RAG Evaluation

Monday, August 4, 2025 10:14 AM

The LLM and RAG System has become a tool for helping the business.

Many people learn to develop the system, but sometimes they overlook the crucial part, which is the evaluation.

Assessing LLMs and RAG systems is vital to ensure that both retrieval and generation components function correctly for many reasons, including accurate response and high-quality results.

There are a few evaluation metrics for the LLM and RAG System, including:

↳ : Measures if the relevant retrieved context is ranked higher than the irrelevant ones.

↳ : Assesses how well the retrieved context aligns with the expected response (ground truth).

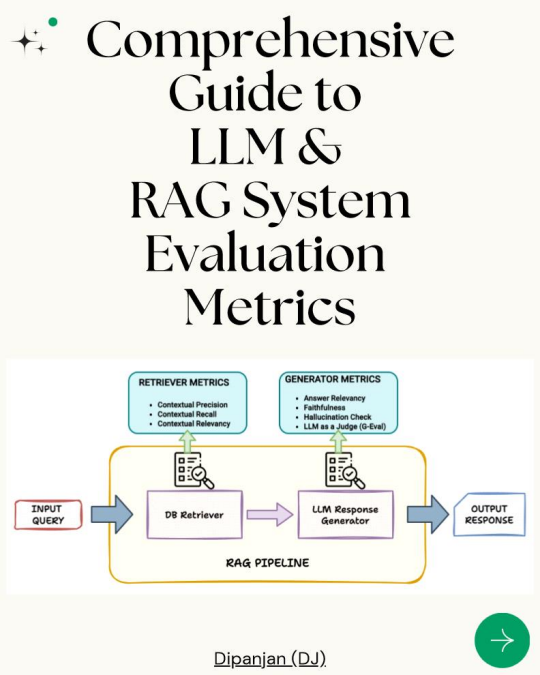
↳ : Evaluates the relevance of the retrieved context to the input query. ↳ ( - ): This metric checks whether the generated response is relevant to the input query using LLM evaluation.

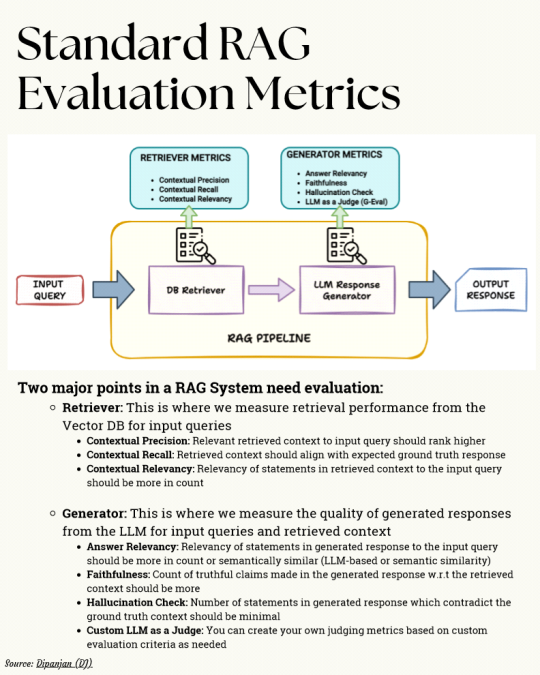
↳ ( - ): This method evaluates relevance by measuring the semantic similarity between the generated response and the input query.

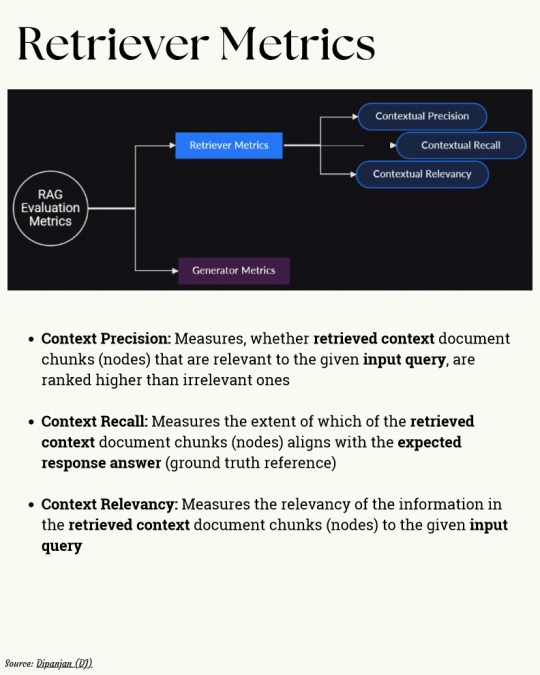
↳ : It makes sure that the generated response is factually accurate and grounded in the retrieved context.

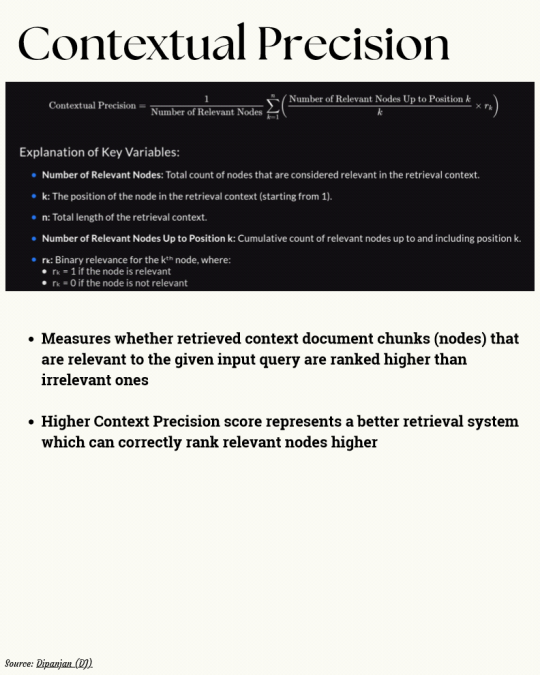
↳ : Measures contradictory or fabricated information in the generated response. ↳ - - - : Allows for custom evaluation criteria using LLMs with chain-of-thought (CoT) reasoning to assess responses.

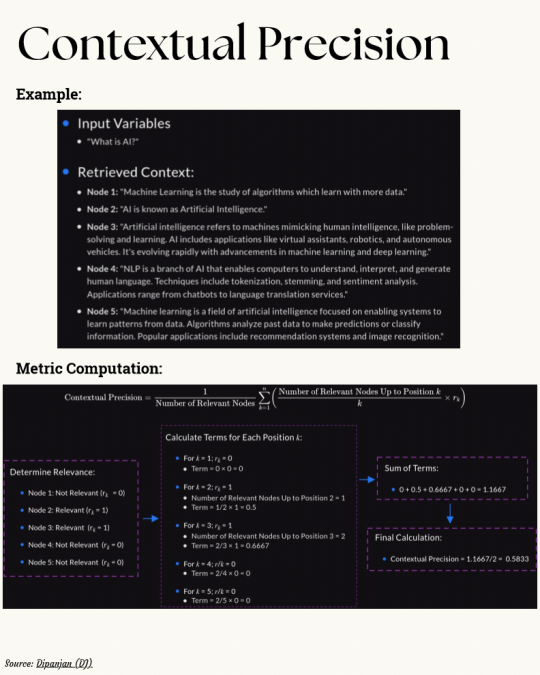
RAG Evaluation Page 1

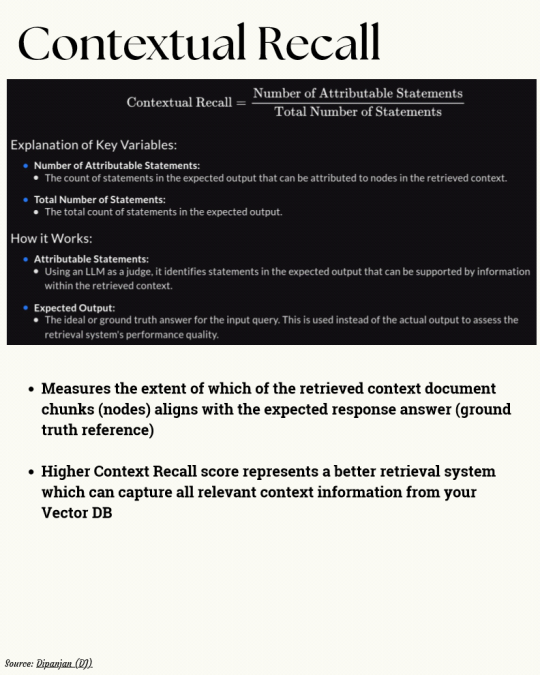
 RAG Evaluation Page 2

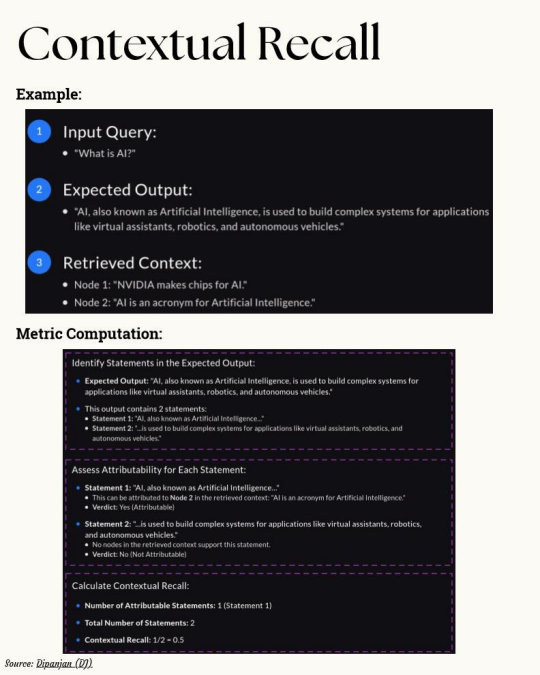
 RAG Evaluation Page 3

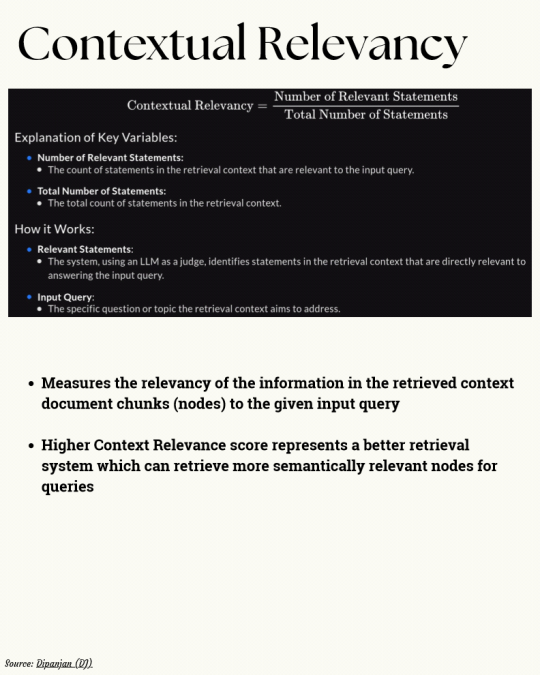
 RAG Evaluation Page 4

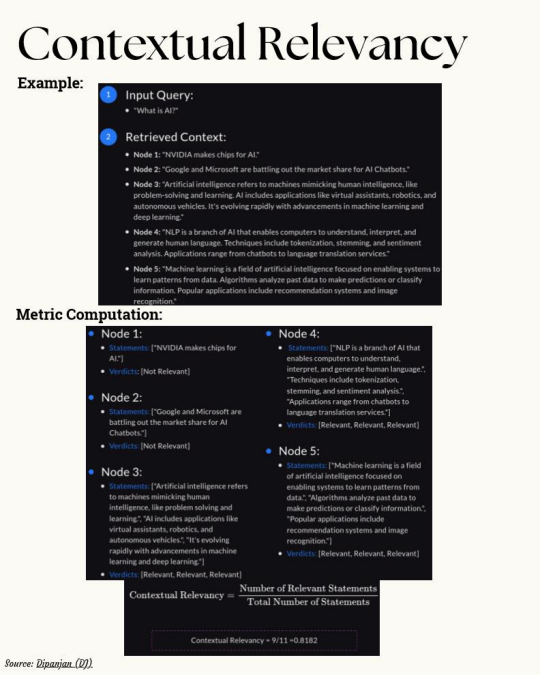
 RAG Evaluation Page 5

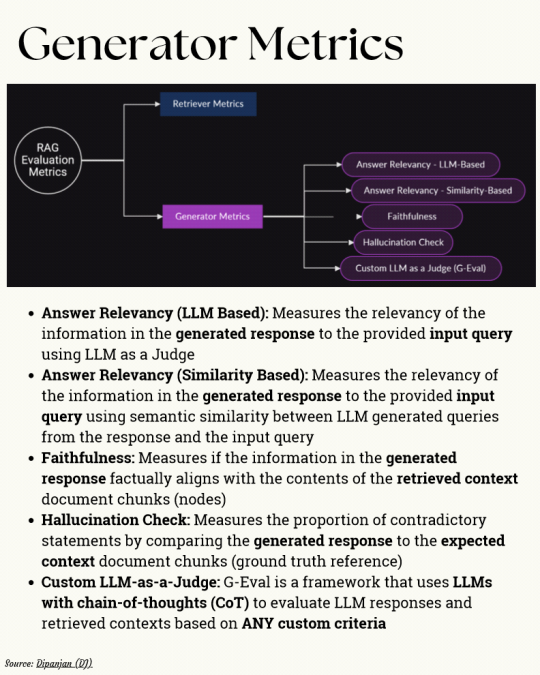
 RAG Evaluation Page 6

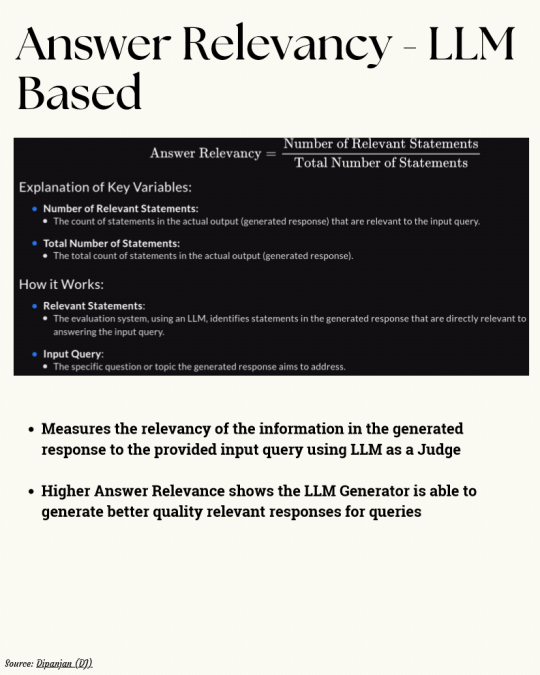
 RAG Evaluation Page 7

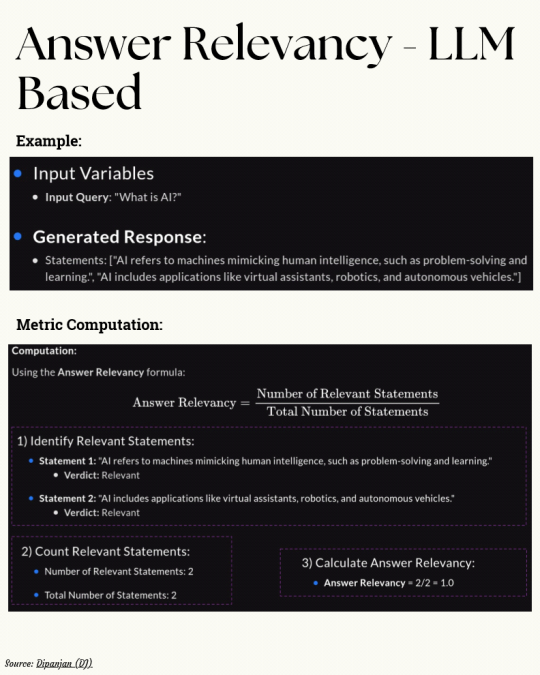
 RAG Evaluation Page 8

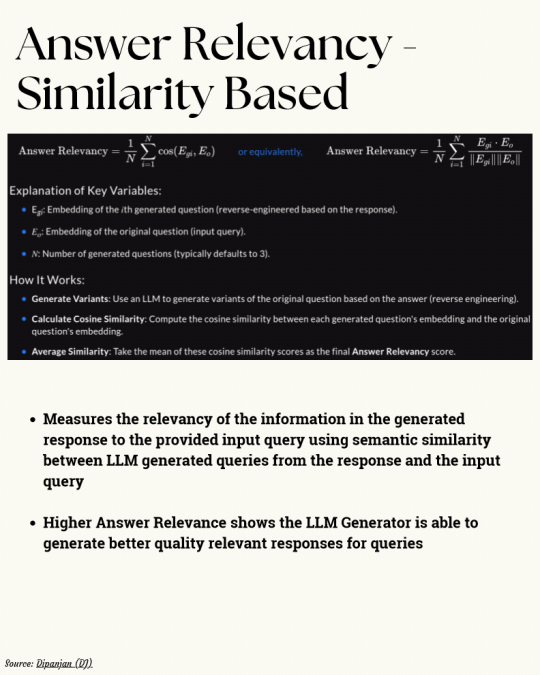
 RAG Evaluation Page 9

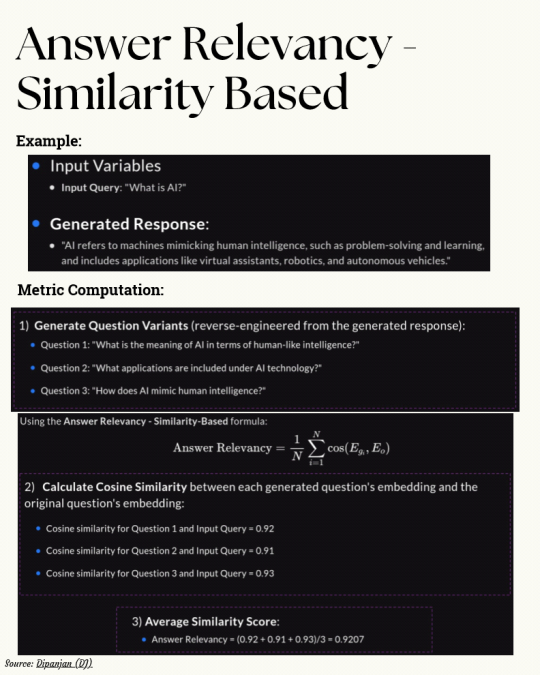
 RAG Evaluation Page 10

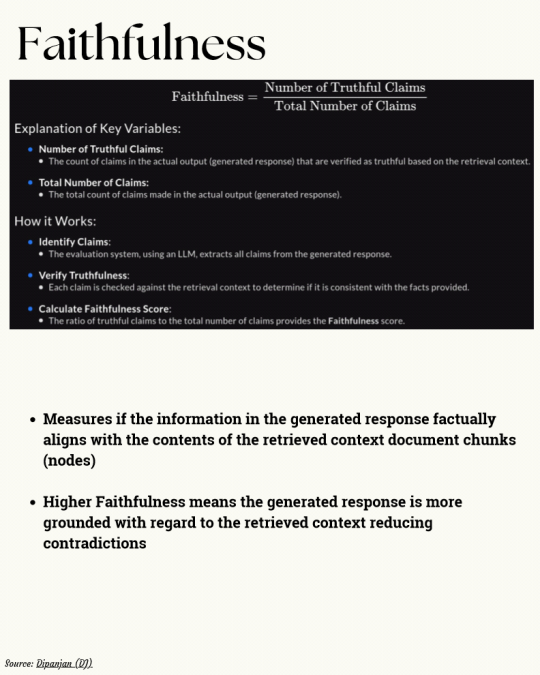
 RAG Evaluation Page 11

 RAG Evaluation Page 12

 RAG Evaluation Page 13

 RAG Evaluation Page 14

 RAG Evaluation Page 15

 RAG Evaluation Page 16

 RAG Evaluation Page 17

 RAG Evaluation Page 18

 RAG Evaluation Page 19

 RAG Evaluation Page 20

 RAG Evaluation Page 21

 RAG Evaluation Page 22

DeepEval

Wednesday, October 1, 2025 5:59 PM

The LLM Evaluation Framework

GitHub repo: https://github.com/confident-ai/deepeval

DeepEval is a simple-to-use, open-source LLM evaluation framework, for evaluating and testing large-language model systems. It is similar to Pytest but specialized for unit testing LLM outputs. DeepEval incorporates the latest research to evaluate LLM outputs based on metrics such as G-Eval, hallucination, answer relevancy, RAGAS, etc., which uses LLMs and various other NLP models that runs locally on your machine for evaluation.

Whether your LLM applications are RAG pipelines, chatbots, AI agents, implemented via LangChain or LlamaIndex, DeepEval has you covered. With it, you can easily determine the optimal models, prompts, and architecture to improve your RAG pipeline, agentic workflows, prevent prompt drifting, or even transition from OpenAI to hosting your own Deepseek R1 with confidence.

Our AI tests pass. Your agent still hallucinates.

The problem with end-to-end LLM testing:

→ Retriever pulls garbage

→ LLM hallucinates

→ Agent loops play telephone

→ Test fails. Zero clue why.

Component-level evaluation fixes this.

deepeval lets you test each piece:

Real-time visibility into every component

AnswerRelevancyMetric for generation

ContextualRelevancyMetric for retrieval

update\_current\_span captures what actually happened

pytest fixtures for LLMs

evals\_iterator tests all components at once

When it breaks, you see exactly where and why:

❌ "the AI is being weird"

✅ "retriever confidence -40% on financial queries"

RAG Evaluation Page 23

 RAG Evaluation Page 24

Most LLM-powered evals have this SERIOUS FLAW:

(and an open-source alternative to fix them)

LLM-generated evals can easily mislead you to believe that one model is better than the other, primarily due to the way they are set up.

For instance, typical procedures like G-E ’ g

time in isolation, without understanding the alternative.

S wh A 0.72 B 0.74, ’ k w wh h ’ .

This is unlike scoring, say, classical ML models, where metrics like accuracy, F1, or RMSE give a clear and objective measure of performance.

Th ’ j , h g h , not opinions.

LLM Arena-as-a-Judge is a new technique that addresses this issue with LLM evals.

In a gist, instead of assigning scores, you just run A vs. B comparisons and pick the better output.

Just like G-E , wh “ ” ( .g., h , concise, more polite), and use any LLM to act as the judge.

LLM Arena-as-a-Judge is actually implemented in DeepEval (open-source with 12k RAG Evaluation Page 25

LLM Arena-as-a-Judge is actually implemented in DeepEval (open-source with 12k stars), and you can use it in just three steps:

- C A T C , w h “ ” h LL interactions.

- Next, define your criteria for comparison using the Arena G-Eval metric, which incorporates the G-Eval algorithm for a comparison use case.

- Finally, run the evaluation and print the scores.

This gives you an accurate head-to-head comparison.

Note that LLM Arena-as-a-Judge can either be referenceless (like shown in the snippet below) or reference-based. If needed, you can specify an expected output as well for the given input test case and specify that in the evaluation parameters.

Why DeepEval?

It's 100% open-source with 12k+ stars and implements everything you need to define metrics, create test cases, and run evals like:

- component-level evals

- multi-turn evals

- LLM Arena-as-a-judge, etc.

Moreover, tracing LLM apps is as simple as adding one Python decorator. And you can run everything 100% locally.

RAG Evaluation Page 26

Evaluate multi-turn conversations with just a few lines of code!

DeepEval lets you build decision-tree based LLM-as-a-judge evals that break down

complex chats step by step.

Most LLM evaluations look only at the final response, giving a single score with

little context. That is not enough when real conversations span multiple turns.

Conversational DAGs (Directed Acyclic Graphs) let you create fully deterministic,

multi-turn evaluations.

You can combine different nodes to return hardcoded scores based on tasks,

judgments, and verdicts, building evaluation flows that are precise, transparent,

and auditable.

H ’ wh :

• Summarize long conversations before scoring

RAG Evaluation Page 27

• Summarize long conversations before scoring

• Add binary checks like “Did the assistant answer the question?” • Add multi-class checks like “Was the tone Rude, Neutral, or Playful?” • Combine these into a deterministic flow that produces clear, auditable scores

This gives you transparency and precision in evaluating entire conversations, something black-box metrics can’t provide.

Perfect for testing agents, chatbots, or any system where both accuracy and behavior matter.

It’s 100% Open Source.

RAG Evaluation Page 28

RAG Evaluation

Tuesday, May 20, 2025 1:03 PM

 RAG Evaluation Pae 29

|  |
| --- |

RAG Evaluation Pae 30

|  |
| --- |

RAG Evaluation Pae 31

|  |
| --- |

RAG Evaluation Pae 32

|  |
| --- |

RAG Evaluation Pae 33

|  |
| --- |

RAG Evaluation Pae 34

|  |
| --- |

RAG Evaluation Pae 35

|  |
| --- |

RAG Evaluation Pae 36

|  |
| --- |

RAG Evaluation Pae 37

|  |
| --- |

RAG Evaluation Pae 38

RAG Evaluation Guide

Monday, August 11, 2025 8:57 PM

RAG Evaluation Pae 39

 RAG Evaluation Pae 40

 RAG Evaluation Pae 41

 RAG Evaluation Pae 42

 RAG Evaluation Pae 43

 RAG Evaluation Pae 44

 RAG Evaluation Pae 45

 RAG Evaluation Pae 46

 RAG Evaluation Pae 47

 RAG Evaluation Pae 48

 RAG Evaluation Pae 49

 RAG Evaluation Pae 50

 RAG Evaluation Pae 51

 RAG Evaluation Pae 52

 RAG Evaluation Pae 53

 RAG Evaluation Pae 54

 RAG Evaluation Pae 55

 RAG Evaluation Pae 56

 RAG Evaluation Pae 57

RAG Evaluation

Friday, August 22, 2025 7:31 PM

Most teams only evaluate the final output.

Th ’ k g g x w h k w g wh q

were asked.

L ’ x h .

H ’ -level RAG evaluation framework for real

world builders.

Every RAG pipeline has 3 critical stages:

→ Are we surfacing the right documents?

→ Is the context high signal and aligned with the query?

→ Is the output fluent, grounded, and accurate?

: Find and rank the most relevant documents.

:

• Relevant info not retrieved

• Retrieved docs drift semantically

• Relevant docs ranked too low

:

• @ / @ — Measures top-K relevance and

coverage

• — How early relevant docs appear

• — Are the most relevant docs ranked highest?

:

↳ Low Recall → Improve embeddings, chunking, or query

rewriting

↳ Low NDCG → Add rerankers or improve scoring logic

: Make sure retrieved context is useful, complete, and

structured.

:

• Topical chunks don’t answer the query

RAG Evaluation Pae 58

:

• Topical chunks don’t answer the query

• Too much noise, not enough signal

• Key facts lost due to truncation or poor chunking

• Weak grounding leads to hallucination

:

• — Does the context match the query intent?

• — Is the output aligned with the retrieved evidence?

• — Signal vs. noise in input

• — Are all required facts included?

:

↳ High relevance + low coverage → Improve semantic chunking

↳ High hallucination despite context → Restructure inputs or

refine prompts

: Generate fluent, accurate, and well-grounded responses.

:

• Fluent but factually wrong

• Poor use of context

• Repetition, incoherence

:

• / — Lexical similarity to references

• — Predictability and fluency

• — Unsupported claims frequency

• — Fluency, grounding, and real-world clarity

:

↳ Hallucination despite good inputs → Add stronger grounding cues or use prompt templates

↳ High perplexity → Simplify prompts or upgrade model quality RAG Evaluation Pae 59

|  |
| --- |

RAG Evaluation Pae 60

|  |
| --- |

RAG Evaluation Pae 61

|  |
| --- |

RAG Evaluation Pae 62

|  |
| --- |

RAG Evaluation Pae 63

|  |
| --- |

RAG Evaluation Pae 64

|  |
| --- |

RAG Evaluation Pae 65

|  |
| --- |

RAG Evaluation Pae 66

|  |
| --- |

RAG Evaluation Pae 67

|  |
| --- |

RAG Evaluation Pae 68

RAG Evaluations

Tuesday, September 23, 2025 1:49 PM

RAG Evaluations: Your Cheatsheet for Success

Are you curious about how we measure the effectiveness of Retrieval-Augmented Generation (RAG)? The metrics behind it are a treasure trove of insights that can elevate your understanding and drive results in your AI projects.

Here's your condensed guide to RAG Evaluations:

RAG Evaluation Page 69

1. Introduction

- RAG Evaluation measures retrieval and generation effectiveness. - It ensures responses are accurate, relevant, grounded, and hallucination-free.

. Key etrics

- Retrieval: Precision, Recall, MRR, NDCG.

- Generation: BLEU, ROUGE, BERTScore, COMET.

- Overall: Faithfulness, Groundedness, Relevance.

. Retrieval valuation

- Exact Match

- Re-ranking

- Embedding Similarity

. eneration valuation

- Lexical: BLEU, ROUGE.

- Semantic: BERTScore, embeddings.

- Human: Faithfulness, coherence.

. roundedness Faithfulness

- Groundedness → Based on retrieved evidence.

- Faithfulness → No hallucinations.

. Automated vs uman valuation

- Automated: Fast, scalable, less accurate.

- Human: Detailed but subjective.

- Hybrid: Combines strengths of both.

. Tools for RA valuation

- LangChain Eval

- TruLens

- LLM-as-a-Judge

- Hugging Face Evaluate

- Ragas

. est Practices

- Use multiple metrics for comprehensive insights.

- Benchmark various LLMs for better comparison.

- Incorporate human review to enhance accuracy.

Mastering RAG evaluations opens doors to more accurate and impactful AI applications. Have you started implementing RAG in your projects? What metrics do you find most effective?

RAG Evaluation Page 70

RAG Evaluation Frameworks

Thursday, October 16, 2025 10:41 PM

RAG evaluation frameworks all look the same until your production system fails

According to the State of AI report 2025, 73% of teams admit their RAG evaluation is "ad-hoc at best." The real problem? There are 5+ frameworks out

RAG Evaluation Pae 71

evaluation is "ad-hoc at best." The real problem? There are 5+ frameworks out there, and most developers pick one randomly without understanding what they actually measure.

The reality is: Deepchecks isn't the same as RAGAS. TruLens solves different problems than Future AGI. Using the wrong framework means you're measuring the wrong things.

Scope overlap doesn't imply interchangeability ❌

The pattern most miss:

Building your first RAG? Start with RAGAS for research and pipeline testing. Already in production? Deepchecks or TruLens for continuous monitoring.

Debugging complex workflows? Future AGI traces every decision your system makes.

Running CI/CD pipelines? DeepEval integrates directly into your automation. The framework you choose depends on where you are in your RAG journey, not which one sounds most impressive/familiar.

RAG Evaluation Pae 72

RAG Evaluation

Wednesday, October 22, 2025 7:32 PM

F , ’ A E guide, https://lnkd.in/eSzXfGYi

 RAG Evaluation Pae 73

 When incorporating a new AI pipeline component, thorough evaluation and testing are essential. In the realm of AI, testing goes beyond evaluation by utilizing crucial data and metrics from offline assessments to establish a test framework for monitoring pipeline modifications. While metrics for workflow agents or prompting systems hold significance, this post is on introducing metrics for RAG components.

RAG systems are evaluated on both retrieval and generation. Retrieval evaluation assesses how efficiently relevant information is extracted, and generation evaluation examines the quality of the AI model's final response. Both are crucial for reliable results.

Outlined below diagrams are some fundamental RAG system evaluation metrics. RAG Evaluation Pae 74

Further we can classified metrics into four categories: exact match, semantic match, fuzzy, and LLM judge metrics,

Exact Match Metrics (Word-for-Word):

\* Precision@K

\* Recall@K

\* MRR

\* NDCG@K

\* MAP

Fuzzy Match Metrics (Keyword-Based):

\* BLEUScore

\* COMET

Semantic Match Metrics (Meaning-Based):

\* BERTScore

\* COMET

\* Factual Consistency Score

LLM judge Metrics

\* COMET is a key LLM judge metric, Utilizes a neural model, trained on human ratings, to assess how well generated output aligns with desired quality and meaning

\* Factual Consistency Scores:These are determined by using LLMs to compare generated content against verifiable data, ensuring outputs do not contain fabricated or unsupported claims.

\* Prompt Effectiveness : often relies on an LLM to gauge whether a prompt elicits accurate, complete, and contextually appropriate responses.

RAG Evaluation Pae 75

RAG Evaluation Frameworks

Friday, November 21, 2025 7:19 PM

g h g h A … but real AI reliability comes from systematic evaluation.

This cheat sheet breaks down the top evaluation frameworks by Sairam, every AI engineer should know, especially if you're building agents, RAG systems, or end-to end LLM workflows.

H ’ h k w

h k

Best for end-to-end audits of agent planning and responses.

→ Works well with LangChain and OpenAI pipelines

→ Great for multi-agent QA regression tests

→ Needs system logs, retrieved contexts, and agent outputs

A AS

Best for evaluating retrieval quality and grounding accuracy.

→ Measures context precision, recall, and faithfulness

→ Ideal for comparing retrievers and benchmarking RAG setups → Uses reference answers, contexts, and question sets

T L

Best for RAG monitoring and iterative feedback optimization.

→ Tracks context relevance, safety, and groundedness

→ Built for LangChain, LlamaIndex, OpenAI APIs

→ Needs trace logs, feedback functions, and context logs

F A

Best for workflow debugging and agent reliability audits.

→ Tracks tool provenance, consistency, and traceability

→ Integrates with observability and tracing tools

→ Uses system traces and tool-call histories

E

Best for automated testing in development pipelines.

→ Regression, factuality, and critique-based evaluations

→ CI/CD friendly and works via Python test suites

→ Uses test cases, model outputs, and expected answer sets

If you want your RAG pipelines, agents, or multi-step workflows to perform RAG Evaluation Pae 76

If you want your RAG pipelines, agents, or multi-step workflows to perform consistently, these five frameworks give you the structure and metrics you need.

|  |
| --- |

RAG Evaluation Pae 77

|  |
| --- |

RAG Evaluation Pae 78

|  |
| --- |

RAG Evaluation Pae 79

|  |
| --- |

RAG Evaluation Pae 80