Look Up Before You Make Up - ARAMAI

Abstract. Enterprise generative AI fails in production at a rate that has little to do with model capability. The dominant failure mode is structural : systems generate answers from statistical pattern instead of consulting the authoritative definitions an organization has already committed to. This is schema blindness, and by 2026 the diagnosis has moved from contrarian to consensus. This paper argues for an inversion of the default order of operations — a principle we call look up before you make up (LUBMU) . Before a generative system produces an answer, it should consult the authoritative structure that governs the question and let that structure constrain what it generates; where the ground runs out, it should refuse or enrich rather than improvise. We locate the principle in a generation–verification asymmetry , describe the three architectural commitments it requires — including structure-preserving retrieval, which we treat as a methodology of its own (Structured Context Retrieval, SCR) in a companion paper — enumerate the scenarios where LUBMU decides success from failure, marshal the 2026 evidence that it is the dominant variable in accuracy and hallucination, delineate where it is mandatory versus optional, and close on what it asks of the people who build.

1. The failure is structural, not the model

A widely cited 2025 study from the MIT NANDA initiative found that roughly 95% of enterprise generative-AI initiatives delivered no measurable return , and located the cause not in model quality but in a "learning gap" between generic tools and the specific organizations deploying them. The reflex response — a bigger, more fine-tuned model — misreads the problem. Frontier models are extraordinarily capable. What they lack, inside an enterprise, is access to the specific authoritative meaning of the organization deploying them.

That meaning is not absent. An enterprise has spent decades encoding it — in database schemas, data dictionaries, dimensional models, content models, and the canonical definitions teams fought over in meetings. The meaning is real, but it is trapped inside systems and unavailable to the model at the moment of generation. So the model does the only thing it can: it improvises a plausible answer from its training distribution. Plausible is not correct, and in regulated, financial, or clinical contexts the gap between them is precisely where the cost lives.

By 2026 this reading is mainstream. As the keynote of that year's Semantic Layer Summit put it, the reason most enterprise AI fails to deliver business impact "isn't the quality of the LLM — it's the absence of a governed, trusted business context behind

the answer." The model was never the bottleneck. Structure was.

2. The principle and the asymmetry beneath it

Look up before you make up. Before a system generates, it should consult the authoritative structure that governs the question, and let what it finds constrain the answer.

Stated plainly the principle sounds obvious; its force is in the inversion. The default architecture of retrieval-augmented generation adds context and hopes it helps — it rarely verifies that the added context is authoritative or that it actually reduced the system's uncertainty. LUBMU is a stronger commitment: the canonical reference is consulted as a precondition of answering, not an optional garnish.

The reason this works rests on an asymmetry the agentic era makes acute. Generation is cheap, fast, and effectively unconstrained — a model will produce a fluent answer to almost anything. Checking that answer against authoritative structure is comparatively cheap and far more reliable than producing it unaided. A system that consults authority before it speaks exploits that asymmetry deliberately: it spends the model's fluency on expressing a grounded answer rather than on inventing one. Authority is not a brake on capability. It is what converts capability into trust.

3. What looking up requires

LUBMU is not a prompt technique. It is an architectural commitment with three parts, and it sits beneath the governed-semantics and catalog layers most enterprises already recognize — closer to the foundation than to the application.

Consult authoritative structure, not prose. The thing a system looks up should be the organization's canonical definitions and relationships — its semantic layer — treated as the first source of truth. When a question maps to a defined metric or entity, the system returns the canonical value rather than reasoning its way to a plausible one. Trust is ranked: authoritative structure outranks lineage, which outranks loose prose context.

Preserve structure through retrieval. Much enterprise meaning is relational — it lives in how typed things connect, constrain, and exclude one another. Retrieval that collapses that structure into undifferentiated similarity commits a premature semantic compression: it discards the very relationships that distinguish a correct answer from a believable one. We treat this as a methodology in its own right — Structured Context Retrieval (SCR) — and develop it in a companion paper.

Check conformance to known shapes. When a system consumes data it did not produce, trust should rest on whether that data conforms to a declared, machine-checkable contract — a known shape — rather than on hope. The semantic-web community has expressed such contracts for a decade through SHACL and ShEx; the gap is not the standard but the discipline of treating conformance as a precondition. Known shapes turn trust from an assumption into a check.

None of this replaces the model. It surrounds the model with authority, so that fluency is spent on a grounded answer rather than a fabricated one.

5. The evidence

The discipline is no longer aspirational. Several independent lines of 2026 evidence converge: authoritative structure, consulted first, is the dominant variable in accuracy and hallucination.

The clearest case is Anthropic's own. In June 2026 Anthropic described running most of its internal analytics through Claude agents. The most trusted source the agents consult is a curated semantic layer of canonical definitions, and the team made consulting it the mandatory default: "before you guess an answer or write any SQL, check the semantic layer first." Without that layer the agents were right 21% of the time; with it, 95% , and in some domains near 99% . An attempt to auto-generate the definitions from raw data failed — it "reproduced the underlying ambiguity rather than resolving it"; letting the model write them by hand was "its single biggest source of wrong answers." Dimensional modeling, they reported, proved "just as important as it ever was." The bottleneck "wasn't access to prior work — it was structure."

Independent benchmarks converge on the same delta. A 2026 paired evaluation across frontier models found semantic-layer routing lifting accuracy from the mid-80s to 98–100% on enterprise suites. A typed knowledge-graph grounding study (AssetOpsBench, KDD 2026) raised an industrial-operations agent from 65% to 99% by using the model to read and route and the typed graph as the source of truth. A conversational-analytics review documented progression from 10–20% raw-schema accuracy to 90–99% as each authoritative layer was added. The delta is structural — and structure-preserving retrieval (SCR) is how it is captured rather than compressed away.

Hallucination falls in proportion. Deployments consulting governed structure first routinely report 80–90%+ reductions in error; governed-RAG studies show drops on the order of 87%. In our own life-sciences knowledge engagement, schema-grounded

retrieval reduced interpretation error by roughly 87%. Stanford HAI's 2026 AI Index documents that hallucination persists across frontier models despite scaling — model size alone does not close the gap. Structure does.

And the economics now favor it. Beyond accuracy, the semantic-layer-first pattern delivers large compute savings and auditability. Independent 2026 analysis attributes three-year ROI of 141–551% to governed semantic infrastructure against an average "silent hallucination" cost measured in the millions; a 2026 survey found 59% of organizations directing incremental budget to semantic layers as critical AI infrastructure.

6. Where it is mandatory, where it is optional

A foundational discipline earns trust by being honest about its scope. In regulated, multi-hop, or safety-critical work — financial controls, clinical decision support, supply-chain obligation, anything where a confident wrong answer is the failure — LUBMU should be mandatory and auditable , and where a question falls outside what the system can ground, it should refuse or enrich rather than generate. This is the emerging 2026 consensus: for such work, schema-grounded generation is no longer optional; it is infrastructure.

In creative or exploratory work — ideation, drafting, divergent generation — making things up is the feature, not the bug. There the discipline is weighted, not absolute . And it generalizes beyond text: as agents take in images and sensor streams, "authoritative structure" widens accordingly, but the order of operations does not change. Consult what is known; then generate.

7. What it asks of builders

Adopting LUBMU reframes roles rather than eliminating them. The unglamorous work of dimensional modeling and definition stewardship turns out to be more important in the agentic era, not less — it is the substrate the agents consult. The authority layer this implies is best understood not as a replacement for conceptual modeling but as its successor : it recovers the discipline that decayed when "schema-on-read" made structure feel optional, and gives it a new consumer — the agent. Domain experts are no longer asked to invent meaning from a blank page; they adjudicate which of an organization's competing definitions is canonical, and bless it.

In practice, that translates into a handful of concrete commitments:

• Make "check the canon first" the default path , not an opt-in — so grounding is what happens unless someone deliberately overrides it.

• Treat the semantic layer as a maintained product with an owner — not a one-time artifact and not a byproduct of a migration.

• Budget definition stewardship as ongoing work. Authority drifts the moment the business changes; currency is a running cost, not a project that closes.

• Instrument the lookup. Log when the system grounded against authority versus improvised, so "did it look up?" is observable rather than assumed.

• Make refusal and enrichment first-class responses. A system that can say "this is outside what I can ground" is more trustworthy than one that always answers.

• Adopt structure-preserving retrieval (SCR) wherever relationships carry the meaning — similarity-only retrieval discards exactly what those questions depend on.

• Version authoritative structure for non-human consumers , so agents adapt rather than break when a definition changes.

8. The discipline that makes it durable

Authority must be governed live. Anthropic reported watching its own agent accuracy drift from 95% to 65% in a single month as the curated layer fell behind the data. Consulting authoritative structure is only as good as the structure's currency; a lookup discipline without a governance discipline degrades silently.

Lookup is necessary but not sufficient. Consulting a current semantic layer answers "what does this mean," but not "where did this come from," "when was it true," or "how confident should the system be." The next primitives in this stack — provenance as first-class grounding, structure evolution, and measurable grounding — extend the discipline rather than replace it, alongside SCR. They are the subject of forthcoming work.

9. Conclusion

Enterprise AI underdelivers not because models cannot reason but because we let them generate before they consult. The remedy is an order of operations, applied as infrastructure rather than as a prompt: consult authoritative structure, preserve it through retrieval, check conformance, and only then generate — refusing or enriching where the ground runs out, and governing the authority so it stays true. The 2026 evidence is no longer suggestive; it is convergent. Look up before you make up. It is

the difference between a system that sounds right and one that is.

· References

• Challapally, A., et al. The GenAI Divide: State of AI in Business 2025. MIT NANDA Initiative, MIT Media Lab, July 2025.

• Anthropic. "How Anthropic enables self-service data analytics with Claude." claude.com, June 2026.

• dbt Labs. Semantic Layer vs. Text-to-SQL Benchmark. April 2026.

• Rumiantsau, M., et al. "Semantic Layers for Reliable LLM-Powered Data Analytics." arXiv:2604.25149, 2026.

• "AssetOpsBench: Typed Knowledge-Graph Grounding for Industrial-Asset LLM Agents." arXiv:2605.26874, KDD 2026.

• Stanford HAI. Artificial Intelligence Index Report 2026 , Responsible AI chapter.

• AtScale. Semantic Layer Summit 2026, opening keynote, May 2026.

• Futurum Group. Enterprise Data & Analytics Survey 2026.

• Colrows. "The ROI of a Company Brain." 2026.

• Seekr. "The Hallucination Tax." May 2026.

• Promethium. Conversational Analytics Architecture Review , 2026.

• Lakera. Structured/governed-RAG hallucination studies, 2024–2026.