# The Missing Layer

A market problem analysis · 5 November 2026

*Published by Entelligage. Entelligage is a company operating in the enterprise AI market; this paper analyses a market problem and does not describe any product or service.*

---

## The problem in brief

Enterprises are spending heavily on generative AI and, by the best public evidence available, most are not seeing measurable returns. The usual explanations point at the model: it is not capable enough, not safe enough, not cheap enough. This paper argues that the evidence points somewhere else. The difficulty sits between the organization's data and the model, in how information is prepared, labelled, kept current and selected at the moment a question is asked. That layer is rarely designed as a distinct piece of work, and nobody is usually assigned to own it.

Each statistic below is stated as its primary source states it, with the sample and its limits. Where a figure could not be checked against a primary source, it has been left out.

## 1. The implementation gap

In July 2025, MIT's Project NANDA published preliminary findings drawn from a review of more than 300 publicly disclosed AI initiatives, interviews with 52 organizations, and a survey of 153 senior leaders. Its headline: despite $30 to 40 billion in enterprise investment in generative AI, 95% of organizations are getting zero return. Just 5% of integrated AI pilots are extracting millions in value, while the rest stay stuck with no measurable P&L impact. [1]

The authors' reading is the one that matters here. The divide "does not seem to be driven by model quality or regulation, but seems to be determined by approach." [1] The report describes itself as preliminary findings, and its sample is limited relative to the headline. It should be read as a serious signal, not a settled measurement.

## 2. Retrieval won, and moved the hard problem

For knowledge that changes, most enterprises have settled on retrieval-augmented generation (RAG): the model is given relevant passages from the organization's own content at question time, instead of having that content built into its weights. In a 2024 survey of 600 US IT decision-makers at enterprises with 50 or more employees, Menlo Ventures reported that RAG "now dominates at 51% adoption, a dramatic rise from 31% last year," while only 9% of production models were fine-tuned. [2]

The consequence is straightforward. RAG decouples knowledge from the model, so the quality of an answer depends heavily on what is retrieved, and what is retrieved depends on how the underlying

content was prepared, structured and indexed. Choosing RAG did not remove the hard problem. It relocated it from the model to the data layer.

Agents sharpen this. An agent issues several retrievals and chains the results, so a weakness in the data layer can compound across steps instead of staying inside one answer.

## 3. Adoption is not the same as reliability

Retrieval is often assumed to solve hallucination. A preregistered Stanford study of three commercial legal research tools that use RAG found otherwise: the authors report that the tools "each hallucinate between 17% and 33% of the time," and that providers' claims of eliminating hallucination were overstated. They also found substantial differences between systems in responsiveness and accuracy. [3] Law is a useful test domain because errors can be checked, but the study covers only those tools and that task, and it does not by itself show what causes the errors.

What it does show is that putting a retrieval step in front of a capable model does not make the system reliable. How retrieval is built matters.

## 4. Four ways the data layer fails

This paper groups retrieval-layer problems into four categories. The grouping is the author's analytical framework, not a finding from a cited study.

1. **Naive chunking.** Content split at fixed lengths can separate an idea from the context that gives it meaning.
2. **Missing metadata.** Without reliable labels for owner, date, status or type, a system cannot scope a search or prefer current material over old.
3. **Stale or contradictory content.** Superseded policies and expired prices look as well formed as current ones. Similarity search has no inherent way to tell them apart.
4. **Context pollution.** Passing more retrieved material to the model is not automatically better. Irrelevant or conflicting passages can degrade the answer.

On the fourth, the authors of ChunkRAG report that evaluating and filtering retrieved passages at the chunk level reduced hallucination and improved factual accuracy compared with existing RAG methods. [5] It is a single preprint, but it supports the point that precision of context matters, not only quantity.

## 5. The missing layer is shared structure

A common response to missing metadata is to add tags. The difficulty is that tags produced ad hoc do not compose. One document is labelled "expired," another "outdated," another "superseded."

People and automated tools each produce their own label vocabulary, so a filter written against one will miss the others.

The structural alternative is a controlled vocabulary: an agreed set of entity types, attributes, relationships and categories, derived from the organization's own content, against which both documents and questions are classified. With one shared scheme, filters become reliable, a question can be mapped to the same categories the documents use, and rules in the scheme can flag retrieved material that violates them. This is a design argument, and it is the central claim of this paper.

There is research support, with limits. Sharma, Kumar and Li (arXiv, December 2024) describe OG-RAG, a method that grounds retrieval in a domain ontology. Compared with baseline methods across four language models, they report a 55% increase in recall of accurate facts, a 40% improvement in response correctness, 30% faster attribution of responses to their context, and a 27% boost in fact-based reasoning accuracy. [4] These are the authors' own results on their own benchmarks, from a preprint. They show that structure can help under test conditions, not that any given enterprise will see the same gains.

The vocabulary mismatch is also acknowledged in vendor documentation. One major cloud provider's retrieval documentation describes the case of a user asking about the "PTO policy for remote workers hired after 2023" when the documents say "time off," "telecommute," and "recent hires." [6] The question and the content are written in different vocabularies, and something has to bridge them.

## 6. Why the layer goes unowned

Three reasons stand out, and all three are the author's assessment rather than measured findings. First, the data layer falls between teams: data engineering owns storage, application teams own the interface, and neither owns the meaning of the content. Second, the layer is invisible in demonstrations, which usually run on a small, clean set of documents. Third, it is recurring work. Vocabularies drift as the business changes, so the effort is never finished, which is hard to budget as a project.

## 7. Limits of this analysis

The sources above are a preliminary report, an industry survey, one academic evaluation in a single domain (a preprint) and two further preprints. None establishes by itself that missing structure causes the stalls described in section 1. The claim of this paper is narrower: the available evidence is consistent with the problem lying upstream of the model, the mechanisms are plausible and documented, and the layer deserves explicit design and ownership. Better models may reduce some of these failures, and the question of how much remains open.

## Conclusion

The visible failure is that AI projects stall. The evidence suggests that the model is not the only place to look, and perhaps not the first. Between the organization's data and the model there is a layer of labelling, vocabulary, currency and scoping that determines what the model is allowed to see. Most organizations have not named that layer, designed it, or assigned anyone to it. That is the missing layer.

## Sources

All sources were opened and checked on 5 November 2026. Quoted wording is exact.


- [1] Challapally, Pease, Raskar and Chari, MIT NANDA. "The GenAI Divide: State of AI in Business 2025." July 2025 (preliminary findings). MIT's NANDA group does not host the report on its public pages (access is by request); the copy read for this paper is a third-party mirror: https://mlq.ai/media/quarterly_decks/v0.1_State_of_AI_in_Business_2025_Report.pdf
- [2] Menlo Ventures. "2024: The State of Generative AI in the Enterprise." 20 November 2024. https://menlovc.com/2024-the-state-of-generative-ai-in-the-enterprise/
- [3] Magesh, Surani, Dahl, Suzgun, Manning and Ho. "Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools." arXiv:2405.20362, 30 May 2024 (preprint). https://arxiv.org/abs/2405.20362
- [4] Sharma, Kumar and Li. "OG-RAG: Ontology-Grounded Retrieval-Augmented Generation For Large Language Models." arXiv:2412.15235, 12 December 2024 (preprint). https://arxiv.org/abs/2412.15235
- [5] Singh et al. "ChunkRAG: Novel LLM-Chunk Filtering Method for RAG Systems." arXiv:2410.19572, submitted 25 October 2024 (preprint). https://arxiv.org/abs/2410.19572
- [6] Microsoft Learn. "RAG and Generative AI - Azure AI Search." https://learn.microsoft.com/en-us/azure/search/retrieval-augmented-generation-overview