Evaluating Gen AI Responses

Evaluation Metrics for Regression problems

- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- R-squared (coefficient of determination)
- Adjusted R-squared

Evaluation Metrics for Classification problems

- Accuracy
- Precision
- Recall
- F1 score
- Specificity
- Log Loss

Model evaluation for classification : why isn't accuracy sufficient?

Let us take example of a classification model that classifies if a transaction is fraudulent or not
The training data is labeled as:
"yes" means the transaction is fraudulent (True positive)
"no" means transaction is not fraudulent (True negative)

Actual Values

Output	Count
Yes (Fraudulent)	30
No (Genuine)	970

Predicted Values

Output	Count
Yes (Fraudulent)	1
No (Genuine)	999

Accuracy

$$\frac{True\ Positives\ +\ True\ Negatives}{Total\ Predictions}$$

$$\frac{971(?)}{1000} = 97.1\ \%$$

The model has a best-case **accuracy** of **97.1%**, even when it didn't detect 29 out of 30 fraudulent transactions

Hence other evaluation techniques such as confusion matrix are used
False negatives and False positives should be close to 0

Challenges in evaluating Gen AI solutions

Subjectivity
Multiple valid outputs make it hard to define a single correct answer.

Fluency vs. Factuality
Models may produce fluent but factually incorrect or hallucinated outputs.

Diversity and Creativity
Creative outputs like stories or images may be novel, making comparison with references difficult.

Human Preference and Use-Case Fit
Standard metrics often don't align with what humans prefer or what the use-case demands.

Bias and Toxicity Detection
Generated content may reinforce stereotypes or generate harmful text, which is hard to detect with metrics.

BLEU

BLEU stands for **Bilingual Evaluation Understudy**.

It is primarily used for evaluating machine translation systems.

The main idea behind BLEU is to compare the output of a machine translation model (the candidate translation) with one or more reference translations created by human translators.

BLEU measures the overlap of n-grams (a sequence of n items) between the candidate and reference translations.

The score ranges from 0 to 1, with 1 indicating a perfect match with the reference translation.

ROUGE

ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. It is commonly used for evaluating summarization tasks but can also apply to translation and other NLP tasks

ROUGE measures the quality of the generated text by comparing it to a reference summary.

ROUGE evaluates the overlap of n-grams, similar to BLEU, but **it focuses more on recall rather than precision**

Scenario – BLEU and ROUGE

Scenario:

A media company like **BBC News** or **The New York Times** wants to develop an **AI-powered summarization system** that condenses long news articles into short summaries for their mobile app.

BLEU

BLEU focuses on precision—how much of the generated text matches the reference exactly in n-grams.

ROUGE

ROUGE focuses on recall—how much of the key information from the original text appears in the summary.

Which one will be better suited for this scenario?

Scenario – BLEU and ROUGE

Scenario:

A company like **Google Translate** or **DeepL** is developing an **AI-powered translation system** to translate formal business documents **from English to French**. The goal is to ensure **grammatically correct and precise translations** that closely match human-translated reference texts.

BLEU

BLEU focuses on precision—how much of the generated text matches the reference exactly in n-grams.

ROUGE

ROUGE focuses on recall—how much of the key information from the original text appears in the summary.

Which one will be better suited for this scenario?

PERPLEXITY

How confidently does the model, typically a LLM, predicts on a sample of data

The confidence is measured by how far apart the probability distributions are
(close distribution means model is not confident)

Mathematically, it is the exponentiated average negative log-likelihood of a test dataset under the model.

Intuitive interpretation:

If a language model is very confident in the correct next token (giving it a high probability), the perplexity will be lower; if it spreads probability across many tokens (not sure which token is correct), the perplexity will be higher.

Range:

Limitations:
Cannot measure task performance
(Q&A accuracy, factual correctness, summarization quality, or dialogue coherence).

BERT Score

Reference Text:

“people like Western cuisine.,”

LLM response:

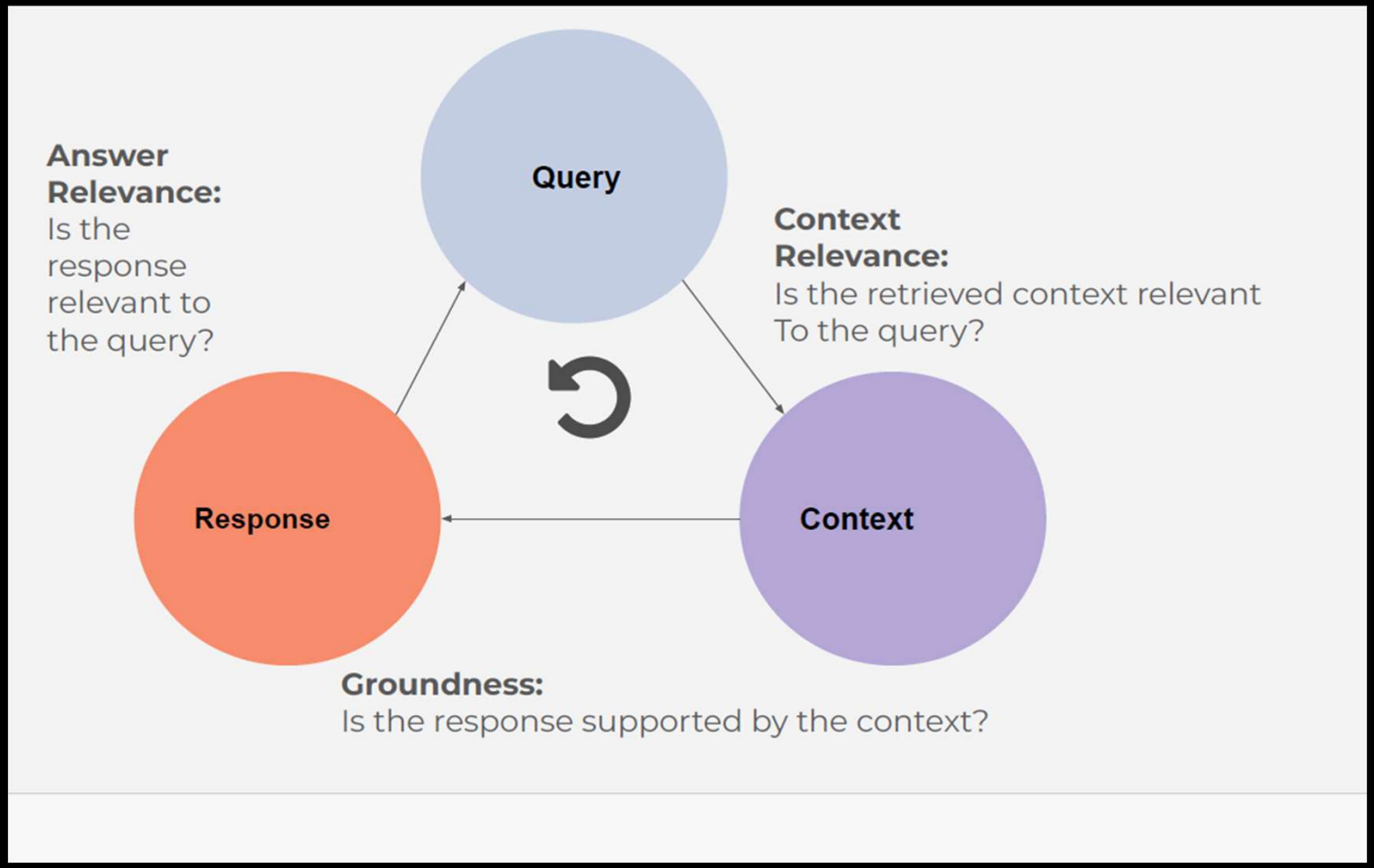
A: “consumers prefer imported dishes.”.

B: “people like global flavours”

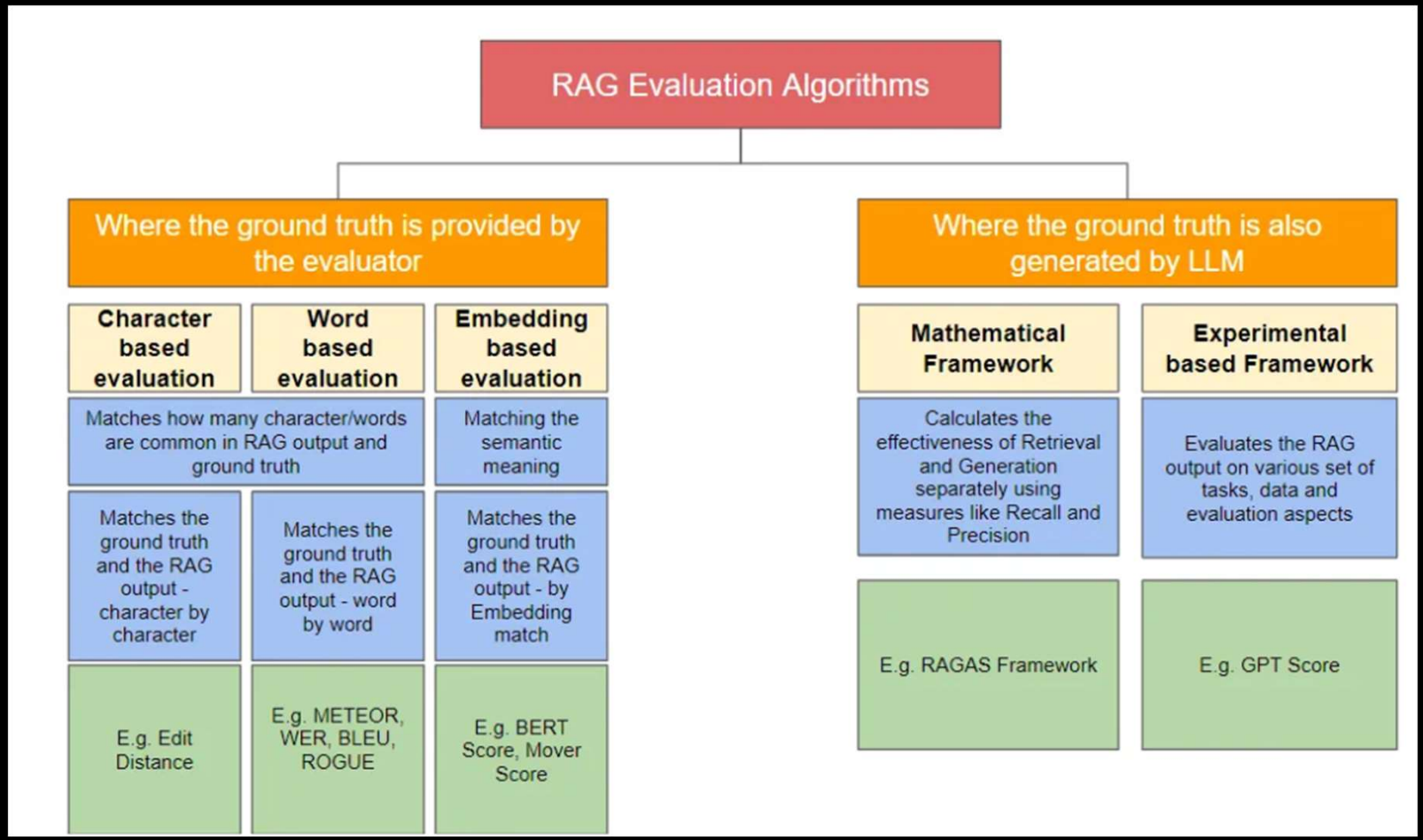
Which response will the n-gram based metrics rate higher?

In BERTScore, the similarity between two sentences is computed as the sum of the cosine similarities between their token embeddings, thereby providing the capability to detect paraphrases.

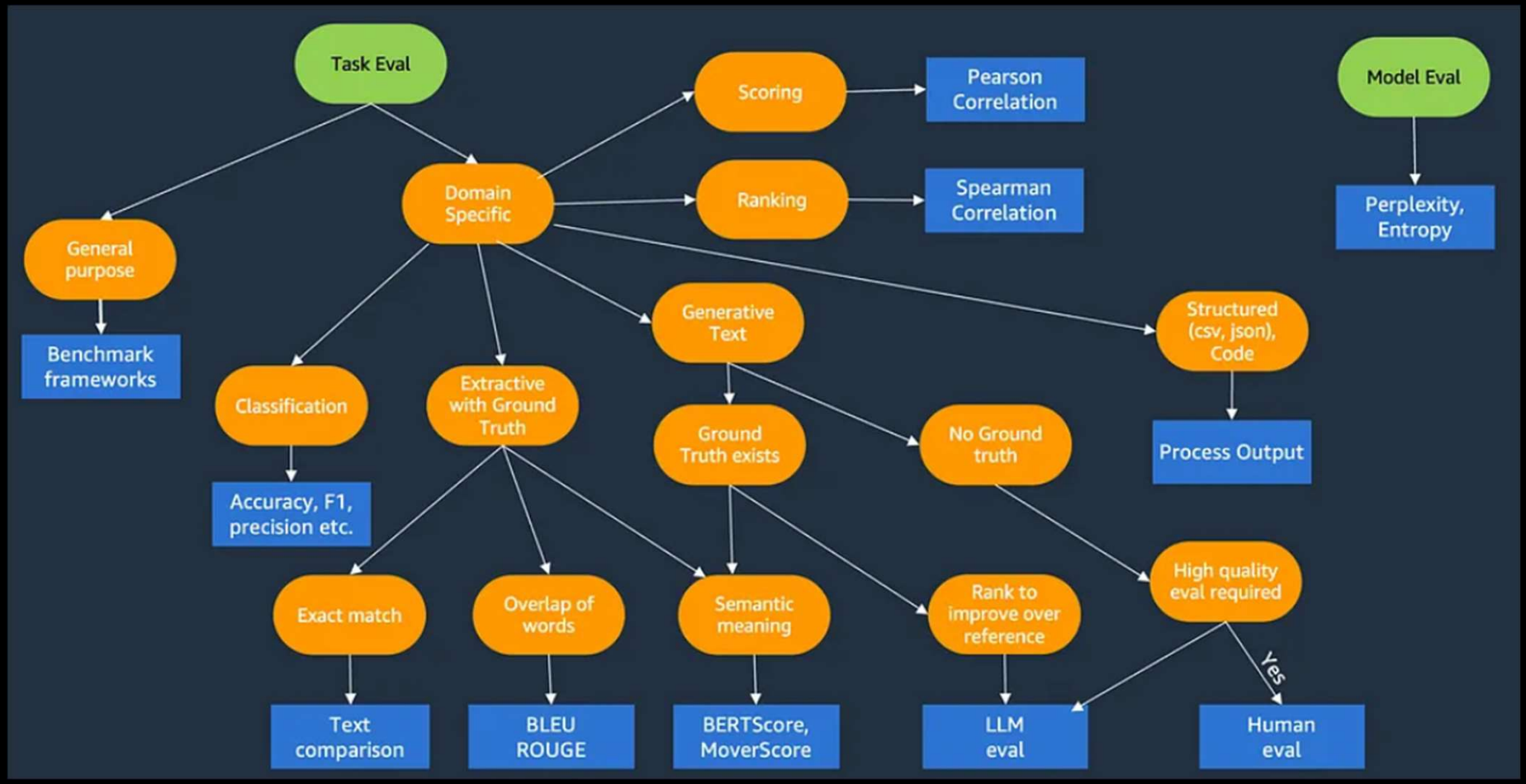
The RAG triad



Evaluate GenAI responses - approaches



Evaluating LLMs



The RAG triad - example

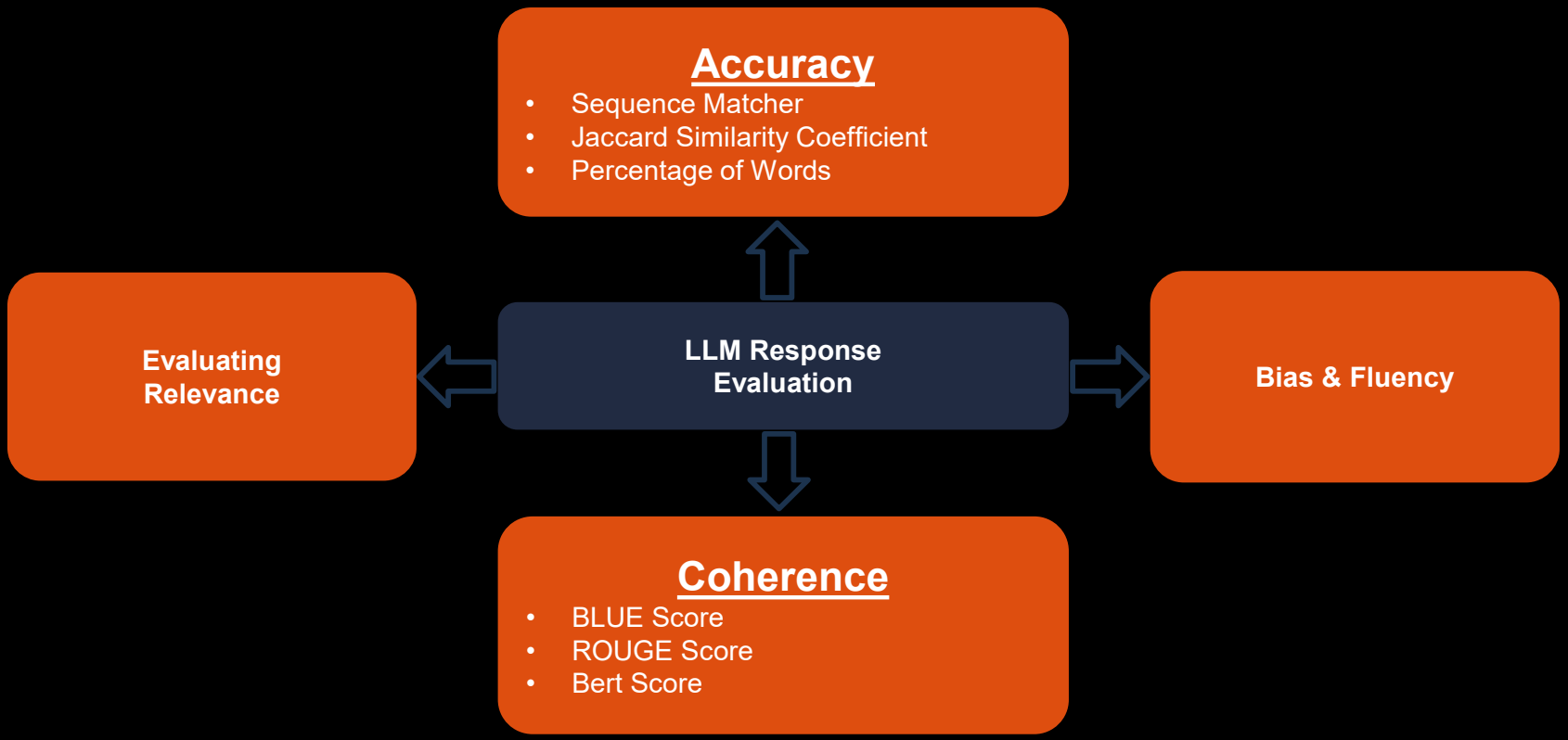
Scenario: M&A Analysis Request
A financial analyst asks a generative AI system: *“What are the strategic synergies of the proposed merger between Company A and Company B, based on recent financial reports and industry trends?”*

Groundedness
Is the answer grounded in factual retrievable sources?
If the system generates an answer based on outdated or incorrect documents, such as financial reports from several years ago or irrelevant M&A deals in different industries, this would indicate a failure in groundedness

Answer relevance
The answer highlights relevant synergies, like how Company A’s manufacturing efficiency complements Company B’s distribution network, leading to cost reductions and competitive advantage
if the answer mentions "synergies in marketing departments" when neither company has a strong marketing focus, the relevance fails.

Context relevance
Recent industry reports suggest demand for eco friendly products. Both companies have that capacity. By merging their R&D departments, they can leverage this trend better

GenAI Evaluation parameters and metrics



LLM Evaluation framework / metrics

Ragas Score

Generation

faithfulness

How factually accurate is the generated answer

answer relevancy

How relevant is the generated answer to the question

Retrieval

context precision

The signal to noise ratio of retrieved context

context recall

can it retrieve all relevant information required to answer the question

Evaluating LLMs

Metric/Method	Task / Use Case	Typical Range	Interpretation	Description (Recap)	Pros	Cons	Typical Evaluation Approach
Accuracy	Classification	0 to 1 (or 0% to 100%)	Closer to 1 is better (perfect=1)	Fraction of correct predictions over total predictions.	- Simple to understand and compute - Works well for balanced data	- Can be misleading for imbalanced classes - Does not differentiate which classes are misclassified	- Ratio of correct predictions to total predictions
Precision & Recall	Classification	0 to 1 (or 0% to 100%)	Closer to 1 is better (perfect=1)	Precision: among predicted positives, fraction that are truly positive Recall: among actual positives, fraction predicted correctly.	- More informative than Accuracy - Better for imbalanced data	- Must be viewed together or with F1 - Single metric alone can be insufficient	- Compare predicted vs. ground truth labels, focusing on "positive" class - Aggregate metrics (macro/micro) if multi-class

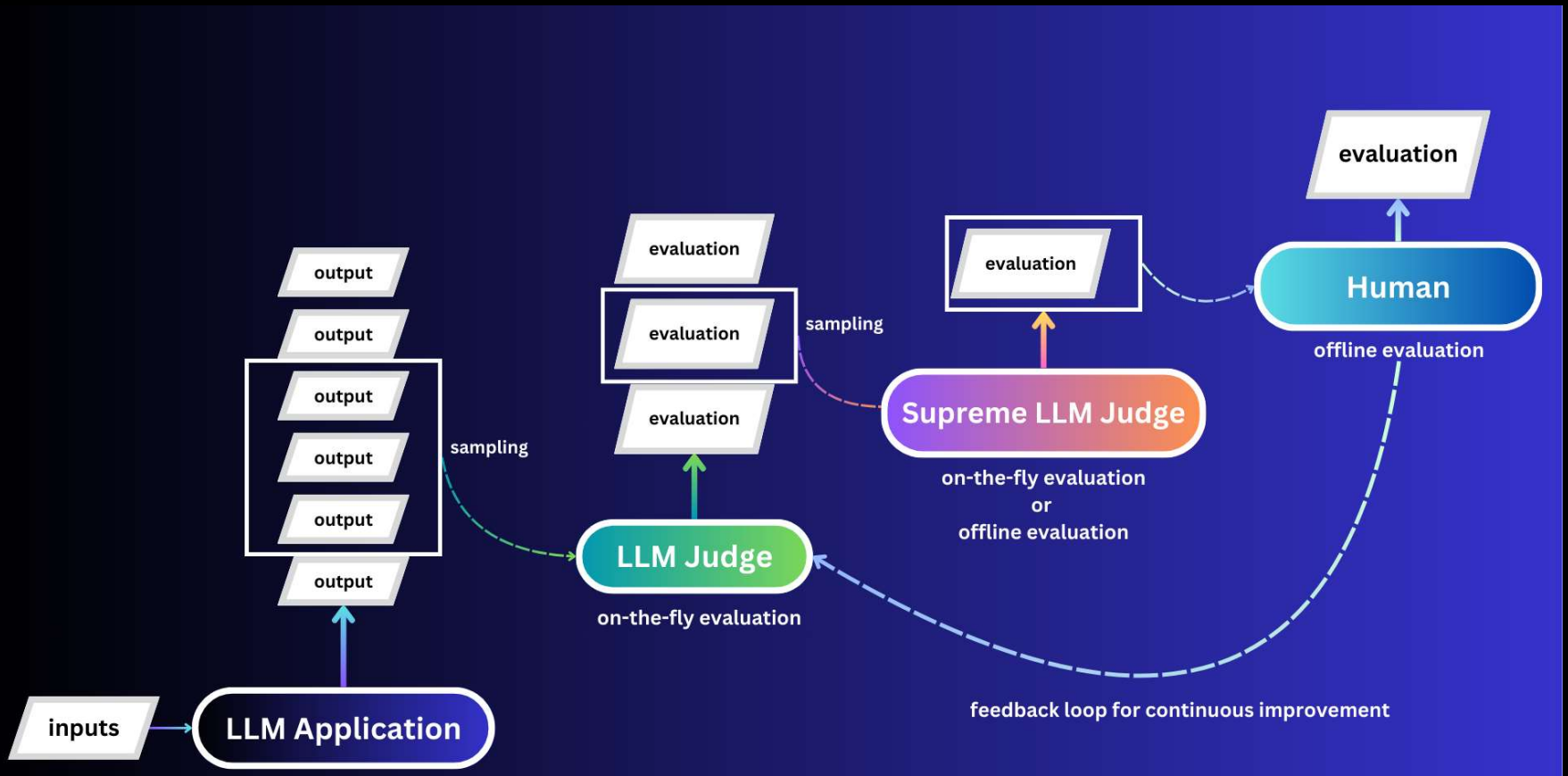
Evaluating LLMs

Metric/Method	Task / Use Case	Typical Range	Interpretation	Description (Recap)	Pros	Cons	Typical Evaluation Approach
ROUGE (e.g., ROUGE-1, ROUGE-L)	Summarization, Generation, QA	0 to 1 (or 0% to 100%)	Closer to 1 is better (perfect=1)	Overlap-based metric that counts matching n-grams or longest common subsequence between system output and reference.	- Standard in summarization - Straightforward, fast to compute	- Focuses on surface-form overlap - Less effective at measuring semantic equivalence	- Compare generated text to reference text(s) - Calculate n-gram or LCS overlap
BLEU	Machine Translation, General Generation	0 to 1 (often reported as 0-100)	Closer to 1 (or 100) is better (perfect=1)	Counts overlapping n-grams (up to 4-grams) between candidate and reference, with a brevity penalty.	- Widely used in MT - Quick to compute	- Surface-based, does not capture synonyms/paraphrase well - Single reference can cause underestimation	- Compare generated text to one or more reference translations - Compute n-gram matches and brevity penalty

Evaluating LLMs

Metric/Method	Task / Use Case	Typical Range	Interpretation	Description (Recap)	Pros	Cons	Typical Evaluation Approach
BERTScore	Summarization, QA Generation, Chatbots	0 to 1 (often 0.6–0.9 range)	Closer to 1 is better (perfect=1)	Uses contextual embeddings (e.g., BERT) to compute similarity between each token in candidate vs. reference.	- Better at capturing semantic similarity than n-gram overlap - More robust to paraphrasing	- Depends on pretrained language model choice - Slower than n-gram metrics for large datasets	- Embed candidate and reference texts with BERT (or similar) - Compute pairwise token similarities and find best alignment - Average or F1-like aggregate of token similarities
Perplexity	Language Model Evaluation	1 to ∞	Lower is better (perfect=1)	Exponentiated average negative log-likelihood of test set tokens under the model.	- Standard measure of LM fluency - Easy to compute for models that produce probabilities	- Not always aligned with human notions of "quality" - Doesn't address factual correctness or style	- Evaluate probability assigned to each token in a test set - Compute average negative log probability, then exponentiate

LLM as a judge



https://towardsdatascience.com/judge-an-llm-judge-a-dual-layer-evaluation-framework-for-continous-improvement-of-llm-apps-7450d0e81e17/

Open AI **evals** framework

OpenAI Evals is an open-source evaluation framework designed to test and benchmark LLM-based applications.

Instead of providing fixed metrics (like BLEU or ROUGE), it allows custom test cases and evaluation methodologies for different types of tasks.

It is especially useful for LLMs that generate open-ended text, where traditional token-based metrics (BLEU, ROUGE) may not capture true quality.

https://medium.com/@rudresh.narwal/openai-evals-dea94f7f2012

https://platform.openai.com/docs/guides/evals

Langchain **Evaluate** framework

A built-in evaluation module inside LangChain that allows you to assess chains, prompts, retrieval systems, and full LLM workflows.

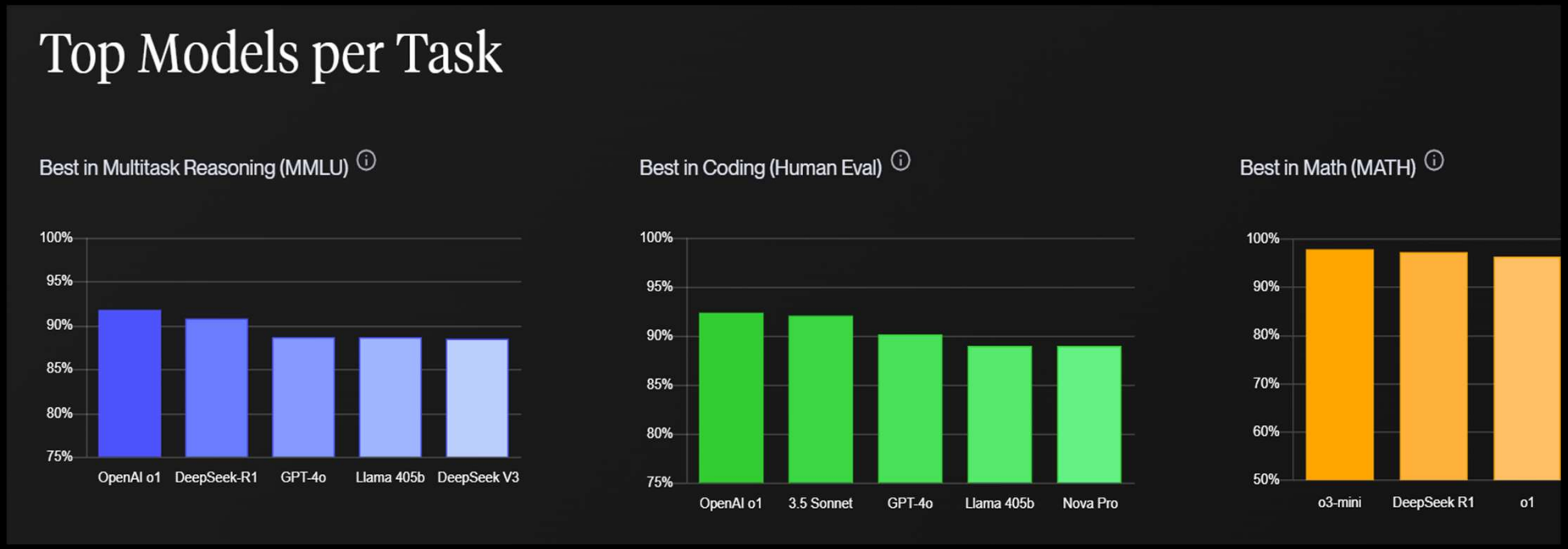
Supports various evaluation techniques, including:
- LLM-based scoring (self-critique).
- String-based comparisons (BLEU, ROUGE, BERTScore).
- Heuristic evaluation (e.g., checking if a response contains specific keywords).

Capabilities:
- Works with LangChain agents and retrieval systems (RAG).
- Easily integrates with existing LangChain applications.
- Custom evaluators can be defined using Python.

Role of Human reviewers in assessing response quality

- Human reviewers play a crucial role in evaluating how well the responses from LLM align with the intended goals and expectations
- It makes sure that LLM is not only informative but also produces human-like text
- Researchers assess LLM for empathetic responses and effective addressing of user concerns

LLM Leaderboard



MMLU: Massive Multitask Language Understanding

https://www.vellum.ai/llm-leaderboard

Implementation challenges

Accuracy
May occasionally miss critical details or emphasize less important information, affecting the accuracy of the output

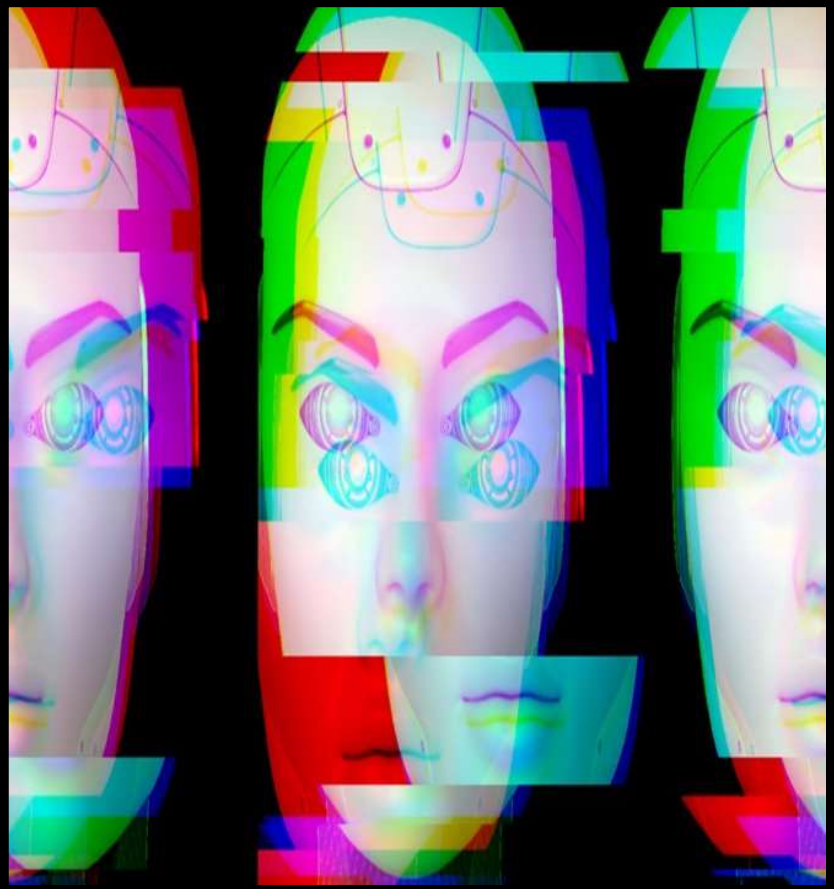
Bias and fairness
Inherits potential biases in training data

Lack of deep understanding and coherence
LLMs do not understand text like humans which can sometimes result in incoherent or logically flawed responses

Consistency and reliability
Can sometimes produce inconsistent or unreliable reasoning outputs, particularly if the prompt is ambiguous or if the task requires multi-step reasoning

Errors / Hallucinations
Incorrect information presented as factually correct (misleading)

Hallucination



Hallucinations occur when AI model generates incorrect or misleading information but presents it as if it were a fact.

Example: LLM designed to generate summaries of news articles may produce a summary that includes details not present in the original article, or even fabricates information entirely.

Reasons

Flawed training data – Inaccurate, Incomplete.

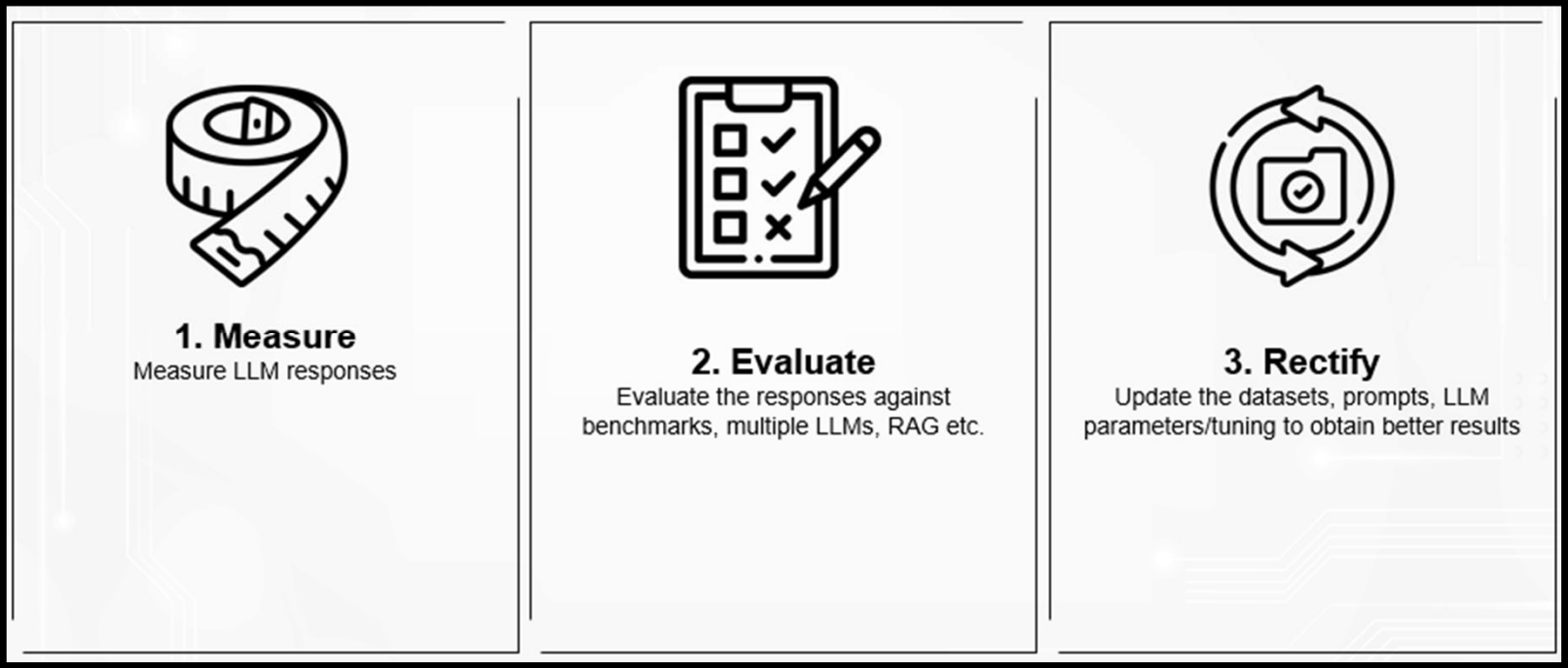
Overfitting to training data

Extrapolation beyond training.

Insufficient context or context switching

Ambiguous prompt

Mechanisms to mitigate hallucinations



Evaluate GenAI responses

- Ensures that the model consistently produces high-quality and accurate responses
- It adheres to ethical standards in various domains like healthcare
- Researchers can pinpoint areas for improvement
- Assess its ability to understand and respond to wide range of queries
- Regular evaluation builds trust
- Error Identification and Correction
- Alignment with User Expectations