Qdrant $\rightarrow$ Vector DB

Vector BB
- $\rightarrow$ col 1
- $\rightarrow$ col 2
- $\rightarrow$ col 3
- $\rightarrow$ col 4

Why ? Metadata Filtering

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 $\longrightarrow$ Frameworks

RAGAS $\rightarrow$ 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 :- The first super bowl hosted on Jan 15, 1967
Response   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

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

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)
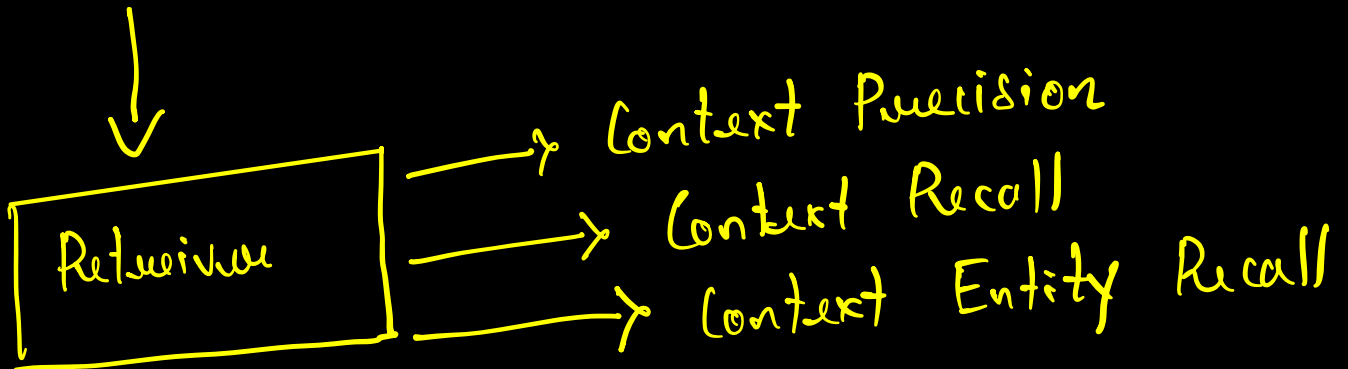   not in context
   LLM is hallucinating

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

Score Interpretations

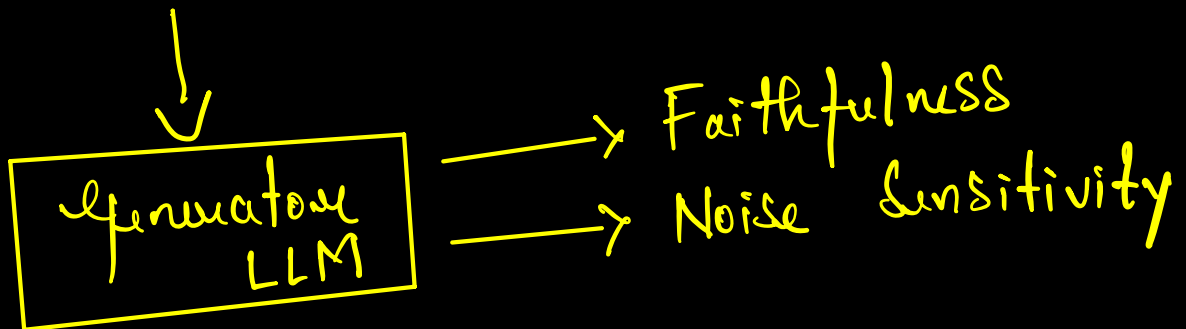1 : Purfect (All claims have R.C)

0 : Answer made up by LLM.

RACG Metrics Mental Model
_____

User Question

↓

┌─────────────┐
│   │ ────→ Context Puecision
│ Retueivee │ ────→ Context Recall
│   │ ────→ Context Entity Recall
└─────────────┘

↓

Retueived
Thunks

↓

┌─────────────┐
│ Genmatoue   │ ────→ Faithfulness
│   LLM   │ ────→ Noise Sensitivity
└─────────────┘

↓

Response ────→ Answee Relevancy

Response
Answer Relevancy

Measures :- Address whether the answer actually answering the question

Q1 What is the capital of France?

R1 The Eiffel Tower is beautiful.

R2. Paris is a beatiful city

(Totally Irrelevant but factual partially Irrel)

Working :-

1) Generate hypothetical answer 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 first S.B was held on Jan 15, 1967?

R.C :- The first 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 1889r

4) Pizza originated Italy.