key terminologies in RAG LLM:

- **User Query**:
  The natural language question or input provided by the user, which initiates the RAG process to retrieve relevant contexts and generate a response..
- **Response**:
  The final answer or output generated by the LLM based on the user query and retrieved contexts.
- **Retrieved Contexts**:
  The top-K relevant documents or snippets fetched from the knowledge base to support the LLM in generating a grounded response.
- **Reference (Ground Truth)**:
  The factual, verified answer or dataset used as a benchmark to evaluate the correctness and relevance of the response and retrieved contexts.

RAG LLM Testing Scenarios

Document Retrieval Testing

- **Scenario**: Test whether relevant documents are retrieved.
  Metric:-Context Recall ensures that the retrieval system **does not miss any relevant information**, even if individual documents might be less semantically similar.

- **Scenario**: Test how well the LLM integrates retrieved documents into the response.
  Metric : Context Precision - Assesses the proportion of retrieved context used accurately in the response.

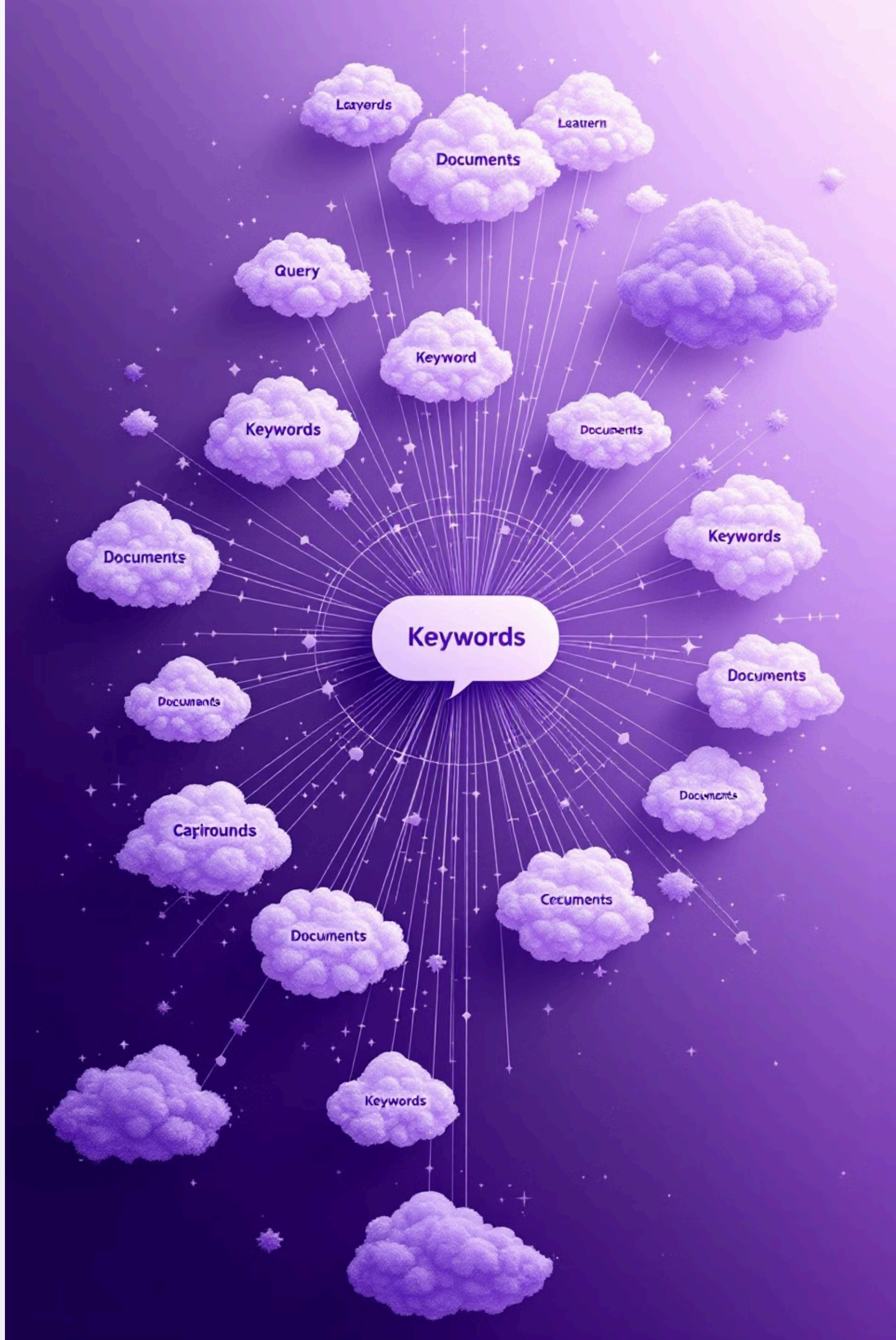
User Input and Document Matching

□ **Scenario: Test semantic similarity between user query and retrieved documents.**

Metric: Measures how semantically close the generated response is to the expected answer.

□ **Scenario: Test relevance of retrieved documents to user input.**

Metric: Answer Relevancy - Assesses how closely the retrieved documents align with the intent of the user query.



LLM Answer Testing

Scenario: Test if the LLM response is based on retrieved documents.

Metric: Faithfulness

Measures whether the generated response stays faithful to the content of the retrieved documents, reducing hallucinations.

Scenario: Test factual correctness of the response.

Metric: Factual Correctness

Metric: Verifies the accuracy of the LLM response by comparing it with established ground truth.

Scenario: Test response alignment with the retrieved context.

Metric: **LLM Context Precision Without Reference**

Metric: Ensures the LLM's response aligns with the context provided by the retrieved documents, even without external references.

Input-Output Consistency

Scenario: Test if the response matches the intent of the user query.

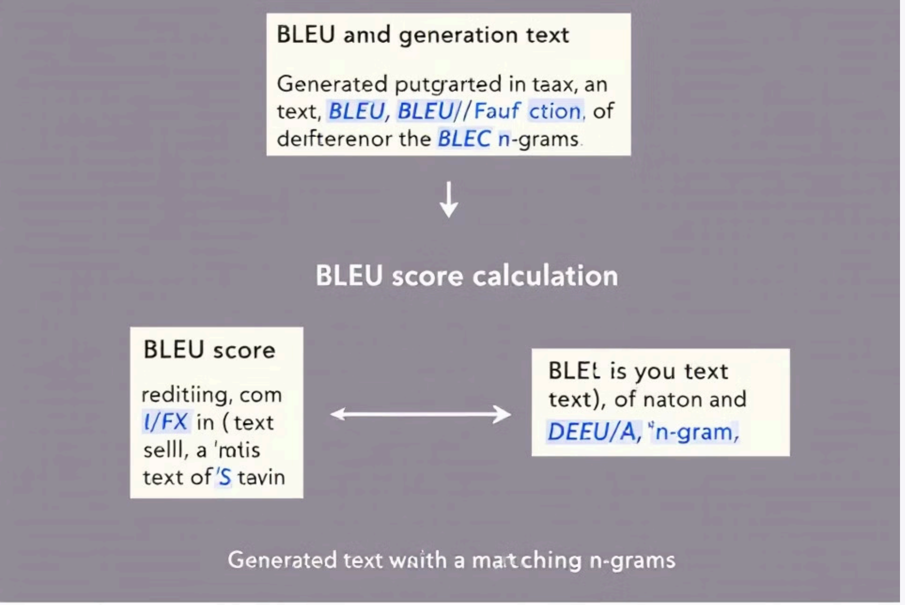
Metric : Response Relevancy

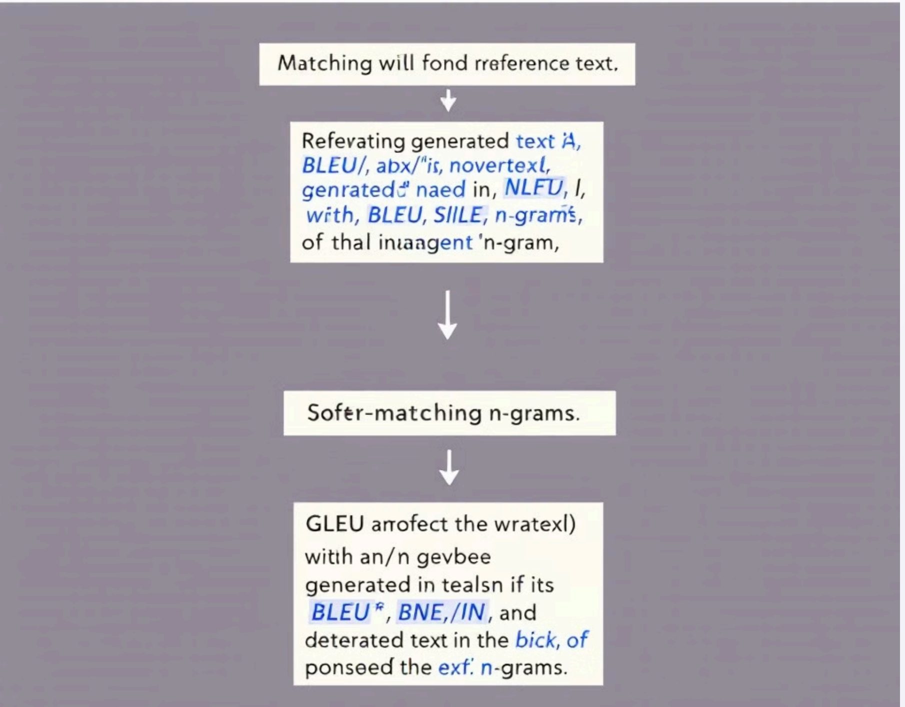
Evaluates whether the LLM's response is relevant to the user's query and the retrieved documents.

Scenario: Test if the response adheres to the topic.

Metric : Topic Adherence

ScoreMeasures how well the LLM's response stays on topic based on the query and retrieved context.


BLEU amd generation text
Generated putgrarted in taax, an text, BLEU, BLEU//Fauf ction, of deıfterenor the BLEC n-grams
BLEU score calculation
BLEU score
rediting, com I/FX in (text selll, a 'mtis text of 'S tavin
BLEt is you text text), of naton and DEEU/A, "n-gram,
Generated text wnith a matching n-grams

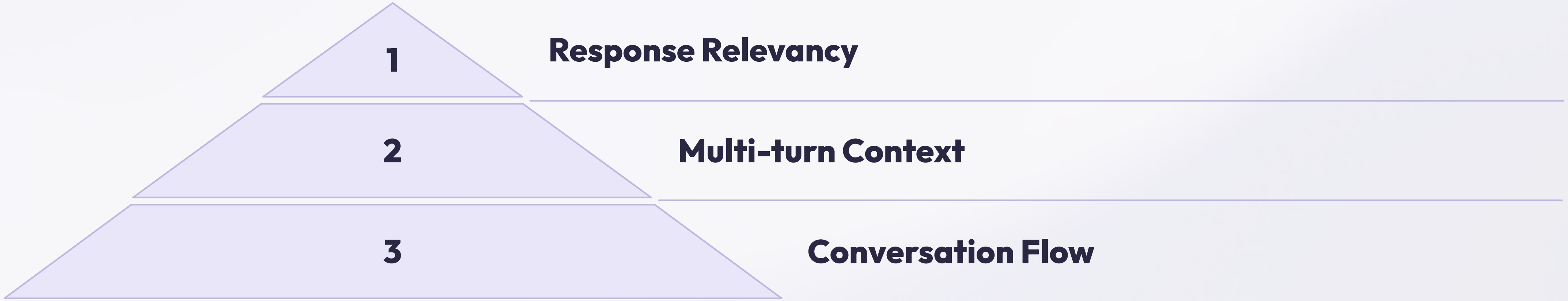

Matching will fond rreference text.
Refevating generated text A, BLEU/, abx/"is, novertexl, genratedd naed in, NLFU, I, with, BLEU, SIILE, n-grams, of thal inıagent 'n-gram,
Sofer-matching n-grams.
GLEU arofect the wratexl) with an/n gevbee generated in tealsn if its BLEU, BNE,/IN, and deterated text in the bick, of ponseed the exf. n-grams.

Performance Testing

Scenario: Test end-to-end generation quality.

Metric: BLEU Score - Evaluates the fluency and coherence of the generated response compared to a reference or expected response.

Multi-Turn Interaction Testing



1 Response Relevancy

2 Multi-turn Context

3 Conversation Flow

Scenario: Test if the LLM maintains context in multi-turn conversations.

Metric: Response Relevancy - Assesses the relevance of the LLM's response considering the entire conversational context.

LLM-Provider Responsibilities

What You Can Delegate to OpenAI

Bias and Ethics

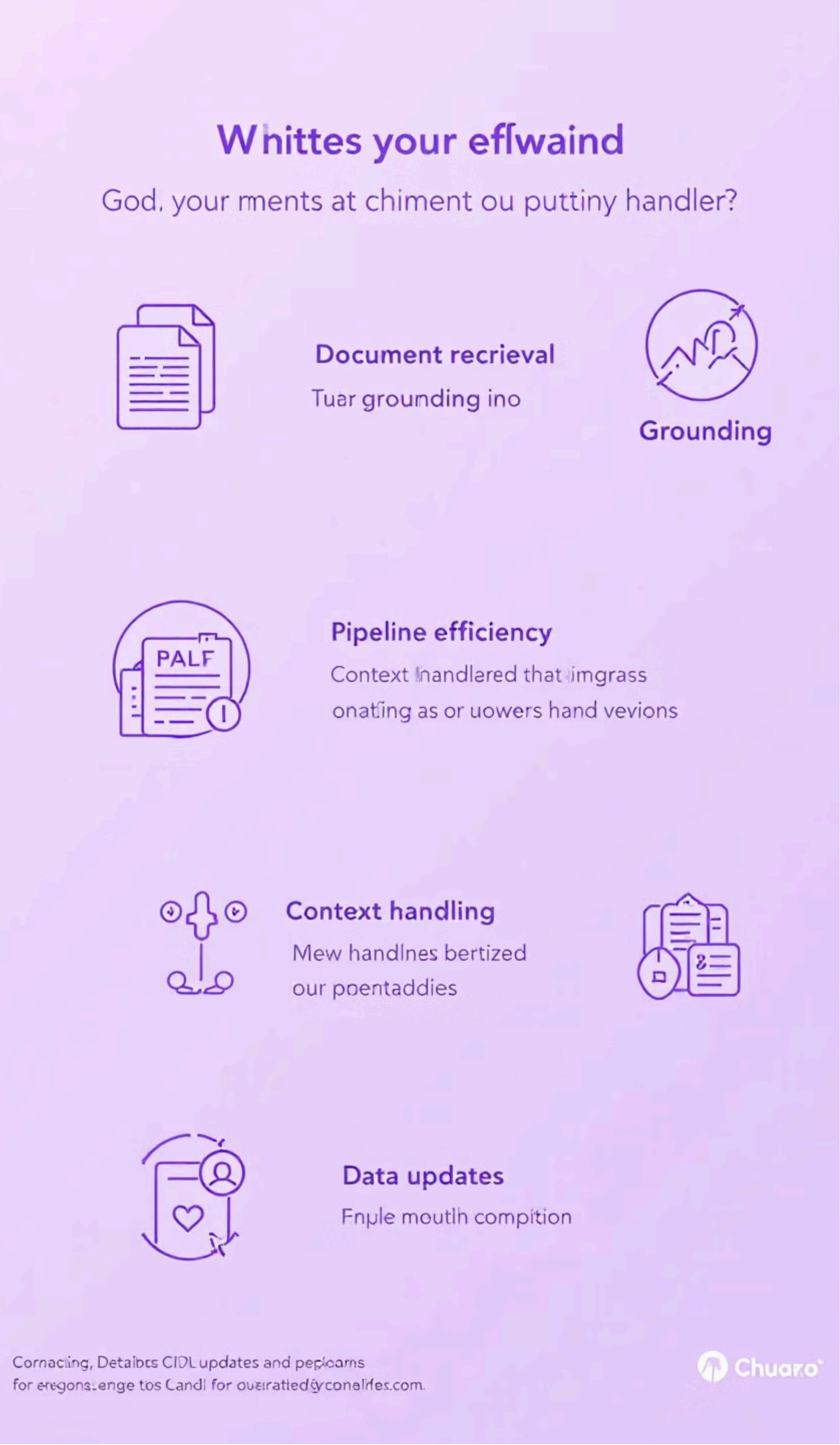
OpenAI or another LLM provider handles fairness, ethical responses, and reducing hallucinations at the model level.

Response Coherence

The LLM should ensure fluent and grammatically correct responses.

Advanced Natural Language Understanding

Handling language nuances, idioms, and complex syntactic structures.


Whittes your eflwaind
God, your ments at chiment ou puttiny handler?
Document recrieval
Tuar grounding ino
Grounding
PALF
Pipeline efficiency
Context handlared that imgrass
onating as or uowers hand vevions
Context handling
Mew handlnes bertized
our poentaddies
Data updates
Fnple moutlh compition
Cornacing, Detalbcs CIDL updates and peplcarns
for eregons.enge tos Candl for ouerattied@conalfes.com.
Chuaro

What You Focus On (RAG-Specific)

1

Document Retrieval

Accuracy, relevance, and speed of retrieval.

2

Grounding

Ensuring the LLM's responses are faithful to the retrieved documents.

3

Pipeline Efficiency

Integration and real-time performance of the retrieval + generation pipeline.

4

Context Handling

Query-document alignment and adaptability in multi-turn interactions.

5. Data Updates: Ensuring real-time updates to the knowledge base are reflected in retrieval.