

✦ REPORT · MARKET INTELLIGENCE

RAG
Trends Report

How retrieval-augmented generation is evolving — from naive vector search to hybrid, agentic, multimodal and permission-aware retrieval — and what it means for teams building knowledge systems.

FORMAT	LEVEL	READ TIME	UPDATED
Report	Intermediate	12 min	Oct 2026

OVERVIEW

What this report covers

Retrieval-augmented generation has gone from a clever pattern to the default way enterprises put their own knowledge in front of language models. Larger context windows did not kill it; they changed it. This report is SplendensLabs' qualitative read on where RAG practice is heading: what has become baseline, what is emerging, where the hype outruns the results, and what we recommend teams do next. It is based on our delivery work and contains no benchmark or market figures.

KEY TAKEAWAYS

- Hybrid search plus reranking is now the baseline, not an advanced technique.
- Long context and RAG are complementary: retrieval decides what goes into the window.
- Agentic retrieval — models planning and iterating their own searches — is moving into production for complex questions.
- Document understanding (layout, tables, images) is where many systems now win or lose.
- Evaluation and permission-aware retrieval have become requirements for enterprise buyers.

WHO IT'S FOR

- Heads of engineering and AI
- AI / ML engineers running RAG systems
- Product owners of knowledge assistants
- Architects planning enterprise search

HOW TO USE IT

Read 01 to recalibrate what 'good' looks like, use 02–03 to spot trends relevant to your system, and take the recommended moves in 06 into your roadmap.

CONTENTS

01 What has become standard
02 Trends in retrieval
03 Long context, documents and multimodality
04 Trust, security and operations
05 Hype vs reality, and what to watch
06 Recommended moves

01 / BASELINE

What has become standard

Practices that were advanced a short while ago are now what buyers and reviewers expect from any serious RAG system.

FIRST-GENERATION RAG	TODAY'S BASELINE
Fixed-size chunks, vector search only	Structure-aware chunks with context headers
Top-k straight into the prompt	Hybrid keyword + vector search and reranking
No citations or refusal behavior	Cited answers and honest 'I don't know'
Quality judged by trying a few questions	Golden sets and automated evals per release

METHOD NOTE

Signals from the field

These are qualitative observations from building and operating RAG systems for clients, not survey or benchmark results. Products in this space change quickly; verify specific features and pricing at the time of reading.

02 / RETRIEVAL

Trends in retrieval

The center of gravity is moving from a single search call to retrieval as a reasoning process.

PRODUCTION-READY

Agentic retrieval

The model decomposes a question, runs several searches, reads results and searches again before answering.

PRODUCTION-READY

Contextual chunk enrichment

Adding document-level context or summaries to each chunk before embedding, so isolated passages retrieve correctly.

SELECTIVE

Graph-based RAG

Entity and relationship graphs help with ‘connect the dots’ questions, at the cost of extraction effort and upkeep.

SELECTIVE

Structured + unstructured

Routing questions to text-to-SQL or APIs for numbers, and to document search for explanations.

Agentic retrieval loop

01 Understand the question → 02 Plan sub-queries → 03 Hybrid search + rerank → 04 Check evidence is sufficient → 05 Search again or answer → 06 Cite sources

Agentic retrieval improves hard, multi-part questions but adds latency and cost. We route simple questions to a single retrieval pass and reserve the loop for questions that need it.

03 / CONTEXT

Long context, documents and multimodality

Bigger windows and better document understanding are reshaping what goes into retrieval and what goes into the prompt.

TREND	WHAT IS CHANGING	IMPLICATION
Long context windows	Whole documents fit in a prompt	Retrieve documents or sections, not just snippets, where it helps
Prompt caching	Repeated context is cheaper to resend	Stable reference material can live in a cached prefix
Vision-capable models	Pages can be read as images	Tables, charts and scans become retrievable
Layout-aware parsing	Better extraction of structure from PDFs	Ingestion quality is now a key differentiator
Multi-vector retrieval	Richer per-token or per-region representations	Better matching on complex pages, with more storage

OUR VIEW

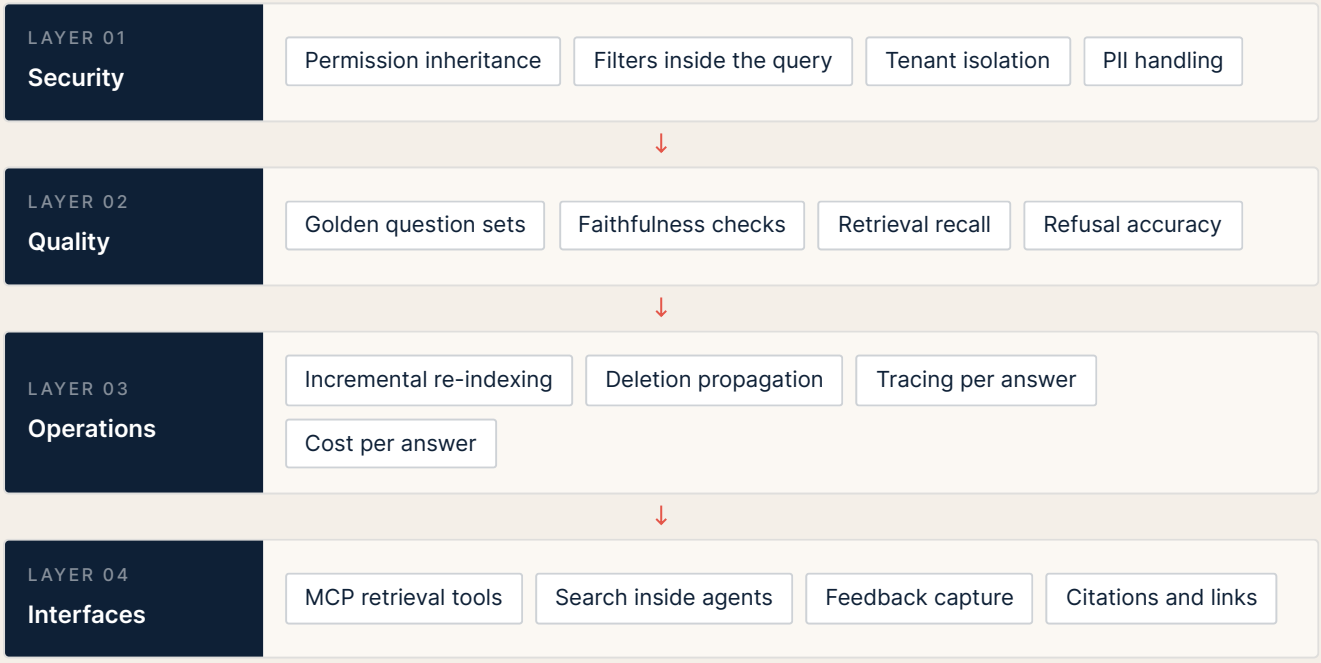
Long context does not replace retrieval

Even with very large windows, sending everything is slower, costlier, and harder to permission and cite. Retrieval remains the step that decides what the model should see and what the user is allowed to see.

04 / ENTERPRISE

Trust, security and operations

As RAG systems answer customers and staff in regulated settings, the non-functional requirements are catching up with the functional ones.



Signals we are seeing

- Security reviews now ask how permissions are inherited from source systems, not just whether the app has a login.
- Vector search is converging into general databases and search engines, reducing the need for a separate store.
- Retrieval is increasingly exposed as a tool that many agents call, rather than baked into one chat app.
- Teams expect to re-embed and re-index as models improve, so pipelines are being built for repeatability.

✦ 05 / OUTLOOK

Hype vs reality, and what to watch

Not every new technique earns its complexity. Here is how we rate the current conversation.

IDEA	OUR READ	WHEN IT IS WORTH IT
'RAG is dead'	Overstated	Never as a blanket claim; retrieval still selects and permissions context
GraphRAG everywhere	Useful, narrow	Entity-rich corpora with cross-document questions
Agentic retrieval	Real, with costs	Complex questions where single-pass recall falls short
Fine-tuning instead of RAG	Different job	Style and format; not for facts that change
Bigger embedding models	Incremental	After chunking, hybrid and reranking are fixed

What to watch

01

Retrieval as a service

Platforms bundling parsing, indexing, reranking and permissions as one managed layer.

02

Memory for agents

Retrieval over an agent's own history and user preferences, with privacy controls.

03

Standard eval practice

Shared methods for judging faithfulness and retrieval quality across vendors.

✦ 06 / RECOMMENDATIONS

Recommended moves

What we would prioritize for an existing or planned RAG system in the next two quarters.

FIX THE BASELINE

- [ ] Hybrid search with a reranker on every query path
- [ ] Structure-aware chunking with context headers
- [ ] Citations on every answer and a refusal threshold

MEASURE

- [ ] Golden question set built from real user questions
- [ ] Retrieval and answer metrics tracked per release
- [ ] Per-answer tracing and user feedback captured

SECURE

- [ ] Permissions inherited from sources and filtered inside the query
- [ ] Deleted documents removed from the index promptly
- [ ] Data residency confirmed for every component

EVOLVE

- [ ] Pilot agentic retrieval on the hardest question category
- [ ] Improve parsing for tables, scans and images
- [ ] Expose retrieval as a reusable tool for agents

> **The model gets the headlines; retrieval decides whether the answer is right.**
>
> SPLENDENSLABS PERSPECTIVE

✦ GET IN TOUCH

Bring your RAG up to date

We build and modernize production RAG systems — hybrid and agentic retrieval, document understanding, permissions and evals. Bring us your corpus and your hardest questions.

- RAG health check against today's baseline
- Retrieval and evaluation upgrade plan
- Agentic and multimodal retrieval pilots

BOOK AN AI STRATEGY SESSION →

EMAIL
hello@splendenslabs.com

WEB
splendenslabs.com

OFFICE
Sona Incubation Foundation, Salem