

RAG Is Not Enough: Retrieval Needs Evals, Routing, and Observability

RAG | Mar 07, 2026

Reliable retrieval systems need measurable answer quality, retrieval quality, traces, reranking, and failure review.

Article

Retrieval-augmented generation is still one of the most useful patterns in AI products, but basic RAG is not enough. Uploading documents, chunking text, and asking a model to answer from retrieved snippets is only the starting point.

The real engineering work is deciding what happens when retrieval fails.

Strong RAG systems include:

1. Clear source boundaries: which documents are allowed for which users.
2. Good ingestion: cleaning, metadata, chunk strategy, and document lifecycle.
3. Hybrid retrieval: dense search, keyword search, and reranking where needed.
4. Answer rules: cite sources, admit uncertainty, and avoid inventing facts.
5. Evals: test questions, expected evidence, and scoring for retrieval and answer quality.
6. Observability: traces that show what was retrieved, what was ignored, and why the answer was produced.

The most common mistake is judging RAG by a few impressive demos. A production RAG system needs a repeatable evaluation set that includes edge cases, outdated documents, conflicting documents, and questions that should not be answered.

For internal company workstations, RAG should be treated as a knowledge access layer. It should help an AI system know what it can rely on, not just provide more text to the model.

Web link

<https://www.dropticks.com/articles/rag-is-not-enough-retrieval-needs-evals-routing-and-observability>

Related links

<https://developers.llamaindex.ai/python/framework/>

<https://docs.ragas.io/en/stable/>

<https://arize.com/docs/phoenix>