GENERATIVE AI SERIES: DAY 23

Evaluating Your RAG System: Is It Actually Working?

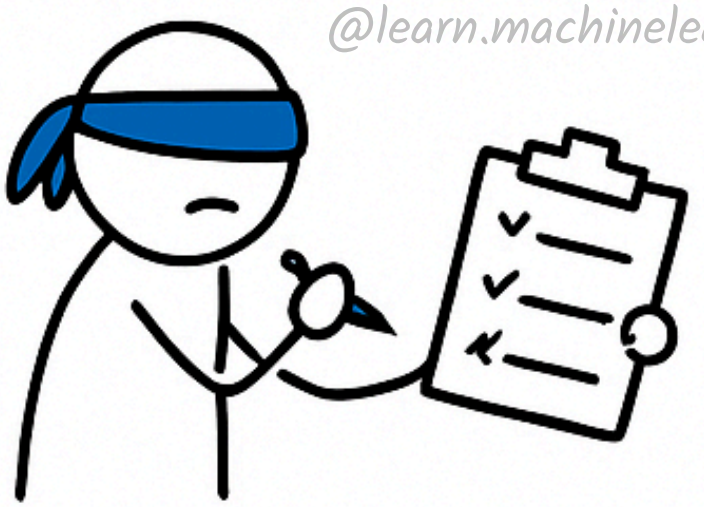
You've built your brilliant RAG system. But... how do you prove it's good? Let's dive into the most critical step: measurement. 📈

Without Data, You're Just Guessing!

Relying on a few test questions to see if your system works is like flying blind. For a reliable, professional system, you need objective numbers.

Guesswork

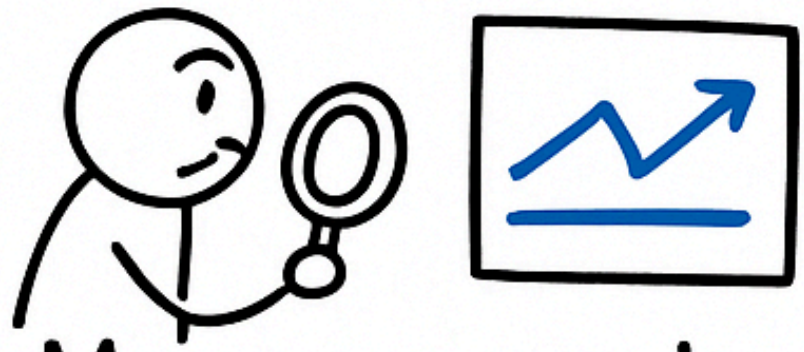
- The answers feel pretty accurate.



Guesswork

Measurement

- Our answer relevancy score is 0.92, an improvement of 15% after re-ranking.



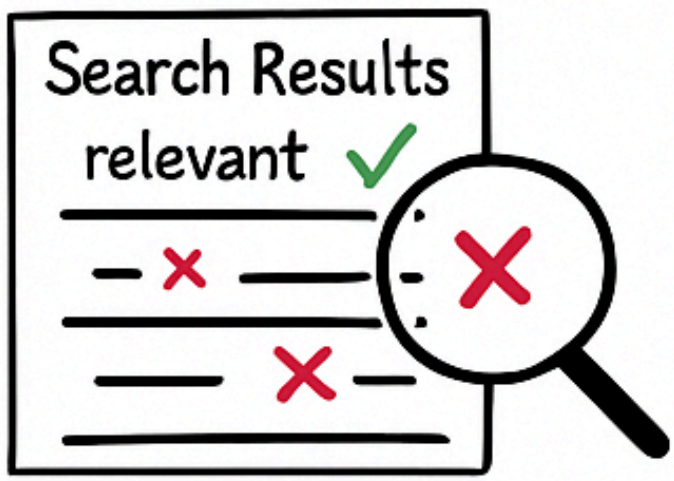
Measurement

Evaluation turns your gut feeling into hard data.

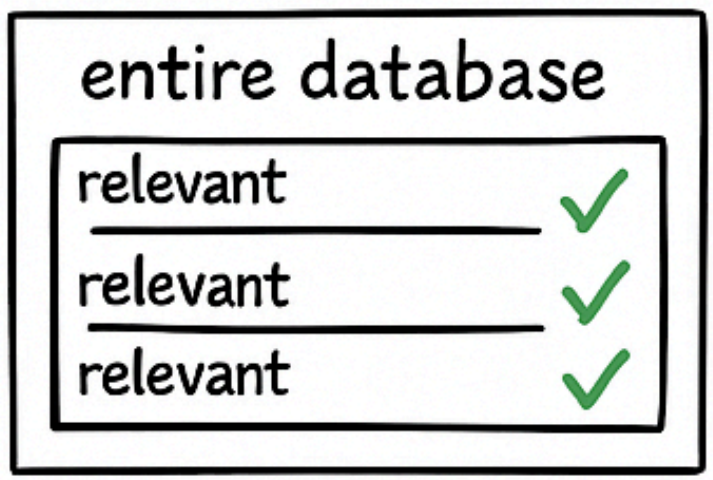
Metrics 1 & 2: Context Precision & Recall

Before we even get an answer, we must check the quality of our retrieved context.

- **Context Precision:** Of all the chunks we found, how many were **actually relevant**? (Are we finding junk?)
- **Context Recall:** Of all **the relevant** chunks that **exist** in the database, how many did we **actuallyfi**-nd? (Are we missing important info?)



Context Precision



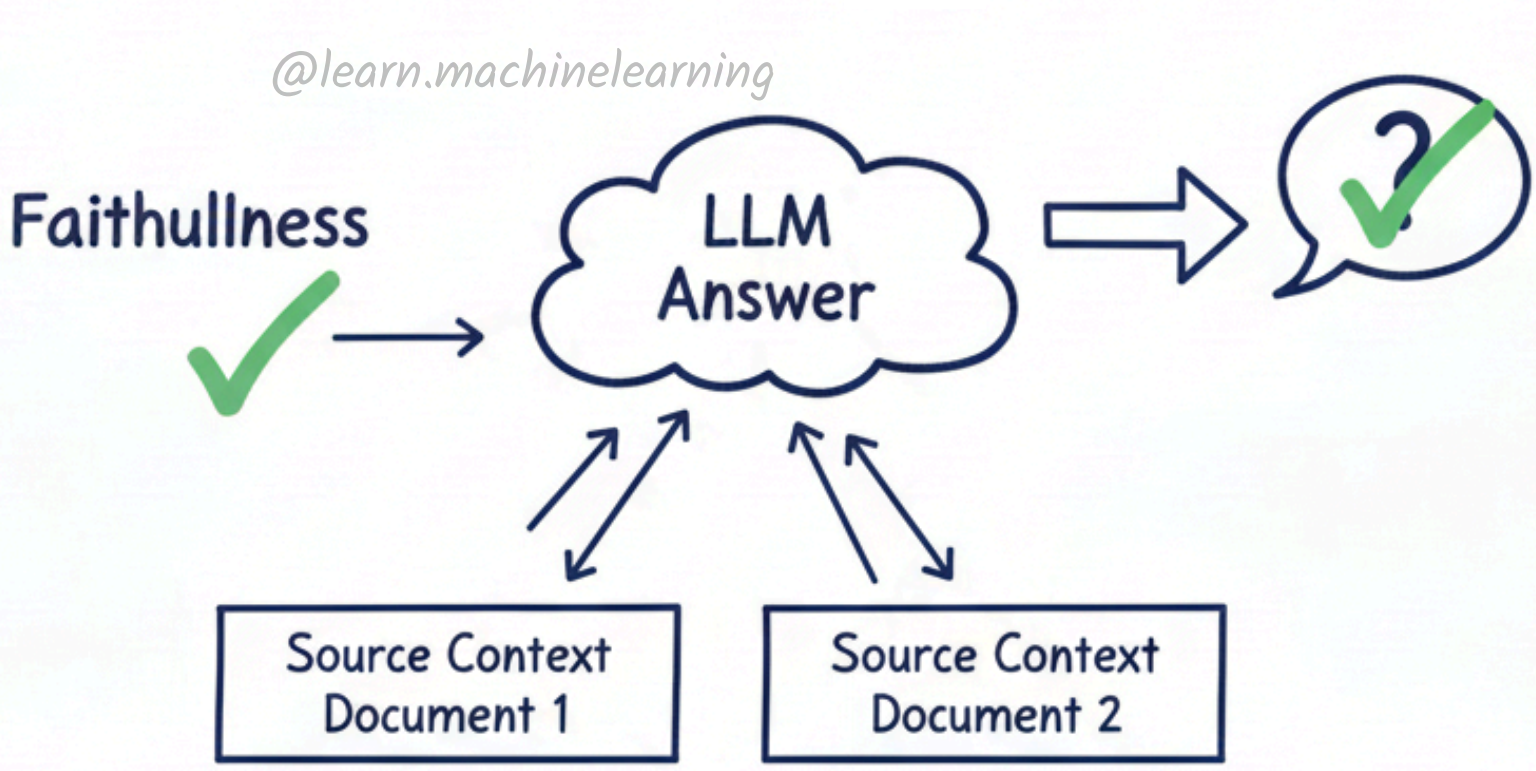
Context Recall

Goal: High Precision AND High Recall

Metrics 3 & 4:
Faithulness & Relevancy

Answer Faithfulness: Is the answer only based on the provided context? (Is at making stuff up / halluistating?)

Answer Relevency: Does answer directly adress the use user's original question? (Is it on-topic?)

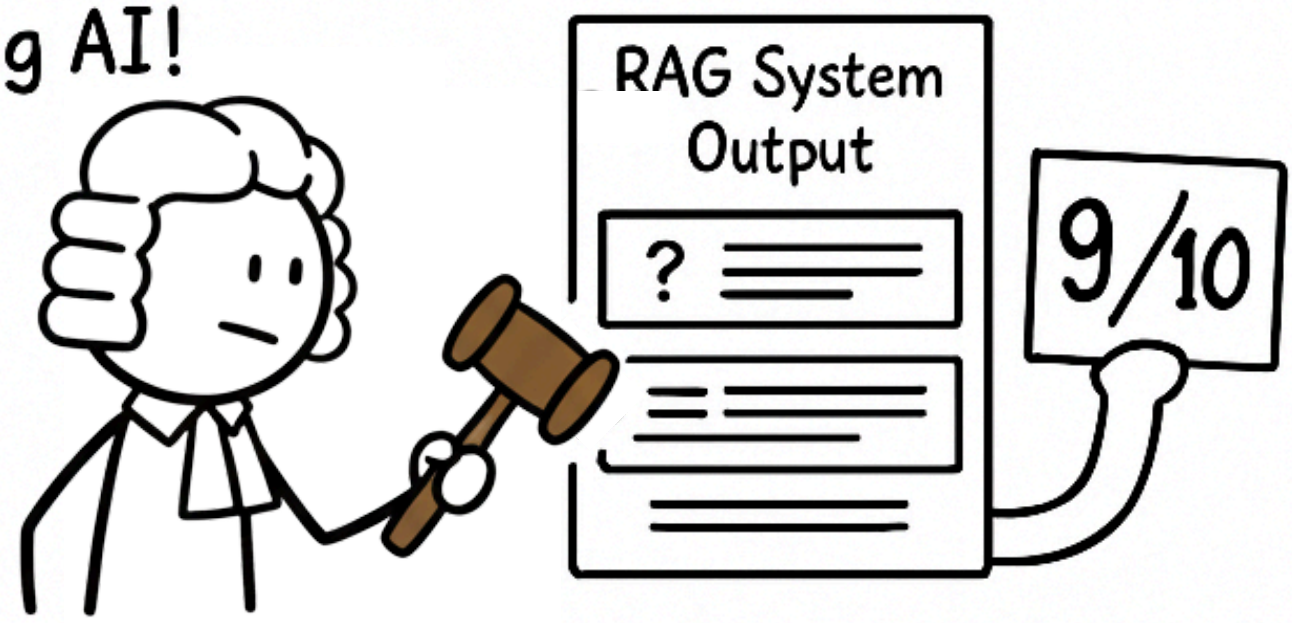


An answer can be factual but completely irrelvant!

How to Automate This? Meet RAGAs!

- Checking these metrics manually is impossible at scale. So, we use modern tools!

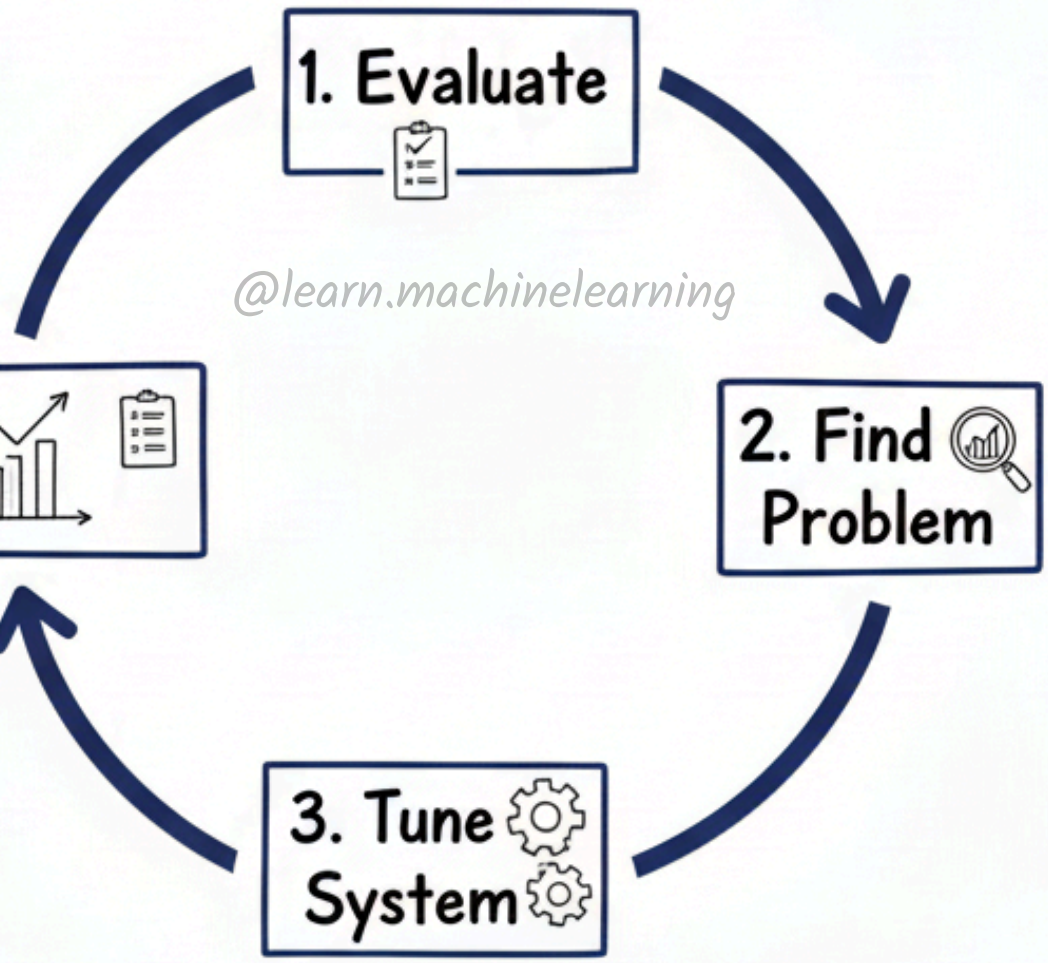
- ✔ RAGAs (RAG Assessment): An open-source framework that uses LLMs to evaluate other LLMs
- ✔ How it works: You provide the question, the retrieved context, and the final answer. RAGs then uses clever prompting to automatically score your system on faithfulness, relevancy, and more
- ✔ It's AI judging AI!



Build → Evaluate → Tune → Repeat

Evaluation isn't a one-time step;
it's a cycle.

Run Evaluation:
Identify Weakness:
Tune Your System:
Re-Evaluate



Day 23 Takeaway: What You Can't Measure, You Can't Improve!

Evaluating your RAG system with clear metrics
(like Precision, Recall, Faithfulness, and Relevancy)
is what transfroms a cool experiment into
a trustworthy, production-grade application.
It is the most critical step for building reliablde AI.

Next Up: Day 24: The Full RAG Blueprint Anateint: Anatomy of a Production-Ready System

From your experience, what's the biggest sign that an AI-powered is untrustworsthy? ⇩

Follow @learn.machinelearring for the final day of our RAG Deep Dive! Let's discuss!

SAVE this post! These metrics are the key to building quality AI. 💬 🔖