LIBRA A.I. Technologies

Yannis Kopsinis, PhD
CEO and Co-Founder

Agentic RAG in production

What actually works, what it costs, and how you'll know



1

What is RAG?

and why you need it

What RAG is: an open-book exam



Closed book

The model answers from what it learned in training.

Open book: RAG

It looks things up first (your documents, web sources, etc.) and answers from what it found.

R **Retrieval**
search the documents

A **Augmented**
add the best passages to the question

G **Generation**
write the answer from them, with sources

RAG retrieves external information and uses it to support a generated answer.

Three premature obituaries

“The Death of RAG”

D. Dahl · DEV Community · Feb 2024

Proposed

Put everything in a long context window

Test it when the authorised documents fit

“RAG is dead, long live agentic retrieval”

LlamaIndex · May 2025

Proposed

Let an agent choose indexes, and chunks or whole files

Still retrieval, under agent control

“Is Grep All You Need?”

Sen et al. (PwC) · arXiv · May 2026

Proposed

Let the agent grep the raw text; skip the vector index

Grep is retrieval too, and the harness mattered as much

Retrieval remains useful when the relevant evidence must be selected from a large, changing, or access-controlled collection.

Why Retrieval Evolves

Plain Questions

one focused retrieval may be enough

Can a consumer repay a mortgage early?

Research

the next step depends on what you find

"A customer took a mortgage in March 2019, paid it off early, and was charged 1% fee. Was this allowed by law?

Question complexity

Look-up

Find a specific piece of information.

Multiple sources

Combine information from different documents.

Versions

Documents may change over time

Cross-references

Follow links to related provisions or materials.

Evidence-driven next step

Decide what to investigate next based on what you find.

More focused → More investigative

The target: an acceptable answer, within budget

Evidence support

Material claims link to supporting passages

Applicable sources

Right version, Right jurisdiction and authority; precedence explained, conflicts flagged

Uncertainty

Abstains when evidence is missing; asks when ambiguity would change the answer

Access control

Only evidence this user may see, always

Within budget

The required quality at an acceptable total cost and latency.

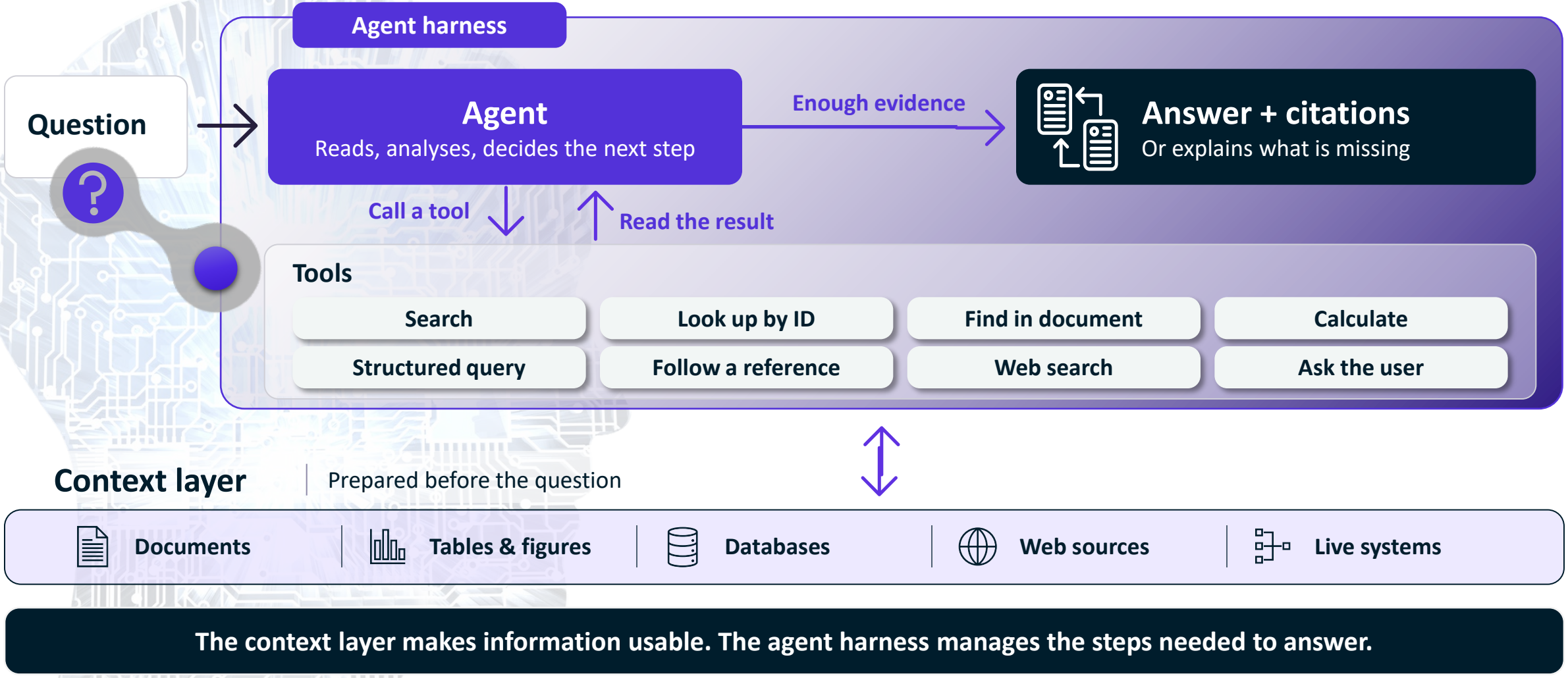
Main lever: selective context, enough evidence, not everything.



For our question:

“A customer took a mortgage in March 2019, paid it off early, and was charged 1% fee. Was this allowed by law?

The system: Context layer and Agent harness



The context layer makes information usable. The agent harness manages the steps needed to answer.

2

Preparation & Retrieval



Quiet failures

01 Wrong table parsing

The numbers are correct, but they're assigned to the wrong year.

Original (document)

Loan type	Maximum fee		
	2018	2019	2020
Consumer	1.0%	1.5%	2.0%

After extraction (to the AI)

2018	2019	2020	2021
Consumer	1.0%	1.5%	2.0%

2019 fee:
1.5% → 1.0%

✗

The year labels **shifted** during extraction.

02 Naive document chunking

An important exception can disappear when the text is split.

Original (document)

"The fee is waived, **except for loans under €5,000**."

After splitting (to the AI)

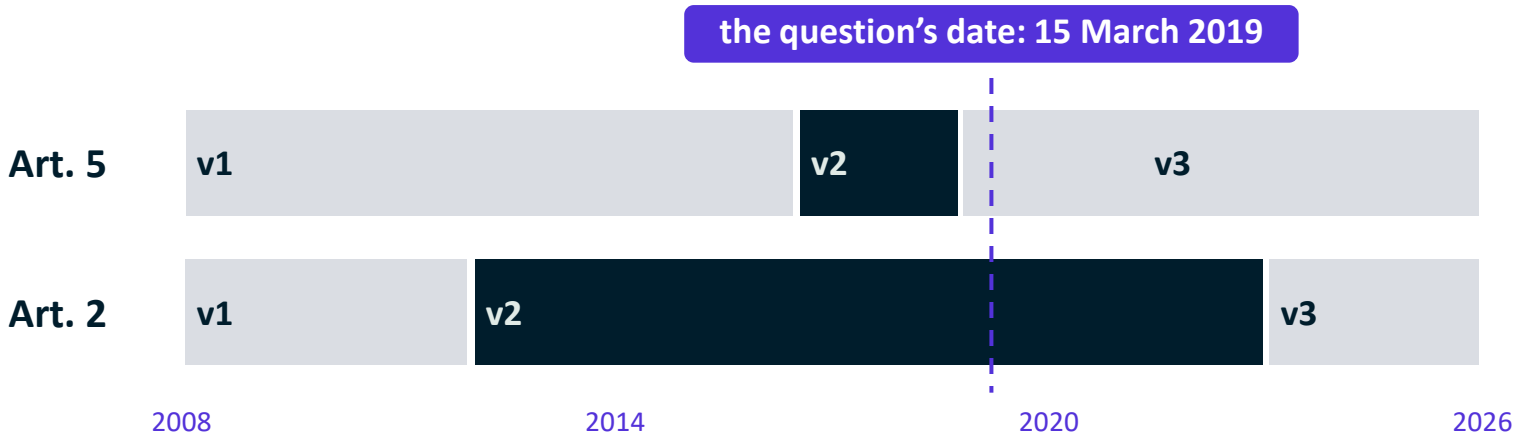
"The fee is waived."

✗

An important exception is lost.

Errors in preparation don't look like errors. They look like answers.

The latest version isn’t always the one you need



Applies: the version in force on the question’s date, not the newest.

Three dates to keep apart

- Published
- In force (effective)
- The date the question is about

2.7% → 98.3%

Cymbler, Guez & Fabre, “**Temporal Misgrounding in Legal RAG**: A Versioned-Corpus Benchmark for **French Tax Law**”, arXiv:2608.09393, August 2026

Dates select the version; cross-references and case law still need reading. Your domain determines the version, authority and reference rules.

Retrieval is three decisions

Where to look 01

Document vector index
Document raw content
Live systems & APIs
Web sources

How to find it 02

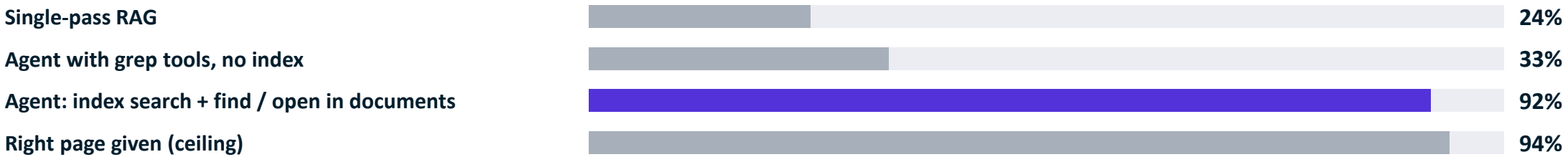
Filtering
Keyword search
semantic search
Hybrid · combine both
Agentic Search

What to keep 03

Remove duplicates
Rerank results
include context
Query-focused summarization
Extractive compression

An agent that searches the index and reads inside documents

Correct answers · 150 questions over long financial filings (FinanceBench)



AgenticRAG (Microsoft preprint), Table 3 · arXiv:2605.05538 · rows from different setups, baselines as reported in the paper · LLM judge + manual review

3

Optimize performance & costs





What drives agentic RAG cost - and how we control it

Three production decisions, not just a token bill

01

Model calls & model size

Use a small model for evidence checks. Reserve larger models for harder reasoning.

02

Repeated context & tokens

Cache shared context. Select and compress evidence before the final, expensive model.

03

Searches & tool rounds

Batch independent searches. Rerank results and stop when evidence is sufficient.

Published token-usage benchmarks

2.6× BRIGHT

7.8× FinanceBench

Agentic vs single-shot · Same model

Cost optimisation is an architectural decision, not just a model-selection.

“A cheaper wrong answer is not a saving.”

Source: AgenticRAG, arXiv:2605.05538 · Benchmark: token usage, not monetary cost · No production savings claim

The golden dataset is an asset

Knowledge experts

write question–answer pairs from real practice, with the supporting passages

Generated, then approved

drafted from the documents; experts approve, correct or reject every item

Cultivated from use

real questions and failures, reviewed and added

Golden dataset

Each item holds

- the question
- the expected answer
- supporting passages, with their version
- the question type
- expected behaviour: answer, abstain or ask

Why it is an asset

- Expert judgement no public benchmark has
- Reused for every change: model, prompt, parser, architecture
- Versioned with the documents; a held-out part never tuned on
- It becomes the specification

Machine drafts, experts decide: in IBM's MTRAG benchmark, annotators repaired the model-drafted response on **92%** of turns.

Evaluation framework is a must

Architecture

fixed or agentic · when to escalate · steps and searches · which checks

Context layer

parsing · splitting · labels and context · keyword, vector or hybrid · reranker

Prompts at every stage

rewriting · sufficiency check · agent and tool instructions · generation · verification

Model mixture

the right model per position: check, extract, rank, reason, judge

Every choice scored on

- quality: accepted answers, coverage
- cost per request
- latency: the slowest 5%

Keep what nothing beats on all three, per question type.

And the future: new models and architectures, tested on the same dataset, adopted or withdrawn fast.

The closing thought



Your standards for trustworthy answers should not.

Own the context and the design.

Own

Context Layer · Evaluation Methods and tools · tailor-made benchmark

LIBRA A.I. Technologies

Thank you!

Any Questions?

Yannis Kopsinis · CEO and Co-Founder · Libra AI Technologies

Yannis.kopsinis@libramli.ai