INTRODUCTION

Agentic RAG

When Retrieval Meets Reasoning

Daniel Schroter Thüm

Senior Consultant / AI Engineer

daniel.schroter.eu

March 2026

OVERVIEW

Agenda

01 RAG Recap & Limitations

02 What is Agentic RAG?

03 Architecture & Design Patterns

04 Research Landscape

05 Frameworks & Tools

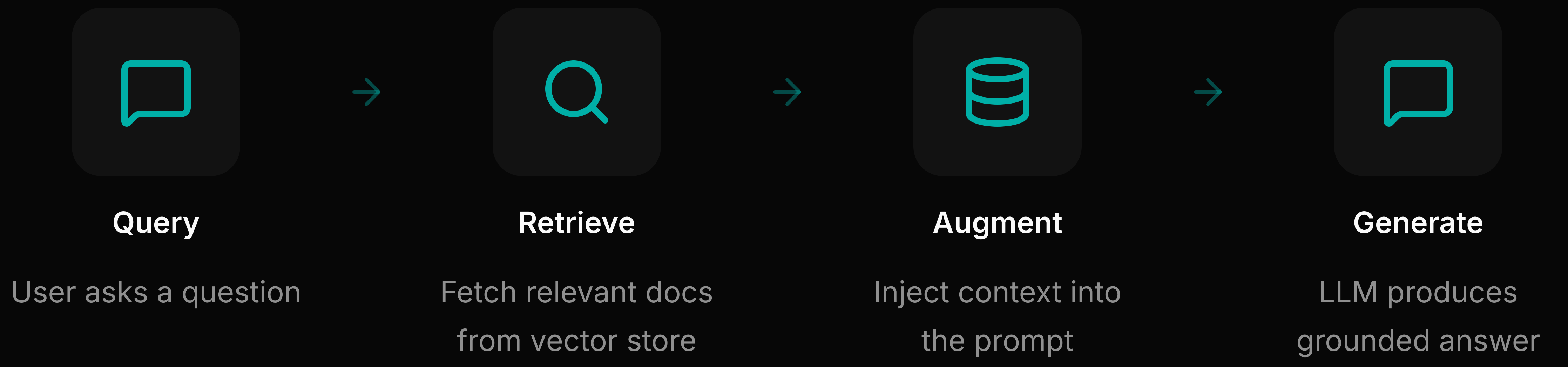
06 Advantages, Challenges & Use Cases

07 Getting Started

FOUNDATIONS

Retrieval-Augmented Generation

The standard pattern: retrieve external knowledge to ground LLM responses



Query

User asks a question

Retrieve

Fetch relevant docs from vector store

Augment

Inject context into the prompt

Generate

LLM produces grounded answer

Key insight: RAG bridges the gap between parametric knowledge (training data) and non-parametric knowledge (external documents), reducing hallucination.

THE PROBLEM

Limitations of Traditional RAG

One-Shot Retrieval

Single query, single retrieval pass. No ability to evaluate or re-retrieve if results are poor.

Single Data Source

Typically limited to one vector store. Cannot combine structured data, APIs, and documents.

No Self-Correction

Cannot assess retrieval quality. Garbage in, garbage out -- no validation loop.

Static Strategy

Same retrieval approach for every query. Cannot adapt to query complexity or type.

DEFINITION

Agentic RAG makes reasoning an integral part of retrieval.

Instead of blindly retrieving and generating, the system **reasons about what to retrieve**, **evaluates what it found**, and **decides what to do next**.

Reflection

Planning

Tool Use

Multi-Agent

HEAD TO HEAD

Traditional RAG vs Agentic RAG

TRADITIONAL RAG	AGENTIC RAG
WORKFLOW Fixed linear pipeline	WORKFLOW Dynamic, iterative reasoning loop
DECISION-MAKING Static rules	DECISION-MAKING Agent decides what/where/how
DATA SOURCES Single vector store	DATA SOURCES Multiple KBs, APIs, tools
COMPLEX QUERIES Struggles with multi-hop	COMPLEX QUERIES Decomposes into sub-tasks
SELF-VALIDATION None	SELF-VALIDATION Scores & re-retrieves
ADAPTABILITY Same strategy always	ADAPTABILITY Adapts in real-time

ARCHITECTURE

Key Architectural Patterns

Taxonomy from Singh et al. (2025)

Single-Agent (Router) `Simple`

One agent manages routing, retrieval, and integration. Centralizes decision-making across multiple knowledge sources.

Multi-Agent `Scalable`

Specialized agents work in parallel: orchestrator coordinates, router directs queries, researchers retrieve and analyze.

Hierarchical `Layered`

Multi-tiered agent structure. Top-tier handles strategic decisions, lower-tier agents execute retrieval tasks.

Corrective & Adaptive `Self-correcting`

Evaluates relevance of retrieved docs, refines queries iteratively. Adaptive variant routes by query complexity.

Source: Singh et al., "Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG" (2025) · arXiv:2501.09136

DESIGN PATTERNS

The Four Pillars of Agentic RAG

Reflection

Agent evaluates its own outputs and retrieval quality, triggering re-retrieval when results are insufficient.

`Self-RAG, CRAG`

Planning

Creates a step-by-step plan before executing retrieval. Decomposes complex queries into sub-tasks.

`Query decomposition`

Tool Use

Calls external tools -- search engines, calculators, APIs, code interpreters -- beyond simple vector retrieval.

`Function calling`

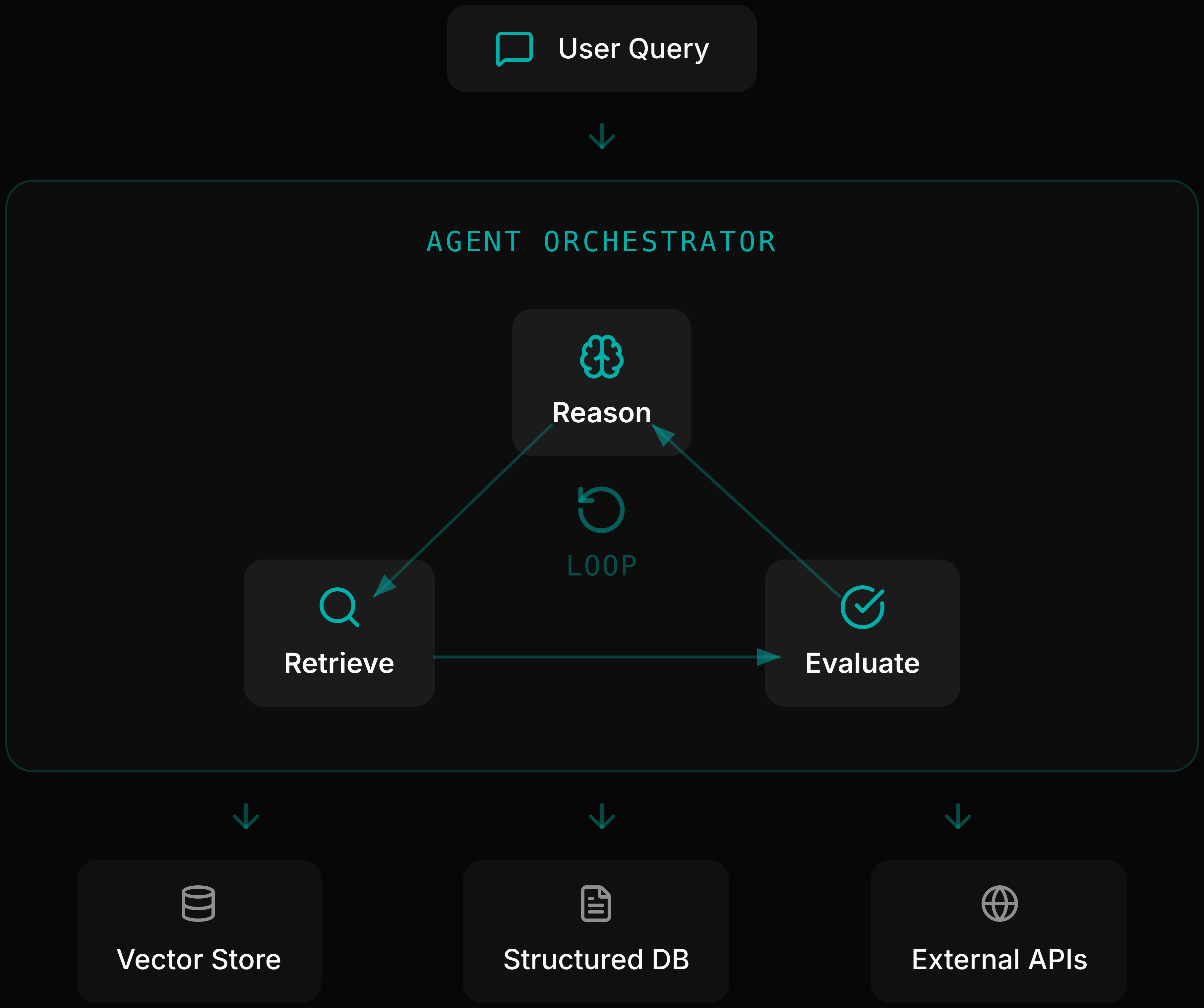
Multi-Agent Collaboration

Multiple specialized agents with different roles work together on complex information needs.

`Manager/worker`

DEEP DIVE

Agentic RAG Architecture



RESEARCH

Key Papers

Jan 2025	Agentic RAG: A Survey	Singh et al.	Foundational survey	arXiv:2501.09136
Oct 2023	Self-RAG: Self-Reflection for RAG	Asai et al.	Key precursor	arXiv:2310.11511
Feb 2026	A-RAG: Hierarchical Retrieval Interfaces	Du et al.	Scaling patterns	arXiv:2602.03442
Jul 2025	RAG-Reasoning with Deep Reasoning	Li et al.	Reasoning integration	arXiv:2507.09477
Jun 2025	Reasoning RAG: System 1 vs System 2	Liang et al.	Industry focus	arXiv:2506.10408
Aug 2025	Agentic Hybrid RAG for Science	Nagori et al.	GraphRAG + VectorRAG	arXiv:2508.05660

Also notable: Corrective RAG (CRAG), Adaptive-RAG, Tiny-Critic RAG (2603.00846), and RAG for Fintech (2510.25518).

EVALUATION

Trade-offs

✓ What You Gain

Higher Accuracy Through Iteration
Self-RAG: 55.8% on PopQA vs 14.7% base model.

Multi-Hop Reasoning
Refines retrieval based on intermediate insights.

Heterogeneous Sources
Vector stores, SQL, APIs, and knowledge graphs in one query.

Adaptive Routing
Routes by complexity: direct generation, single-step, or multi-step.

⚠ What It Costs

Latency
Each iteration adds another LLM call.

Cost
3-10x increase vs traditional RAG.

Reliability
Agent loops can fail without proper guardrails.

Observability
Multi-step pipelines are harder to debug.

Sources: Asai et al. (2023) · Singh et al. (2025)

IN PRACTICE

Real-World Use Cases

Customer Support `66% of queries handled`

Salesforce Agentforce at Fisher & Paykel. The survey also cites Twitch's ad sales system on Amazon Bedrock for campaign and audience retrieval.

Healthcare `68% → 73% accuracy`

Clinical decision support integrating health records with medical literature. Radiology QA improved across 24 LLMs with agentic retrieval.

Scientific Research `Hybrid RAG selection`

Dynamic selection between GraphRAG and VectorRAG per query. Also: research paper synthesis with enriched citations across domains.

Personal AI Assistants `Most widespread today`

AI coding tools (Cursor, Claude Code) and workspace assistants (Copilot, Notion AI) use agentic retrieval across local files, codebases, and web.

Sources: Singh et al. (2025) · Salesforce · arXiv:2508.00743 · Nagori et al. (2025)

KEY TAKEAWAYS

1. Agentic RAG transforms retrieval from a static lookup into a dynamic reasoning loop.

2. Start simple — make reasoning part of retrieval. Add evaluation first, then iterative re-retrieval, then additional sources.

3. Agentic RAG adds latency, cost, and complexity. Use it when queries require multi-step reasoning or cross-source synthesis — not as a default upgrade.

Questions?

Let's discuss.

Daniel Schroter Thüm

Senior Consultant / AI Engineer

daniel.schroter.eu