Qdrant -> Vector DB

Why? Metadata Filtering

Vector DB
- col 1
- col 2
- col 3
- col 4

Memory

In memory (RAM) (No Disk)

Available in Runtime Only

Collection :- Vector Dim (786)

Distance Cosine Distance

Vec 1
786

Vector 2
1536

RAG Evaluation Metrics

Metrics ⟶ Frameworks

RAGAS -> RAG Assessment

* DeepEval *

* Evaluation Dataset *

Metrics

Faithfulness :- Generated content stick to the retrieved facts.

Check if the LLM is hallicinating

Faithful → Source Material.

Examples

Input :- When was the first super bowl?

Response :- Jan 15, 1967

Actual LLM Response :- The first super bowl hosted on Jan 15, 1967 in Los Angelos.

Retrieved Text :- The first AFL-NFL World Championship was played on Jan 15, 1967. in LA.

1) Extract all claims from generated answer
2) Check if each claim is supported by retrieved context.
3) Scoring Mechanisms

$$\text{Score} = \frac{\text{claims supported by Retrieved context}}{\text{Total claims}}$$

Claims	Score
1) 1st S.B was Jan 15, 1967	(supported)
2) It was in LA	(supported)
3) It was a sunny day.	(not supported) not in context LLM is hallucinating

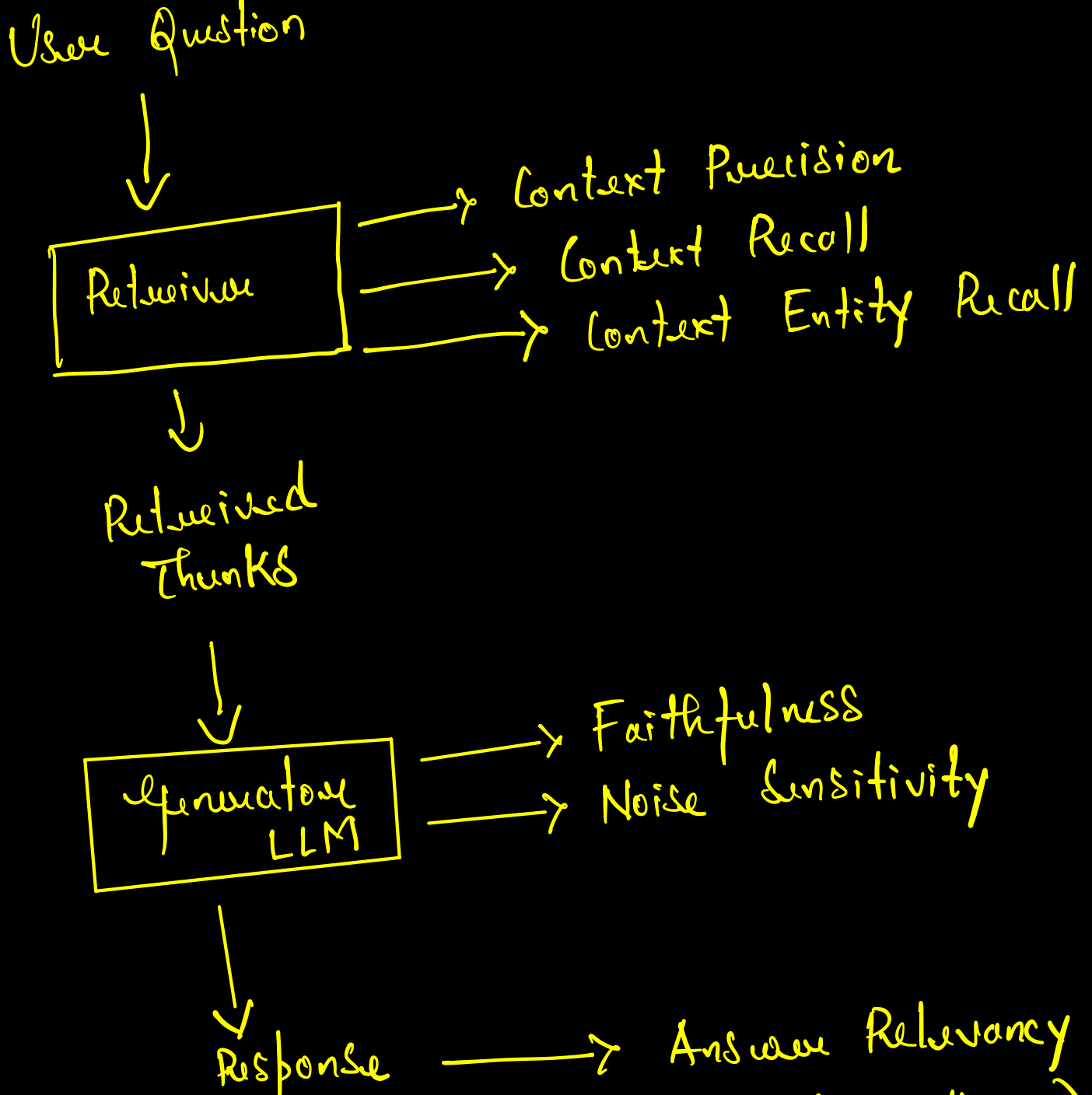
$$\text{Score} = \frac{2}{3} = .67$$

Score Interpretations

1 : Perfect (All claims have R:C)

0 : Answer made up by LLM.

RAG Metrics Mental Model



(address the question)

Response Answer Relevancy

Measures :- Address whether the answer actually answering the question

Q1 What is the capital of France?

R1 The Effiel Tower is beautiful. (Totally Irrelevant but factual partially true)

R2. Paris is a beatiful city

Working :-

1) Generate hypothetical questions that the answer would be suitable.
2) Compare the generated questions with the original questions (Embedding)
3) * Higher similarity = more relevant answer
   * Lower similarity = less relevant answer

Answer Relevancy

1 : Answers directly to the addressed question

0 : completely off-topic answers

0-1 :- Partially off-topic.

Good Relevancy

User Input :- When was the 1st S.B ?

Response :- The finest S.B was held on Jan 15, 1967

R-C :- The finest AFL-NFL Championship was played on 15th Jan, 1967 in LA

R2 :- The S.B is the annual championship game of NFL.

✓ What / ✗ When

Score :- (0.35)

Context Precision

Where is Effiel Tower?

The E.T located in Paris

R.C :- 1) The ET is locate Paris

2) Paris is the capital of France

3) The ET was built in 1889.

4) Pizza originated Italy.

Context Precision

Level :- Retrieval

Measure :- Are the retrieved chunks ranked properly?

Checks the top K highly relevant chunks

Q1

$c_1 \rightarrow w_1$ high

$c_2 \rightarrow w_2$ low

$c_3 \rightarrow w_3$ low

$c_4 \rightarrow w_4$ low

Working :-

1) For each retrieved chunk, determine whether it is relevant to the question

2) Calculate the precision at each position? (weightage by position)

3) Relevant chunks at the top = higher score
Irrelevant chunk at the top = lower.

Input

Where is ET located?

Reference: The ET is in Paris.

R.C = [The ET is located in Paris] Highly ✓✓

[Paris is the capital of France] Some relevant

[The ET was built around 1889] IRRelevant

[Pizza originated in Italy] Completely irrelevant

Context Recall

Level :- Retrieiver

Measures :- Did we retrieve all the necessary info to answer the question.

All chapters of book ⟶ Give an exam

Working :-

1) Breaking down of reference answer into individual claims
2) Check if the claim is attributed to the R·C
3) Score = $\frac{\text{Claims found in the context}}{\text{Total claims in reference response}}$

Example :-

Input :- Tell me about ET

Response :- The ET is in Paris, built in 1889.

R·C = [The ET is a landmark located in Paris (F1)]

[The ET was completed in 1889]

Reference Claims :-

1) location ✓
2) Built ✓

Score = 2/2 = 1

Score Interpretation

1 :- Retrieved contexts are provide the full answer

0 : R-C is completely useless.

Context Recall

Measures :- Did we retrieve the context containing all necessary entities (people, place, org, date)

mentioned in the reference answer.

Analogy :-

Reference :- Paul was going to London on 17 Jan, 2025.

3 Entities

Person :. Paul

Location :. London

Date :. 17 Jan, 2025

Extraction

↓

Reference

Working

1) Extract Named Entity from Reference

2) Extract Name Entity from R.C

3) Score = $\frac{\text{Entities in both (matching with Ref) common}}{\text{Entities in reference}}$

Example:-

Ref: Albert Einstein developed relativity in P.U in 1905.

R.C = [Albert Einstein was a famous physicist in PU.]

A, Entities in Ref :- [[A,B], [PU], [Date]]

B Entites in R.C :- [[A,B], [PU]]

Score :- $\frac{2}{3} = 0.67$

Noise Sensitivity (unwanted)

Jitter

Level :- Generation

Working:-

① Compare the reference answer with generated answer.

② Identify incorrect claims in the response.

③ Determine if the incorrect claims came from (irrelevant) retrieved chunks
noise

④ Score : $\frac{\text{Incorrect Claims from noise}}{\text{Total Claims}}$