the architecture supports it; a public credibility hit when it doesn't.

Most enterprises do not have the public scrutiny that catches a fabricated court citation. They have downstream business processes — customer communications, financial analysis, supply-chain decisions — where the same architectural problem produces the same kind of confident-wrong output, and nobody catches it until the cost is in the data.

CHAPTER 03

The unit economics of accuracy

The cost story is the part of this picture most enterprise leaders are now seeing firsthand. The combination of falling unit prices and rising aggregate consumption produces a billing dynamic that traditional software TCO models do not capture.

3.1 The Jevons paradox in AI

Inference costs per token have dropped roughly an order of magnitude over the past two years, driven by model efficiency improvements and competitive pressure among cloud providers. The naive expectation would be that total enterprise AI costs are falling.[4]

They are not. Total enterprise AI spend is rising. Nutanix's CIO Greg Sengupta and reporting in VentureBeat have called this **Jevons paradox in AI inference**: when a resource becomes cheaper to use, consumption tends to increase faster than the price drops. Cost per token has dropped about 10× in two years. Token consumption has risen more than 100× in the same window.

Deloitte's January 2026 analysis put the same dynamic in operating terms: "While the unit price of AI tokens is falling, overall enterprise spending on and scaling of AI systems is rising. The number of users, complexity of models, and intensity of workloads will likely drive greater token consumption and, consequently, higher costs."[14]

This is the structural reason enterprise AI bills do not match the falling-price headlines.

3.2 What agentic does to the bill

Microsoft Research published an analysis in 2026 with a direct quantitative comparison. Agentic coding tasks consume roughly **1,000 times more tokens** than a chat conversation completing the same task. Input tokens — not output — drive most of the cost. The same agentic task can vary by **30×** in token spend run-to-run.[15]

Two findings from that work matter for budgeting.

- **Higher spend does not produce higher accuracy.** Accuracy peaks at intermediate cost in most workflows. Adding tokens does not buy reliability past a certain point. Some configurations are spending heavily AND producing worse results than cheaper alternatives.
- **Same task, different bills.** The 30× intra-task variance is the line item finance teams cannot forecast. Token consumption is a function of the workflow design, not just the user's query. Workflows that scale cleanly cost less; workflows that don't cost much more.

In a per-token economy, the hallucinated answer costs the same as the correct one. Often more, because hallucinated reasoning chains run longer before they land. So every hallucination is double damage: the cost of the wrong answer downstream, plus the cost of the tokens that produced it.

3.3 The CFO is now in the room

A widely-reported case from Deloitte's CFO Advisory illustrates the visibility problem. A large healthcare enterprise saw monthly token consumption grow 8–10% month-over-month, reaching roughly a trillion tokens over six months. The total translated to more than **$6 million in unplanned annualized cost** — before finance had visibility into the driver.[14]

PYMNTS reported in May 2026 that CFOs are now pushing back on token volume as a billing unit. Gartner forecasts that agentic AI will reach 30% of enterprise application software revenue by 2035, up from 2% in 2025 — an increase that will run through the CFO's line items.[16]

> Every hallucination is double damage: the cost of the wrong answer downstream, plus the cost of the tokens that produced it.

The structural problem facing enterprise finance is that AI spend is showing up on the invoice without a measurable denominator. Total spend is visible. The fraction of that spend producing defensible output is not.

3.4 Introducing Cost Per Defensible Output (CPDO)

This paper introduces a new metric: **Cost Per Defensible Output.**

$$\text{CPDO} = \text{Total AI spend} \div \text{defensible outputs}$$

An output is defensible when three conditions are met. The contributing sources can be identified (traceability). Each contributing source can be inspected and ranked by influence (source attribution). The output carries a confidence value the operator can act on (confidence scoring).

CPDO is not a precision metric. It is a directional one. The purpose is to force AI spend into two columns instead of one — defensible spend, and everything else. The fraction that is not defensible is the hallucination tax.

Practical guidance for finance and technology leaders is the same: **ask your AI vendors what your CPDO is. If they cannot answer, they cannot give you the architecture to lower it.** That is a procurement criterion.

CHAPTER 04

The 5.5% are doing something different

The 5.5% are not running different models than the other 94.5%. They are running the same models inside a different architecture.

4.1 What the McKinsey data shows

McKinsey's 2025 State of AI surveyed 1,993 organizations across 105 countries. The headline numbers are now widely cited: **88% of enterprises use AI** in at least one business function. 72% use generative AI, up from 33% the prior year.[5] Bain pegs U.S. enterprise generative AI adoption at 95%.[17]

Inside those adoption numbers, McKinsey identified a small cohort — **5.5% of respondents** — that report significant value from AI (defined as more than 5% of EBIT attributable to AI). McKinsey calls these companies "AI high performers." They are not running different models than the other 94.5%. They are running the same models inside a different architecture.

Three differentiators stood out in the McKinsey data:

- **Workflow redesign.** AI high performers are **2.8×** more likely to have fundamentally redesigned workflows around AI capabilities (55% vs. 20%).
- **Agentic scale.** High performers are nearly **3×** more likely to have scaled agentic AI across the enterprise.
- **Human-in-the-loop discipline.** 65% of high performers have defined processes for when model outputs need human validation. Only 23% of others do.

McKinsey calls the gap between high performers and the rest "pilot purgatory" for the 94.5%. The technology has been adopted; production deployment is stuck.

4.2 The architectural read

The McKinsey behavior signals are consistent with a single underlying capability: high performers can see, score, and refine what their models are doing. That is what makes workflow redesign possible. That is what makes agentic scale survivable. That is what makes human-in-the-loop discipline operational instead of theoretical.

> The 5.5% are not running different models than the other 94.5%.
> They are running the same models inside a different architecture.

The high-performer cohort is small for a reason. Most enterprise AI stacks do not give operators the visibility required for this kind of discipline. They give operators outputs. They do not give operators source attribution at the granularity needed to triage problems in production.

This is the architectural gap the rest of this paper describes.

CHAPTER 05

The sourced-AI architecture

Most enterprise AI today is what we'll call **unsourced**. The operator can see the output. The operator cannot see — or score, or refine — what the model reasoned from to produce the output. The architectural defaults that ship with most off-the-shelf RAG plus frontier model stacks treat the model as a closed reasoning engine and the retrieval layer as an undifferentiated input pipe.

Sourced AI inverts this. It treats source visibility, source scoring, and source control as first-class capabilities — not features bolted on for governance review.

5.1 The three capabilities

See

For any output the model produces, the operator can identify the specific upstream content — Q&A pairs from fine-tuning data, retrieved document chunks, agent intermediate outputs — that influenced the response. Visibility is at the **token level**: which tokens in the output trace to which contributing sources, with confidence scores attached. Operators do not have to ask the model to justify itself. The system shows the influence directly.

Score

Every contributing source carries an influence score. Internal evaluator models rank sources as **High, Medium, Low, or Irrelevant** impact on the specific output. Operators can act on the scoring: drill into high-impact sources to verify, prune irrelevant ones, escalate low-confidence outputs for review.

Tune

The operator can refine what the model is allowed to reason from — narrowing retrieval scope, pruning low-value training data, reweighting source influence, re-fitting principle alignment — in production. **Without retraining the underlying model.** The control loop is closed at the architectural layer, not at the model layer.

> See. Score. Tune. The three together constitute sourced AI.

5.2 What this is not

Sourced AI is often confused with three adjacent capabilities. The distinctions matter.

- **RAG is not sourced AI.** RAG is a retrieval pattern. Most RAG implementations cannot tell the operator which specific retrieved chunks drove which parts of an output. They retrieve and they generate; what happened in between is opaque.
- **Observability is not sourced AI.** Observability tells you what happened. It records inputs, outputs, latency, error rates. It does not let you intervene at the source level to change what the model reasons from.
- **Eval is not sourced AI.** Eval tells you whether an output was correct on a test set. It does not give you the architectural control to fix the underlying reason an output was wrong in production.

All three are valuable. None of them closes the loop.

5.3 Architectural alternatives: ontology grounding

Source-level control is not the only architectural pattern available for reducing hallucination in production enterprise AI. A second pattern — most prominently advanced by Palantir under the term **Ontology-Augmented Generation**, or OAG — has gained traction in enterprise AI buying conversations and warrants direct examination both for its technical merits and for its limits.

What ontology grounding is

An ontology, in this context, is a software-modeled representation of an enterprise's real-world concepts: customers, distribution centers, supply routes, aircraft, contracts, patients. It captures objects and their properties — the "nouns" of the business — along with the relationships between objects, the actions that can be taken on them (the "verbs"), and the security rules that govern those actions. The result is what Palantir describes as a "digital twin" of the organization: a unified semantic and operational model that LLMs can query as structured tools rather than reason over as free text.

Under the ontology-grounding pattern, an LLM is given access to structured queries against the ontology. When a user asks which distribution centers can fulfill a particular order, the model does not retrieve free-text descriptions of distribution centers and stitch an answer together. It invokes a query against the ontology's Distribution Center objects and returns a deterministic result drawn from structured operational data. The same pattern extends to actions: once verbs are modeled, an agentic workflow can not only answer questions but execute against systems of record — trigger an approval, update inventory, write a transaction back to the source system.

Where it succeeds

For enterprise AI workflows whose answers map cleanly to structured operational data, ontology grounding reduces hallucination in a way that pure RAG architectures struggle to match. The model is not interpreting text and stitching a response; it is invoking deterministic queries against an authoritative graph of objects and relationships. Fleet operations, supply chain workflows, asset tracking, defense logistics, manufacturing operations — anywhere the question is "which of these specific things in our data matches this condition?" — are workflows where the ontology-grounding pattern has a real architectural advantage over text retrieval.

It also extends to action orchestration. Operational workflows that not only ask questions but execute decisions — rerouting shipments, reallocating inventory, triggering approvals — benefit from the same structural rigor. The ontology provides a controlled surface on which agentic systems can act with reduced risk of acting on a hallucinated input.

Where it has limited reach

Three properties of the ontology-grounding pattern bound its applicability.

First, it requires the ontology. Building one that accurately models an enterprise's data, processes, and decisions is a multi-quarter to multi-year engagement involving significant services investment. The model is only as grounded as the ontology is complete and accurate. Most enterprises pursuing AI ROI in 2026 cannot wait through a full ontology build before any AI value flows — and the McKinsey data on pilot purgatory cited earlier in this paper suggests they aren't waiting.

Second, it reaches only the structured slice of enterprise reasoning. Most enterprise AI workflows reason over unstructured content: contracts, clinical notes, regulatory filings, intelligence reports, court rulings, customer correspondence, technical documentation, research memos. This is where the most severe documented hallucination rates live — Stanford RegLab's 69–88% on specific legal queries, the 64.1% measured rate on clinical case summaries, the AI-fabricated court citations that produced $145,000 in Q1 2026 sanctions. Content that does not map cleanly to structured objects falls outside the ontology's architectural reach.

Third, ontology grounding does not in itself eliminate hallucination — it relocates it. Inside the ontology's scope, hallucination rates drop. Outside that scope, the model is still operating as a language model. And the ontology itself is fallible: stale mappings, schema errors, and data quality issues at the source feeds the ontology was built from all flow through into the outputs grounded against it. The marketing framing implies that the ontology is a trusted source. In production, the ontology is one source among many — and the operator's question becomes: how do I see what is actually influencing each output?

The distinction

Ontology grounding and source-level control answer different questions. The first asks: *how do we give the LLM the right structured knowledge to work with?* The second asks: *how do we know what the LLM actually used, and how do we control it?*

Under ontology grounding, the ontology is built once and grounded against. Operators trust the structured graph. Under source-level control, every source is visible, scored, and tunable in production — including, where present, any ontology objects retrieved during a given workflow.

> The choice is not ontology versus source control. The choice is whether the operator has visibility into what the model is reasoning from.

Coexistence

Most enterprise AI estates run both kinds of workflows. Operational queries against structured systems of record live alongside research over documents, regulatory analysis over filings, and agentic reasoning over mixed source types. The two architectural patterns are not mutually exclusive.

Ontology grounding can coexist with source-level control. In production estates that include an ontology — whether built on a vendor platform or maintained in-house — source-level control adds a complementary layer: visibility into how the ontology, alongside every other source the model is reasoning from, is shaping the outputs the enterprise is paying for. For estates that do not have an ontology and cannot afford the build cycle to construct one, source-level control is the lever that operates on the inputs the model is already using today.

> The choice is not ontology versus source control. The choice is whether the operator has visibility into what the model is reasoning from. That visibility is what makes any architecture — including an ontology-grounded one — defensible, explainable, and trusted at scale.

5.4 SeekrFlow

SeekrFlow implements the sourced-AI capability set as a control plane that sits on top of customer-chosen models.

It is model-agnostic and works with BYO fine-tuned models, open-source models, and frontier provider models alongside one another.[18]

- **See** — SeekrFlow traces every output back to the Q&A pairs from the fine-tuning dataset that most influenced it. Operators can click directly through to the original document chunk that generated each training pair, ensuring full provenance. Confidence is exposed at the token level with color-coded scoring.
- **Score** — Each influential Q&A pair is evaluated by an internal LLM and ranked High, Medium, Low, or Irrelevant impact. Irrelevant examples are filtered out automatically. Operators do not have to write the scoring logic.
- **Tune** — Operators prune low-value training data, refine retrieval scope, and reweight principle alignment without retraining the underlying model.
- **Agent observability** — Every agent run captured as a full trace. Every reasoning step, tool call, and output logged with input, output, and timing metadata. Zero configuration.

The result is a system where the operator gets a steering wheel where the rest of the market is selling self-driving cars with no view of the road.

5.5 The economic effect

When source control is real, two things happen to the enterprise AI bill at once.

First, hallucination rates drop on the workflows that matter most — the multi-step, retrieval-heavy, high-stakes workflows where the aggregate benchmarks were never going to help. Because operators can identify the noisy sources and prune them, the model is no longer reasoning over junk.

Second, token consumption drops on the same workflows. Pruning low-value sources doesn't just reduce hallucinations. It reduces the tokens spent generating outputs grounded in noise. Bloated retrieval inflated every prompt before; tighter retrieval costs less per call and compounds favorably across agentic chains.

Both the accuracy line and the cost line bend down together. **CPDO improves on both sides of the equation** — the numerator (spend) falls, the denominator (defensible outputs) rises.

Chapter 06

Case proof

6.1

Arcas — sovereign AI under the EU AI Act

Arcas, a European public-sector AI provider, selected Seekr in early 2026 as its explainable AI partner for sovereign AI deployments.[19] The motivating constraint: the EU AI Act's phased enforcement requires provenance and contestability for AI systems used in regulated sectors. Most frontier model providers cannot meet those requirements at the architectural level without significant scaffolding.

SeekrFlow deploys on customer infrastructure, surfaces source attribution at the Q&A pair level, and supports the data attribution and context attribution frameworks the EU AI Act requires. For Arcas, the partnership made sovereign deployment under tightening EU rules operationally feasible on a timeline that an in-house build could not match.

6.2

GDIT — federal agentic AI

Seekr and GDIT announced a collaboration to accelerate development of secure, trusted agentic AI solutions for the federal government. The federal mission profile makes provenance non-optional: OMB M-25-21 and adjacent directives require AI systems used in government decision support to expose contestability and traceability.

The partnership focuses on agentic course-of-action generation for defense and civilian-agency use cases, where the underlying workflow is multi-step, the stakes are operational, and the auditability requirements are absolute.

6.3

OneValley — commercial outcomes

OneValley, a Seekr customer serving a community of entrepreneurs, reports **60% post-interaction purchase confidence** among users who engaged with their SeekrFlow-powered experience. Users who interacted with a sourced AI system that could explain its recommendations were measurably more confident in the resulting decisions than users on unsourced systems. Trust is operational, not aspirational.

CHAPTER 07

Implementation path

Enterprises that intend to move from unsourced to sourced AI in 2026 will work through three phases. The approximate cadence is 30 / 60 / 90 days, though the specific tempo depends on workflow complexity and existing stack maturity.

DAYS 1–30

Diagnose.

- Audit current AI workflows. Identify the three to five highest-impact production workflows. Document whether operators can see, score, and refine sources today.
- Calculate baseline CPDO. Use total AI spend and a defensibility coefficient based on current source visibility. The number is directional. The point is to make the gap measurable.
- Identify the riskiest workflows. High-stakes domain queries, agentic workflows, and any workflow with regulatory or audit exposure. These are the workflows where the hallucination tax is highest and the architectural payback is fastest.

DAYS 31–60

Deploy.

- Pilot sourced architecture on one workflow. Start with a workflow where the operator can observe outputs and outcomes directly. SeekrFlow can deploy alongside existing model providers — there is no model swap required.
- Establish source-scoring discipline. Train operators to interpret influence scores and act on them. The capability is technical; the discipline is organizational.
- Measure CPDO weekly. Track defensible-spend fraction and hallucination-tax fraction as the pilot matures.

DAYS 61–90

Scale.

- Extend sourced architecture to the next two to three workflows. Prioritize by CPDO impact: workflows with the worst defensibility coefficient get prioritized.
- Build CPDO into AI governance reporting. Defensible spend, hallucination tax, and CPDO become standing metrics on the board AI dashboard.
- Update procurement criteria. Future AI vendor RFPs include the CPDO question. Vendors that cannot answer it get scored accordingly.

Most enterprises will not move all workflows to sourced AI in 90 days. The 90 days is enough to make CPDO measurable, demonstrate the architectural lever, and build the operating discipline. From there, sourced AI extends to the rest of the AI estate on a cadence driven by business priority, not technology constraint.

What comes next

Three things, in order

01

Calculate your CPDO.

Take three minutes with the Seekr CPDO calculator. Inputs are simple. The output gives you a defensible starting number to work from. Available at seekr.com/cpdo-calculator.

02

Audit one high-stakes workflow.

Identify the workflow with the highest combination of cost and risk in your current AI estate. Ask the team that runs it whether they can see, score, and refine the sources behind every output. If the answer is no, that workflow is the pilot.

03

Talk to Seekr.

Twenty minutes with a SeekrFlow engineer is enough to show you whether sourced architecture is the right next step. Request a walkthrough at seekr.com/request-a-demo.

Stop paying for confidently wrong answers.

Score your AI spend →

References

[1] OpenAI o3 and o4-mini System Card (April 2025). PersonQA: o3 33%, o4-mini 48%, o1 16%, o3-mini 14.8%. Coverage: TechCrunch, April 18, 2025. https://techcrunch.com/2025/04/18/openais-new-reasoning-ai-models-hallucinate-more/

[2] OpenAI GPT-5.5 Deployment Safety Hub (April 23, 2026). https://deploymentsafety.openai.com/gpt-5-5 . Analysis of marketing-vs-system-card gap and AA-Omniscience finding: Wire Blog, "GPT-5.5 didn't cut hallucinations 60%." https://usewire.io/blog/gpt-5-5-hallucination-drop-is-a-context-engineering-win/

[3] Stanford RegLab & Stanford HAI, "Hallucinating Law: Legal Mistakes with Large Language Models are Pervasive" (2024) and follow-up in Journal of Empirical Legal Studies (2025). https://hai.stanford.edu/news/hallucinating-law-legal-mistakes-large-language-models-are-pervasive

[4] VentureBeat / Nutanix, "Cheaper tokens, bigger bills: The new math of AI infrastructure" (April 2026). Cost per token –10x in two years; consumption +100x. Jevons paradox in inference. https://venturebeat.com/orchestration/cheaper-tokens-bigger-bills-the-new-math-of-ai-infrastructure

[5] McKinsey, "The state of AI in 2025: Agents, innovation, and transformation" (November 2025). 88% adoption; 72% generative AI; 23% scaling agents; 5.5% significant financial value. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

[6] Anthropic — Dario Amodei at Code with Claude (May 2025). "AI probably hallucinates less than humans." TechCrunch: https://techcrunch.com/2025/05/22/anthropic-ceo-claims-ai-models-hallucinate-less-than-humans/

[7] Vectara HHEM Leaderboard. https://github.com/vectara/hallucination-leaderboard

[8] "The Reasoning Trap," ICLR 2026. SimpleToolHalluBench benchmark. Causal demonstration that deeper reasoning amplifies hallucination.

[9] Wadsworth v. Walmart, U.S. District Court, D. Wyo. (February 24, 2025). Morgan & Morgan sanctioned $5,000 for AI-generated false citations. ABA Journal: https://www.abajournal.com/news/article/no-42-law-firm-by-headcount-could-face-sanctions-over-fake-case-citations-generated-by-chatgpt . Q1 2026 sanction total from public docket compilation.

[10] MedRxiv (2025): Clinical case summary hallucination rate 64.1% without mitigation, 43.1% with structured prompting. Best-performing model (GPT-4o): 23% with mitigation.

References continued

[11] Microsoft Research, “VeriTrail: Detecting hallucination and tracing provenance in multi-step AI workflows” (January 2026). https://www.microsoft.com/en-us/research/blog/veritrail-detecting-hallucination-and-tracing-provenance-in-multi-step-ai-workflows/

[12] AgentHallu: Benchmarking Automated Hallucination Attribution of LLM-based Agents (2026).

[13] Deloitte Australia AI report refund (October 2025). AU$440,000 contract with Australian Department of Employment and Workplace Relations. Fortune coverage: https://fortune.com/2025/10/07/deloitte-ai-australia-government-report-hallucinations-technology-290000-refund/

[14] Deloitte Insights, “AI tokens: How to navigate AI’s new spend dynamics” (January 2026). https://www.deloitte.com/us/en/insights/topics/emerging-technologies/ai-tokens-how-to-navigate-spend-dynamics.html

[15] Microsoft Research, “How Do AI Agents Spend Your Money? Analyzing and Predicting Token Consumption in Agentic Coding Tasks” (2026). 1,000x agentic-vs-chat token consumption; 30x intra-task variance.

[16] PYMNTS, “Enterprises Look Beyond Token Counts to Measure AI” (May 2026). Gartner forecast on agentic AI revenue share.

[17] Bain & Company: U.S. enterprise generative AI adoption at 95% (2025 Bain survey).

[18] Seekr SeekrFlow Explainability product documentation: https://www.seekr.com/seekrflow/explainability/

[19] Arcas selects Seekr as explainable AI partner for sovereign AI under EU AI Act (March 2026). https://www.prnewswire.com/news-releases/arcas-selects-seekr-as-explainable-ai-partner-to-deliver-sovereign-ai-as-eu-ai-regulations-tighten-302730466.html