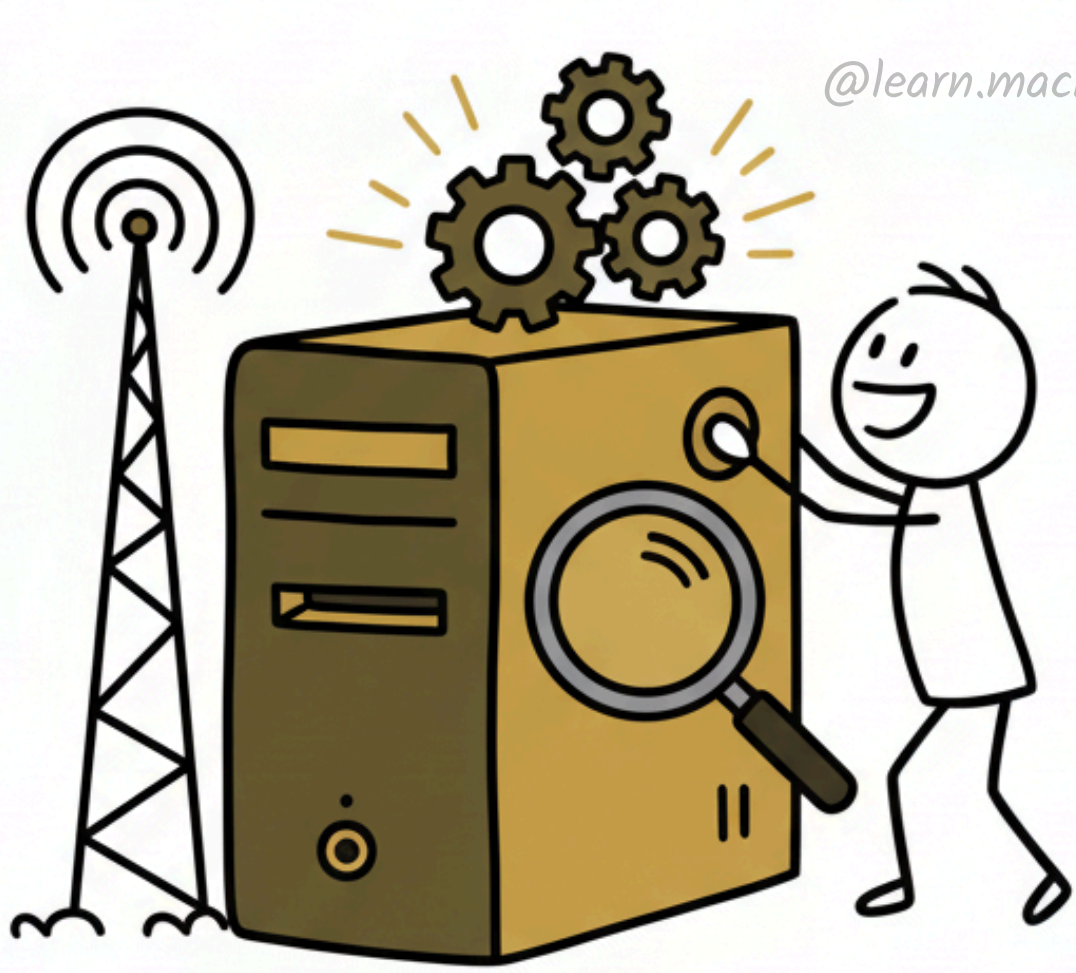
GENERATIVE AI SERIES: DAY 22

Advanced Retrieval: From Good to GREAT

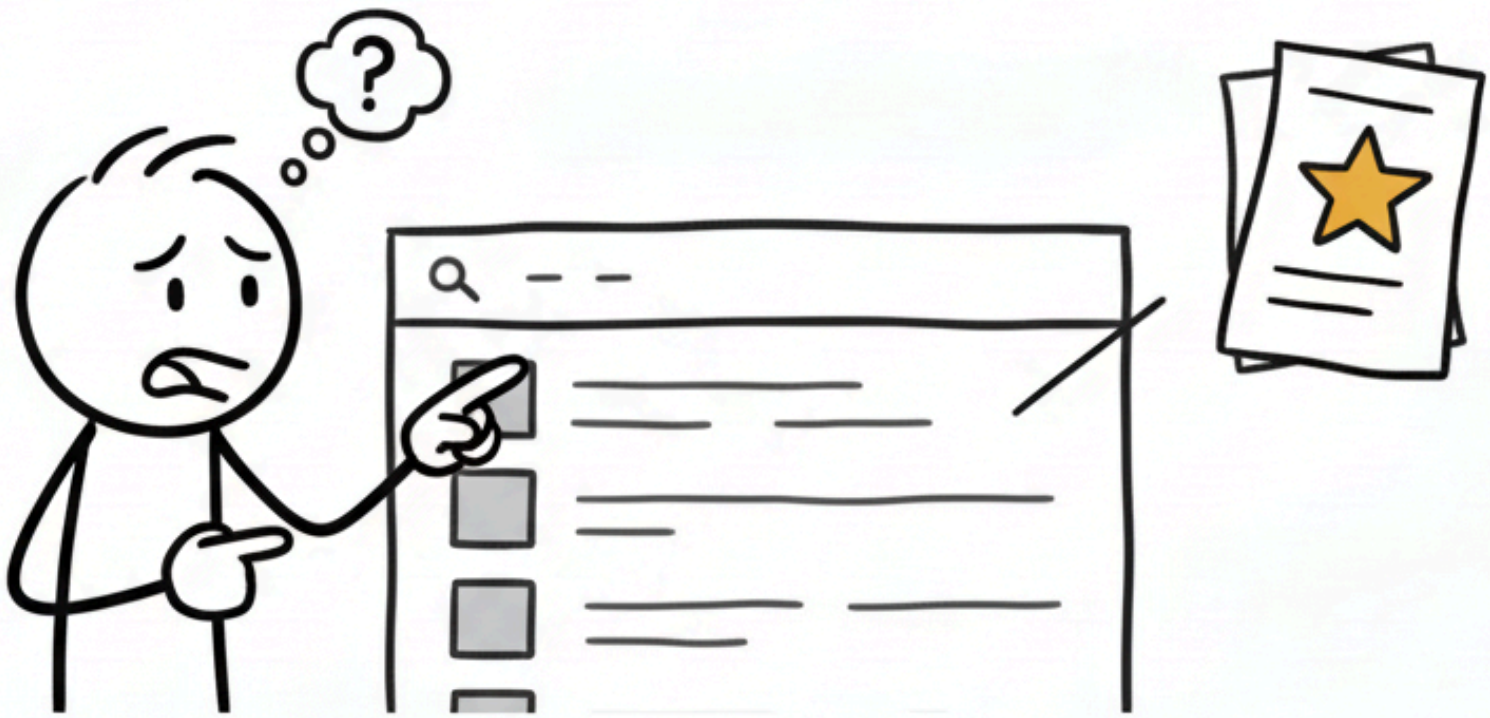
Your basic RAG system works, but how do make it find information with surgical precision? Time to install some powerful upgrades! 🚀



When Basic Retrieval Falls Short

- Sometimes, the most relevant information isn't the most "simillar" in a simple vector search.

- Vague Questions: "What's our company strategy?" might miss a key, recent memo because the wording is slightly different.
- Complex Questions: "Compare our Q1 and Q2 marketing spend for Project Apollo" requires finding multiple distinct pieces of information.

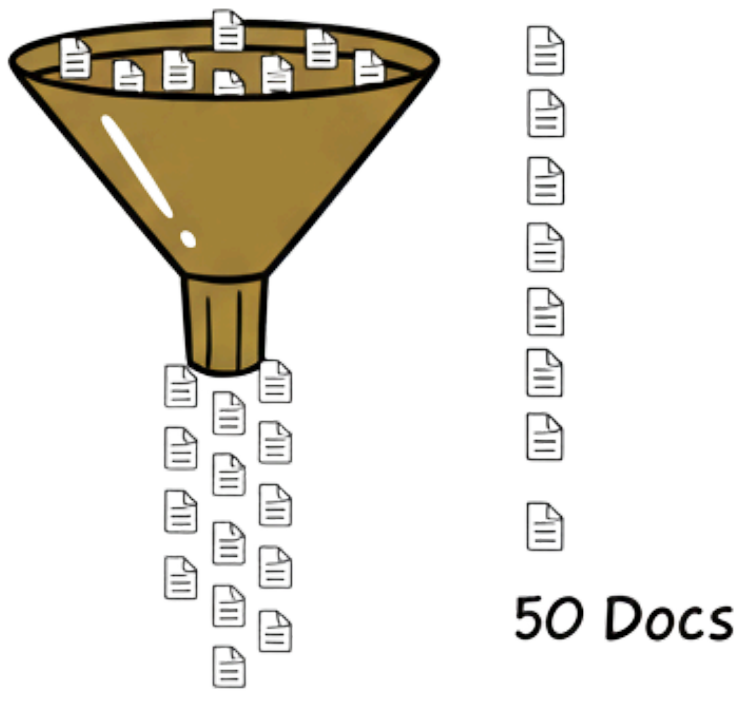


Adding a "Second Opinion": The Re-ranker

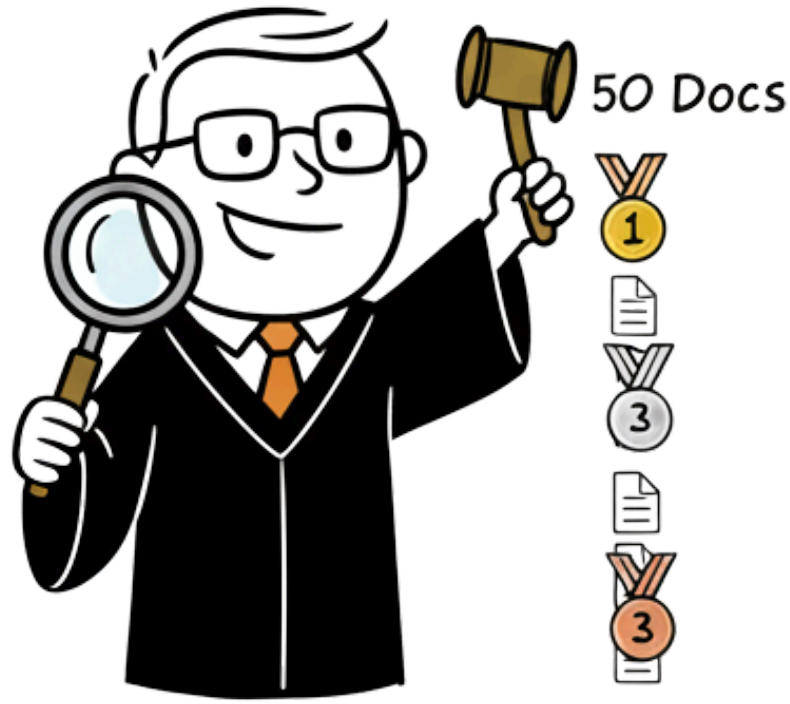
- A Re-ranker refines your initial search results.
- Retrieve Broadly: First, fetch a large number a potential documents (e. g., top 50) using fast vector search.
- Re-rank Precisely: Then, a more powerful (but slower) model closely examines the question against "each of the 50 documents and re-orders them based on true relevance.

- Result: The absolute best 3-5 documents rise to the top.

1. Initial Search



2. Smart Judge

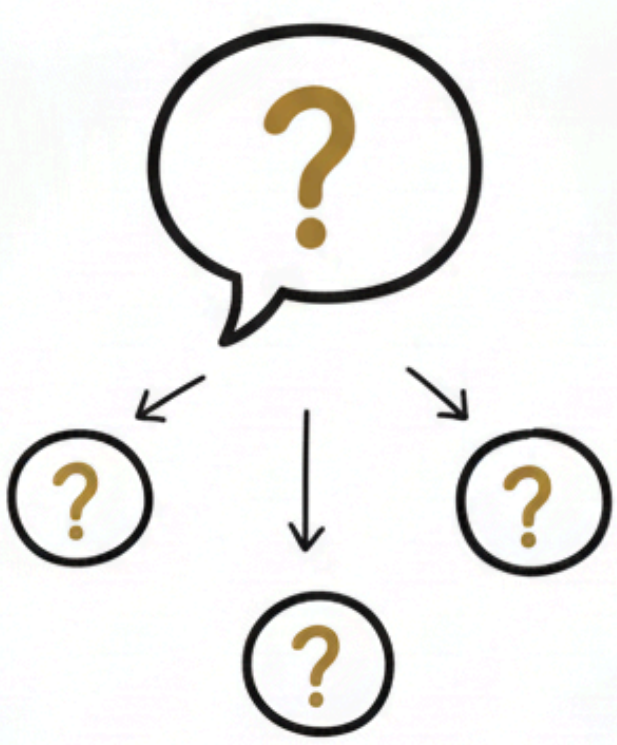


Upgrade 2: Query Transformations

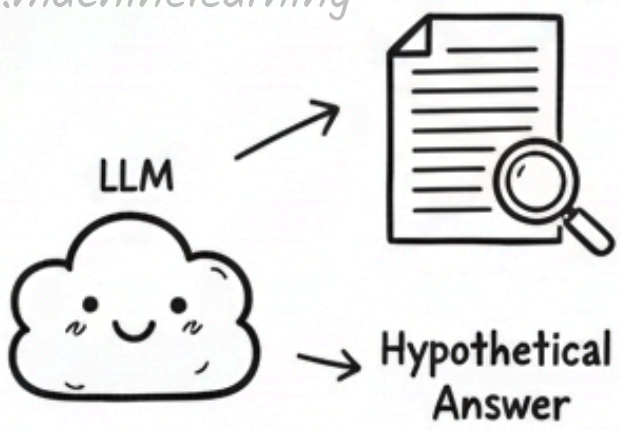
Don't Just Search Your Question... Search BETTER Ones!

Instead of using the user's exact query, we can transform it into somthing more effective for retrieveal.

Multi-Qury: An LLM rewrites a single complex question into several simpler sub-questions. We then search rag to answers to all of them.



Multi-Qury



HyDE (Hypothotical Document Embeding):

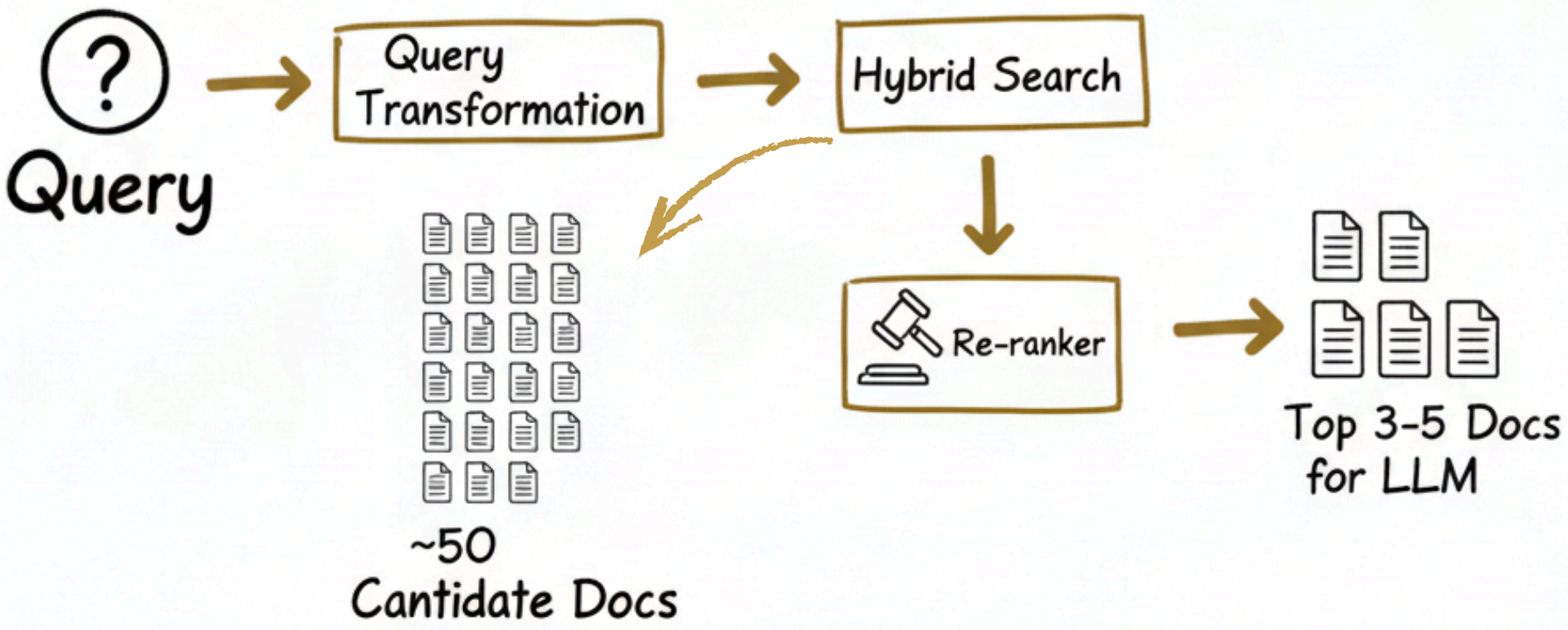
The LLM generates a perfect, hypthotical answer to a question first, and then **searches for real documents** that match that ideal answer.

The Full Advanced Blueprint

Layering Strategies for a Robust System

- Production-grade systems don't choose one method; they combine them into a powerful pineline.

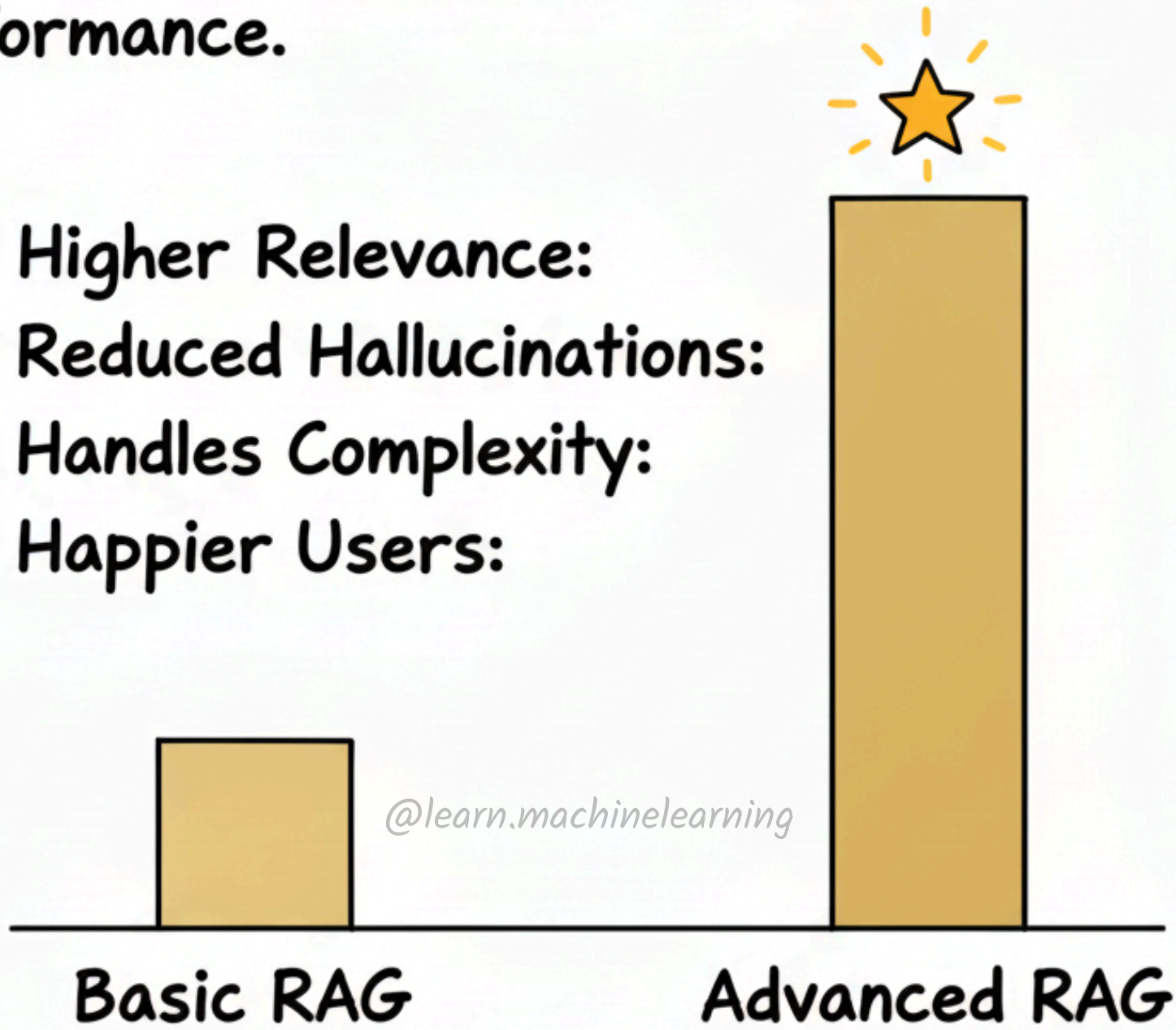
- Transform: Start with the user's query and generate sub-queries.
- Retrieve: Use Hybrid Search (vector + keyword) to get a broad set
- Re-rank: Use a re-ranker to score and select the absolute best documents from the candidate pool



The Impact: Why This Matters

Adding these advanced layers dramatally improves your RAG system's performance.

- ✓ Higher Relevance:
- ✓ Reduced Hallucinations:
- ✓ Handles Complexity:
- ✓ Happier Users:



GENERATIVE AI SERIES: DAY 22

Day 22 Takeaway: Refine Your Search, Refine Your Answers!

Moving beyond simple vector search is key to building a state-of-the-art RAG system. By layering techniques like **Query Transformations** and **Re-ranking**, you create a sophisticated retrieval process that finds the best possible context, leading to truly exceptional answers.

Next Up (Day 23):

Evaluating Your RAG System: Is It Actually Working?

@learn.machinelearning

What's the hardest question you think a simple search fail at? Drop it Drcep below! 👆

Follow @learn.machinelearning for the final lessons of our RAG Deep Dive!

SAVE this post for when 'wen're ready uo level up your AI project!