RAG evaluation and deployment

David Lee | TIM 175

Today

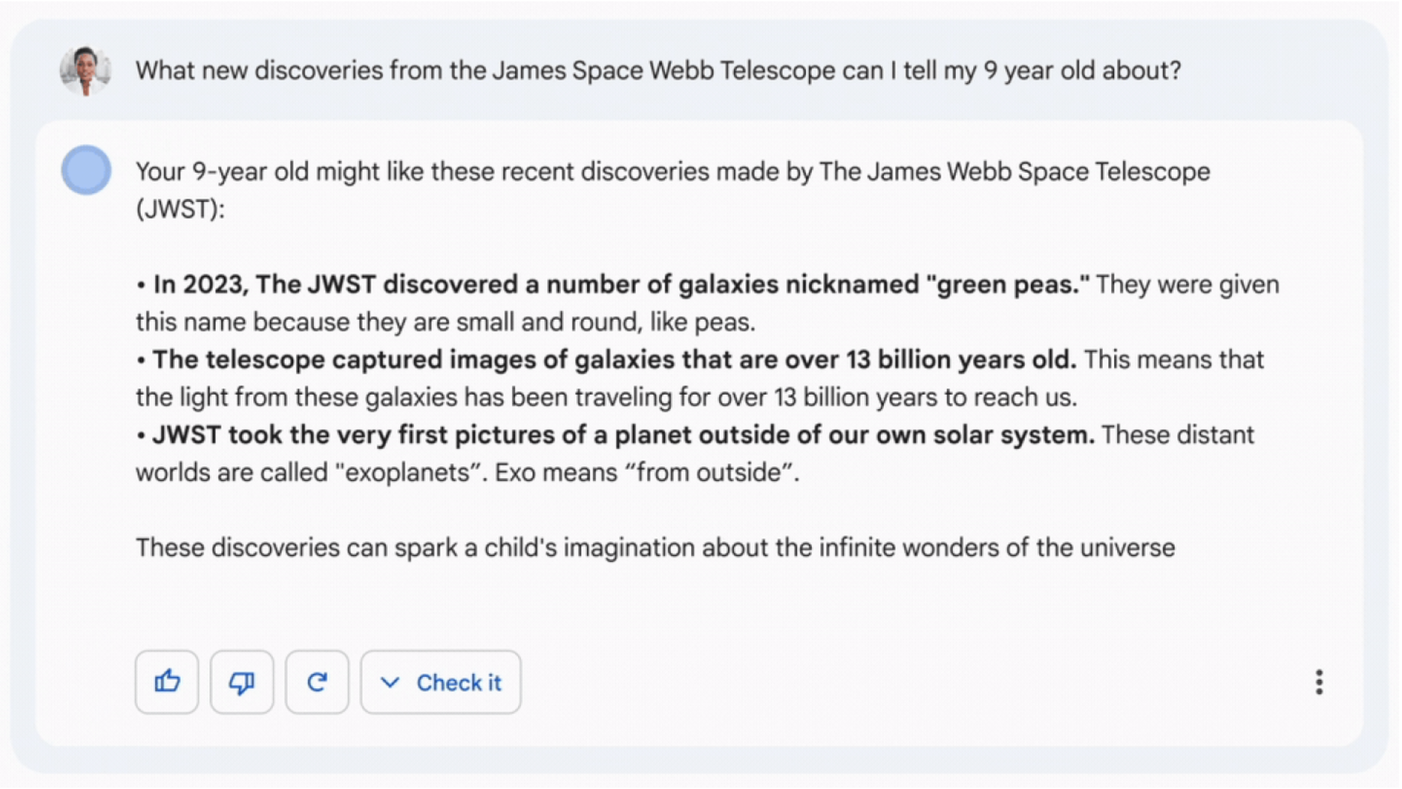
- RAG evaluation and deployment
- HW 8 Prelab
- Upcoming deliverables
  - Homework 8 (prelab) due TODAY at 11:59pm
  - Homework 8 (individual) due SATURDAY at 11:59pm
  - Homework 8 (team) due NEXT MONDAY at 11:59pm

Reminder of evaluation

Google shares lose $100 billion after company's AI chatbot makes an error during demo

By Catherine Thorbecke, CNN

3 minute read · Updated 9:41 AM EST, Thu February 9, 2023



Grant Tremblay
@astrogrant

Not to be a ~well, actually~ jerk, and I'm sure Bard will be impressive, but for the record: JWST did not take "the very first image of a planet outside our solar system".

the first image was instead done by Chauvin et al. (2004) with the VLT/NACO using adaptive optics.

Bruce Macintosh
@bmac_astro

Speaking as someone who imaged an exoplanet 14 years before JWST was launched, it feels like you should find a better example?

5:48 PM · Feb 7, 2023 · **70.2K** Views

Two US lawyers fined for submitting fake court citations from ChatGPT

Law firm also penalised after chatbot invented six legal cases that were then used in an aviation injury claim

TECH / MICROSOFT

Microsoft says listing the Ottawa Food Bank as a tourist destination wasn't the result of 'unsupervised AI'

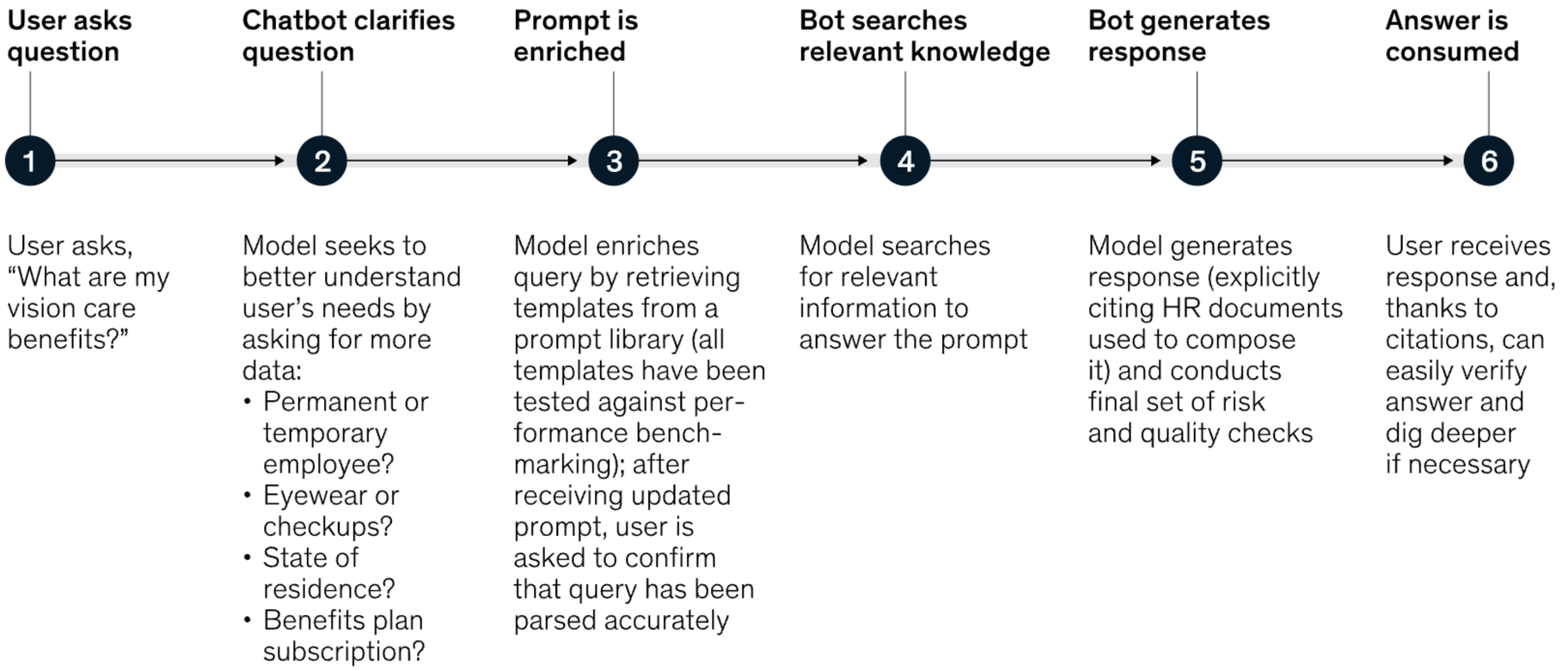
Mushroom pickers urged to avoid foraging books on Amazon that appear to be written by AI

Sample of books scored 100% on AI detection test as experts warn they contain dangerous advice

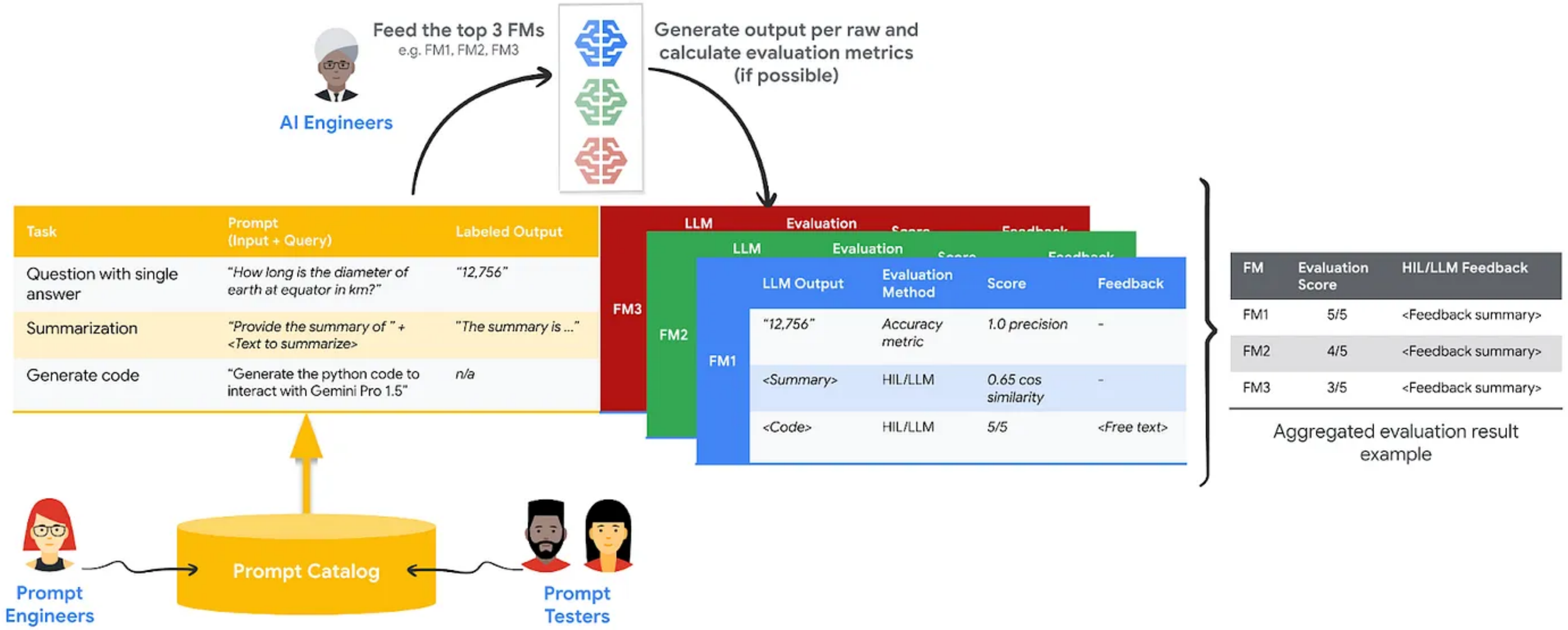
Exhibit 4

Generative AI risk can be mitigated at multiple points across a user interaction.

Sample HR chatbot interaction with built-in checkpoints to catch potential misfires



McKinsey & Company



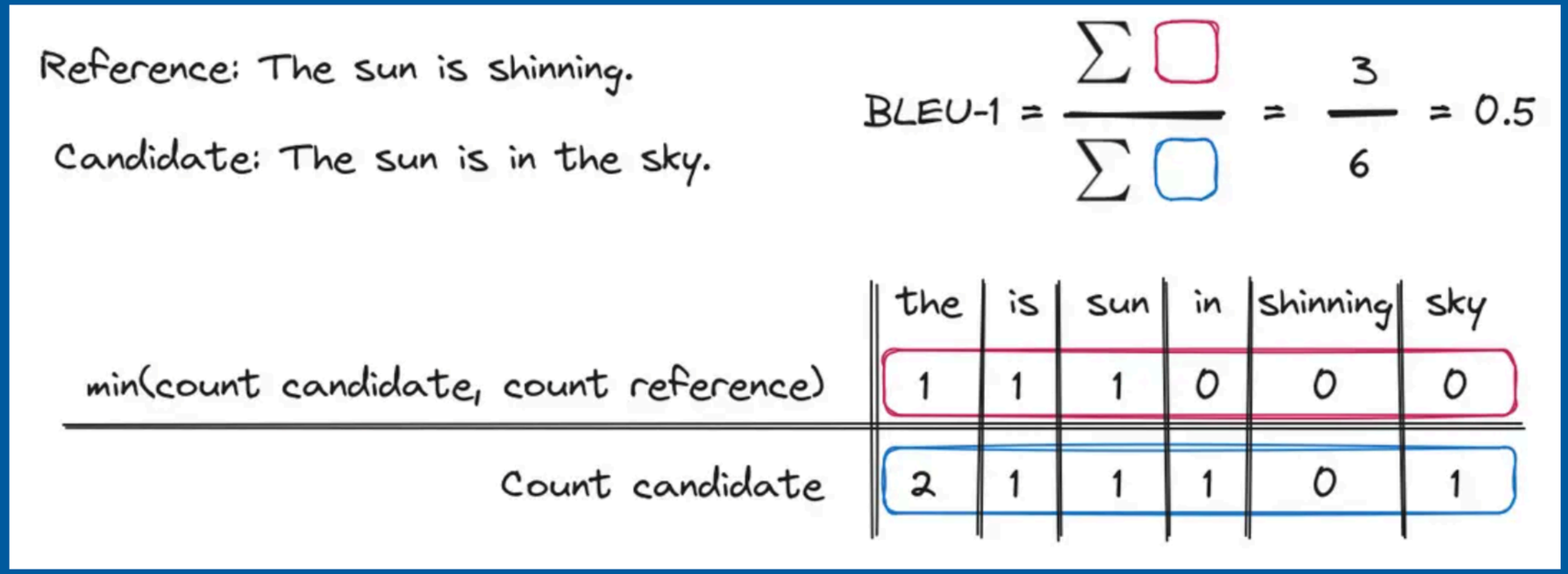
GenAIOps: Operationalize Generative AI
https://medium.com/google-cloud/genaiops-operationalize-generative-ai-a-practical-guide-d5bedaa59d78

Recall previous lecture:

How to evaluate?

Traditional NLP Metrics

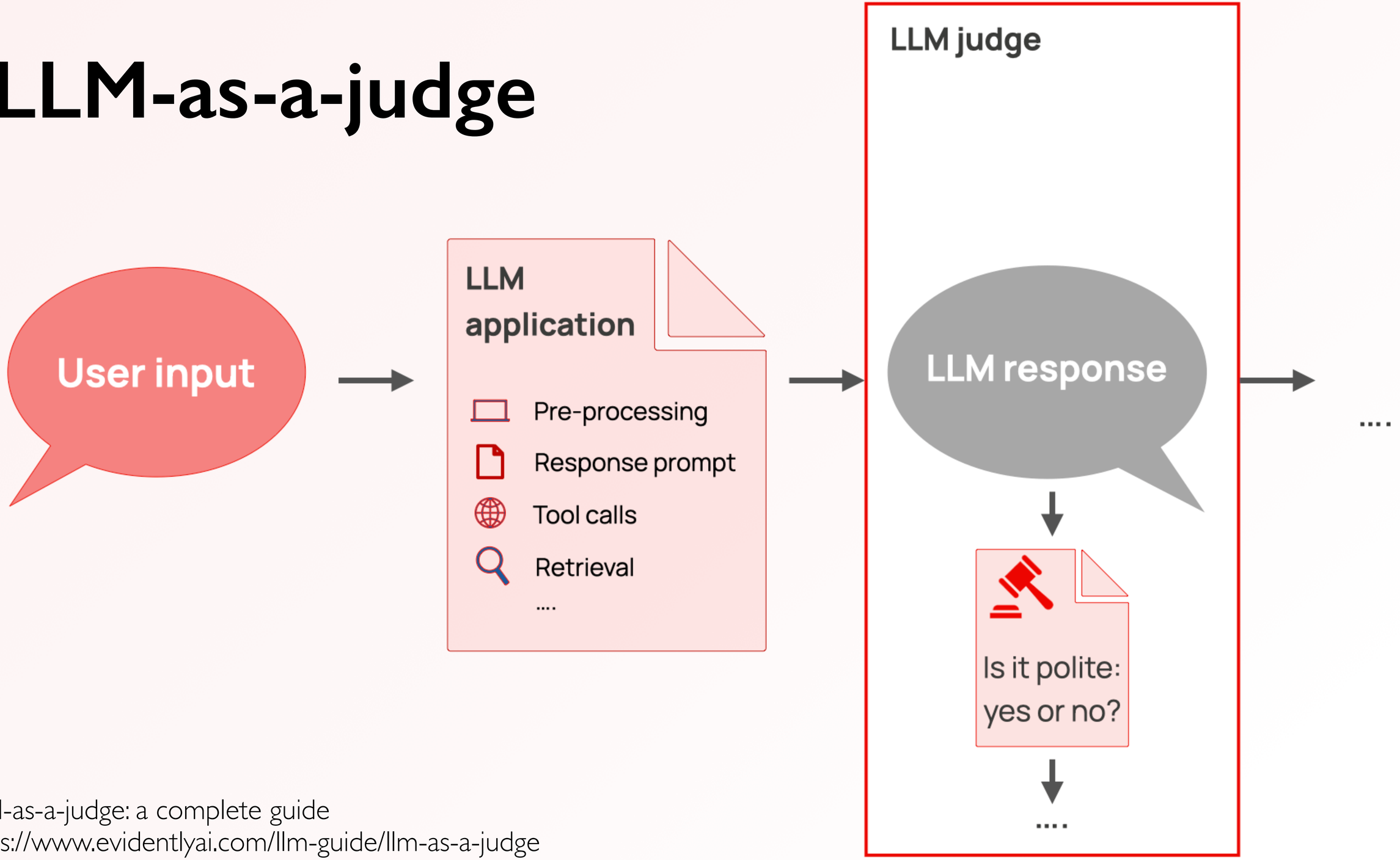
- BLEU - Bilingual Evaluation Understudy



- ROGUE - Recall-Oriented Understudy for Gisting Evaluation

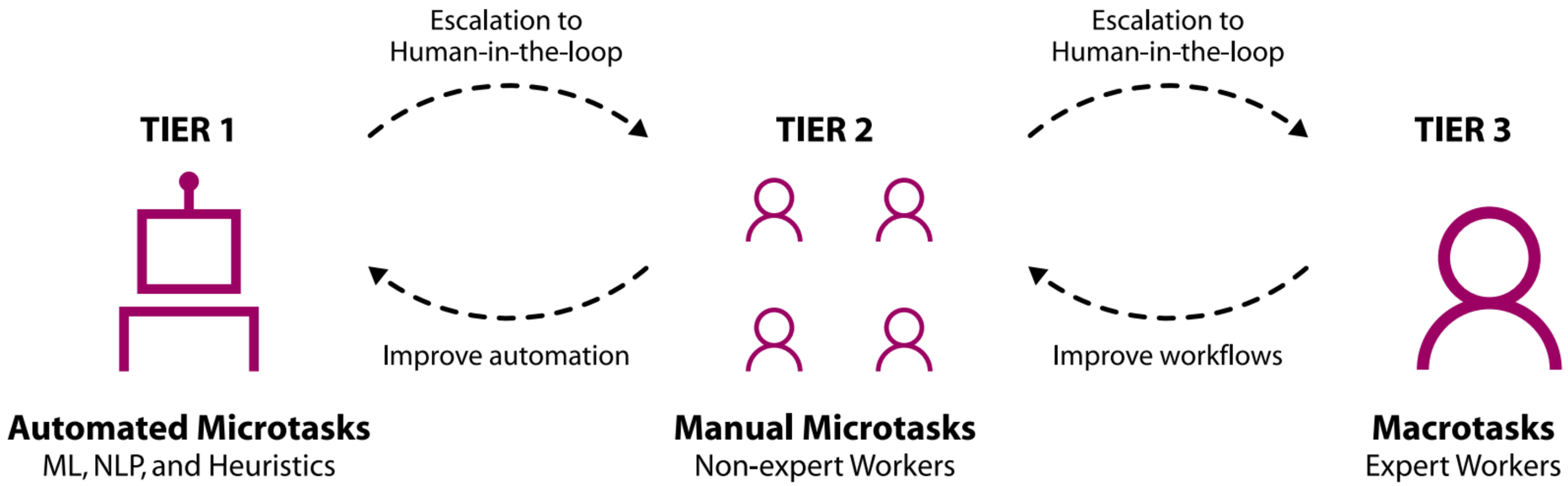
https://www.elastic.co/search-labs/blog/evaluating-rag-metrics

LLM-as-a-judge

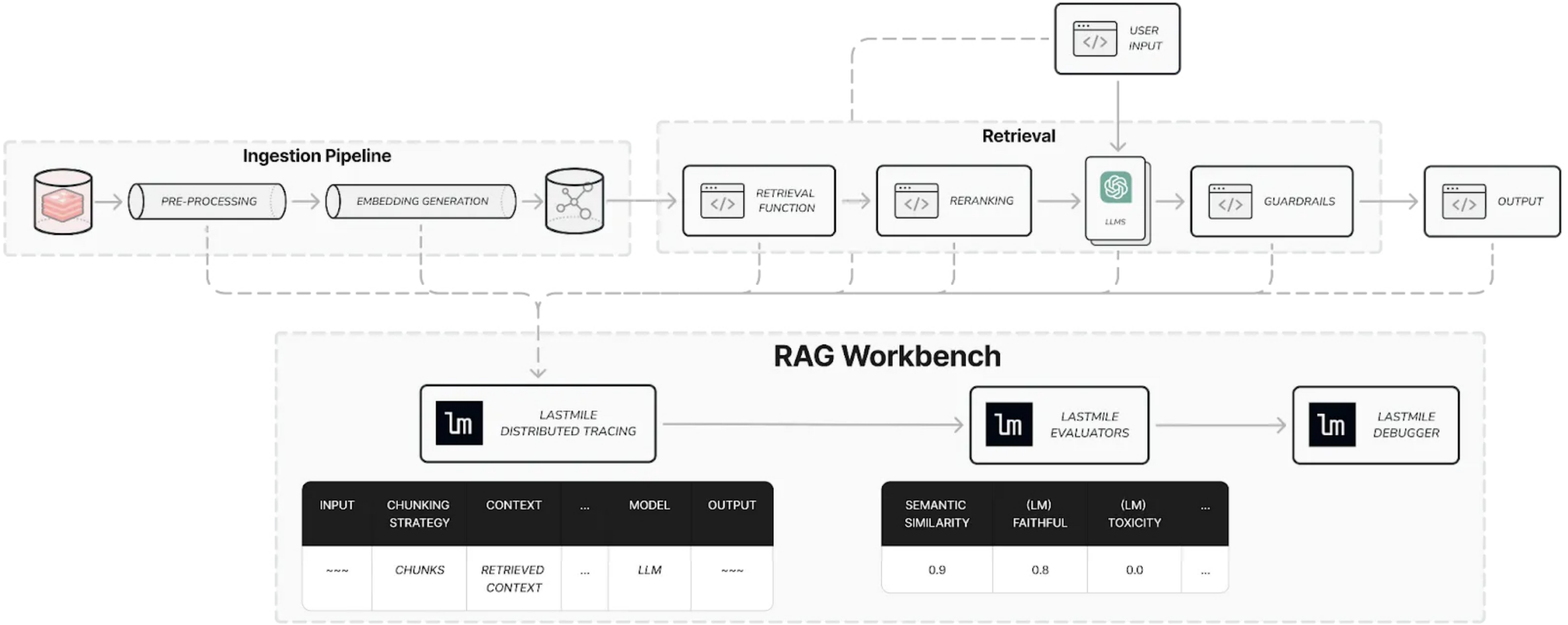


LLM-as-a-judge: a complete guide
https://www.evidentlyai.com/llm-guide/llm-as-a-judge

Human evaluation and annotation



RAG deployment and evaluation


USER INPUT
Retrieval
Ingestion Pipeline
PRE-PROCESSING
EMBEDDING GENERATION
RETRIEVAL FUNCTION
RERANKING
LLMS
GUARDRAILS
OUTPUT
RAG Workbench
LASTMILE DISTRIBUTED TRACING
LASTMILE EVALUATORS
LASTMILE DEBUGGER
INPUT
CHUNKING STRATEGY
CONTEXT
...
MODEL
OUTPUT
~~~
CHUNKS
RETRIEVED CONTEXT
...
LLM
~~~
SEMANTIC SIMILARITY
(LM) FAITHFUL
(LM) TOXICITY
...
0.9
0.8
0.0
...

glad-piano-568

7/12/2024, 9:07:42 PM

Mean Faithfulness	Mean Similarity
0.49	0.7

Search

Create Test Set

INPUT	OUTPUT	GROUND TRUTH	CATEGORY	FAITHFULNESS	SIMILARITY	TRACE ID
What is Einstein famous for in physics?	Einstein is famous for the theory of relativity.	Albert Einstein is famous for the theory of relativity.	Add a Category	0.9832520484924316	0.9	
What instrument did Einstein play?	Einstein played the piano.	Einstein played the violin.	Add a Category	0.0007336565176956...	0.5	

Rows per page 10 1-2 of 2

p-faithful (Hallucination Detection)

p-faithful is a specialized LastMile evaluator that measures the faithfulness of an LLM-generated response to the provided context in a RAG system. It aims to detect hallucination, which occurs when LLMs generate unsupported information despite being given context in the prompt. The p-faithful score answers: *"To what extent is the generated answer faithful to the provided context without introducing unsupported information?"*

Inputs:

- User's question: The input query to the RAG application.
- Retrieved context: The context or information retrieved from a vector DB based on the user question
- LLM-generated response: The output produced by the LLM using the retrieved context.

The p-faithful evaluator outputs a score ranging from 0 to 1. Higher scores indicating that the response is more faithful to the context. Lower scores suggest that the model might have included information not present in the context. Learn more about p-faithful.

	LLM Generated Response	Context Retrieved	User Query	Faithfulness
0	Albert Einstein is famous for the formula E = ...	\n Albert Einstein was a German-born theore...	What is Albert Einstein famous for?	0.66

The p-faithful score of 0.66 suggests that the LLM-generated response is only partially faithful to the retrieved context. While it correctly mentions Einstein's famous equation, $E = mc^2$, it also includes information about Brownian motion, which is not present in the context. The inclusion of this additional information likely contributed to the lower p-faithful score, indicating that the LLM relied on its own knowledge beyond the given context.

```
user_query = [
    "What is Albert Einstein famous for?",
]

context_retrieved = [
    """Albert Einstein was a German-born theoretical physicist who developed
    the theory of relativity, one of the two pillars of modern physics. His
    work is also known for its influence on the philosophy of science. He is
    best known to the general public for his mass-energy equivalence formula
    E = mc², which has been dubbed 'the world's most famous equation'. He
    received the 1921 Nobel Prize in Physics 'for his services to theoretical
    physics, and especially for his discovery of the law of the photoelectric
    effect', a pivotal step in the development of quantum theory."""
]

llm_generated_response = [
    "Albert Einstein is famous for the formula E = mc² and Brownian motion.",
]

# Calculate p-faithful
p_faithful_result = get_rag_eval_scores(
    user_query,
    context_retrieved,
    llm_generated_response,
    api_token=os.getenv("LASTMILE_API_TOKEN"),
)

# Print results
display(pd.DataFrame({
    "LLM Generated Response": llm_generated_response,
    "Context Retrieved": context_retrieved,
    "User Query": user_query,
    "Faithfulness": p_faithful_result['p_faithful']
}))
```

Relevance

The Relevance evaluator measures how well an LLM-generated response aligns with an expected output (ground truth). It aims to assess the topical relevance and contextual alignment of the generated response. The relevance score answers: *"To what extent does the generated response align with the topic and content of the expected output?"*

Inputs:

- LLM-generated response: The response produced by the LLM.
- Expected output (ground truth): The ideal or correct response to compare against.

The Relevance Score evaluator outputs a score ranging from 0 to 1. Higher scores indicate that the LLM-generated response is more relevant and aligned with the expected output. Lower scores suggest that the response may be off-topic or misaligned with the desired content.

	LLM Generated Response	Expected Output	Relevance Score
0	Albert Einstein revolutionized physics with hi...	Albert Einstein transformed our understanding ...	0.94

The relevance score of 0.94 indicates high alignment between the LLM-generated response and the expected output. Both texts cover Einstein's key contributions: the theory of relativity, E = mc², and the photoelectric effect. The score isn't perfect, as the LLM response omits mentions of general relativity and Einstein's broader impact on physics. However, it adds relevant details about light's constant speed. Overall, the high score reflects strong topical relevance and accurate conveyance of Einstein's main scientific achievements.

```
llm_generated_response = [
    """Albert Einstein revolutionized physics with his theory of relativity.
    He proposed that space and time are interconnected and that the speed of
    light is constant in all reference frames. His famous equation E = mc²
    showed that mass and energy are equivalent. Einstein's work on the
    photoelectric effect contributed to the development of quantum theory,
    earning him the Nobel Prize in Physics."""
]

expected_output = [
    """Albert Einstein transformed our understanding of the universe with his
    groundbreaking theories. His special and general theories of relativity
    redefined concepts of space, time, and gravity. Einstein's equation E = mc²
    revealed the fundamental relationship between mass and energy. His
    explanation of the photoelectric effect was crucial to the emergence of
    quantum physics, for which he received the Nobel Prize. Throughout his career,
    Einstein's innovative thinking and scientific contributions reshaped the
    field of physics."""
]

# Calculate Relevance Score
relevance_score = lm_eval_text.calculate_relevance_score(
    llm_generated_response,
    expected_output,
    model_name="gpt-3.5-turbo"
)

# Print results
display(pd.DataFrame({
    "LLM Generated Response": llm_generated_response,
    "Expected Output": expected_output,
    "Relevance Score": relevance_score
}))
```

Summarization

The Summarization evaluator measures the quality of an LLM-generated summary compared to the source document. It aims to assess how well the summary captures the essential information and main ideas of the source document. The summarization score answers: *"To what extent does the generated summary capture the essential information from the source document?"*

Inputs:

- Source document: The full text that is being summarized.
- LLM-generated summary: The summary produced by the LLM based on the source document.

The Summarization evaluator outputs a score ranging from 0 to 1. Higher scores indicate that the summary more effectively captures the essential information from the source document. Lower scores suggest that the summary may be missing key points or including irrelevant information.

	Source Document	LLM Generated Summary	Summarization Score
0	Albert Einstein was a German-born theoretical ...	Albert Einstein, a German-born physicist, deve...	0.88

The summarization score of 0.88 indicates the summary effectively captures key points about Einstein, including his German origin, development of relativity theory, E = mc² equation, 1921 Nobel Prize for the photoelectric effect, and work on unified field theory. It successfully condenses essential information while maintaining accuracy. The score isn't perfect, likely due to omitting Einstein's influence on philosophy of science and his later isolation from mainstream physics.

```
source_document = [
    """Albert Einstein was a German-born theoretical physicist who developed
    the theory of relativity, one of the two pillars of modern physics. His work
    is also known for its influence on the philosophy of science. He is best known
    to the general public for his mass-energy equivalence formula E = mc², 
    which has been dubbed 'the world's most famous equation'. Einstein received
    the 1921 Nobel Prize in Physics 'for his services to theoretical physics, and
    especially for his discovery of the law of the photoelectric effect', a
    pivotal step in the development of quantum theory. In his later years,
    Einstein focused on unified field theory and became increasingly isolated
    from the mainstream of   modern physics."""
]

llm_generated_summary = [
    """Albert Einstein, a German-born physicist, developed the theory of
    relativity and the famous equation E = mc². He won the 1921 Nobel Prize
    in Physics for his work on the photoelectric effect, contributing to
    quantum theory. Later, he worked on unified field theory."""
]

# Calculate Summarization Score
summarization_score = lm_eval_text.calculate_summarization_score(
    llm_generated_summary,
    source_document,
    model_name="gpt-3.5-turbo"
)

# Print results
display(pd.DataFrame({
    "Source Document": source_document,
    "LLM Generated Summary": llm_generated_summary,
    "Summarization Score": summarization_score
}))
```

Q/A on Retrieved Data

The Q/A on Retrieved Data evaluator is a binary measure of whether an LLM-generated response answers a user query based on the retrieved context. It aims to evaluate the accuracy and completeness of the answer in relation to the provided information. The Q/A on Retrieved Data score answers: *"Does the LLM-generated response correctly and completely answer the user query based on the retrieved context?"*

Inputs:

- User's question: The input query to the RAG application.
- Retrieved context: The context or information retrieved from a vector DB based on the user question
- LLM-generated response: The output produced by the LLM using the retrieved context.

The Q/A on Retrieved Data evaluator outputs a binary score. A score of 1 indicates the answer is correct and complete based on the retrieved context. A score of 0 means the answer is incorrect, incomplete, or includes information not present in the retrieved context.

	User Query	Retrieved Context	LLM Generated Response	Q/A on Retrieved Data Score
0	What did Einstein win the Nobel Prize for?	Albert Einstein received the Nobel Prize in Ph...	Einstein won the Nobel Prize in Physics in 192...	1.0

The Q/A on Retrieved Data score of 1 indicates the LLM response fully and accurately answers the query based on the retrieved context. It correctly identifies Einstein's Nobel Prize reason and year, notes it wasn't for relativity, and doesn't add information beyond the context, demonstrating high fidelity to the provided data.

```
user_query = [
    "What did Einstein win the Nobel Prize for?"
]

retrieved_context = [
    """Albert Einstein received the Nobel Prize in Physics in 1921. However,
    contrary to popular belief, he didn't receive it for his theories of relativity.
    Einstein was awarded the Nobel Prize 'for his services to Theoretical Physics,
    and especially for his discovery of the law of the photoelectric effect'."""
]

llm_generated_response = [
    """Einstein won the Nobel Prize in Physics in 1921 for his discovery
    of the law of the photoelectric effect, not for his theories of relativity as
     is often mistakenly believed."""
]

# Calculate Q/A on Retrieved Data Score
qa_score = lm_eval_text.calculate_qa_score(
    llm_generated_response,
    retrieved_context,
    user_query,
    model_name="gpt-3.5-turbo"
)

# Print results
display(pd.DataFrame({
    "User Query": user_query,
    "Retrieved Context": retrieved_context,
    "LLM Generated Response": llm_generated_response,
    "Q/A on Retrieved Data Score": qa_score
}))
```

Context Recall

Context Recall measures how many of the relevant documents (or pieces of information) were successfully retrieved. It focuses on not missing important results. Higher recall means fewer relevant documents were left out. In short, recall is about not missing anything important. Since it is about not missing anything, calculating context recall always requires a reference to compare against.

LLM Based Context Recall

`LLMContextRecall` is computed using `user_input`, `reference` and the `retrieved_contexts`, and the values range between 0 and 1, with higher values indicating better performance. This metric uses `reference` as a proxy to `reference_contexts` which also makes it easier to use as annotating reference contexts can be very time consuming. To estimate context recall from the `reference`, the reference is broken down into claims each claim in the `reference` answer is analyzed to determine whether it can be attributed to the retrieved context or not. In an ideal scenario, all claims in the reference answer should be attributable to the retrieved context.

The formula for calculating context recall is as follows:

$$\text{Context Recall} = \frac{\text{Number of claims in the reference supported by the retrieved context}}{\text{Total number of claims in the reference}}$$

Context Entities Recall

`ContextEntityRecall` metric gives the measure of recall of the retrieved context, based on the number of entities present in both `reference` and `retrieved_contexts` relative to the number of entities present in the `reference` alone. Simply put, it is a measure of what fraction of entities are recalled from `reference`. This metric is useful in fact-based use cases like tourism help desk, historical QA, etc. This metric can help evaluate the retrieval mechanism for entities, based on comparison with entities present in `reference`, because in cases where entities matter, we need the `retrieved_contexts` which cover them.

To compute this metric, we use two sets:

- RE: The set of entities in the reference.
- RCE: The set of entities in the retrieved contexts.

We calculate the number of entities common to both sets ($RCE \cap RE$) and divide it by the total number of entities in the reference (RE). The formula is:

$$\text{Context Entity Recall} = \frac{\text{Number of common entities between } RCE \text{ and } RE}{\text{Total number of entities in } RE}$$

Example

reference: The Taj Mahal is an ivory-white marble mausoleum on the right bank of the river Yamuna in the Indian city of Agra. It was commissioned in 1631 by the Mughal emperor Shah Jahan to house the tomb of his favorite wife, Mumtaz Mahal. **High entity recall context**: The Taj Mahal is a symbol of love and architectural marvel located in Agra, India. It was built by the Mughal emperor Shah Jahan in memory of his beloved wife, Mumtaz Mahal. The structure is renowned for its intricate marble work and beautiful gardens surrounding it. **Low entity recall context**: The Taj Mahal is an iconic monument in India. It is a UNESCO World Heritage Site and attracts millions of visitors annually. The intricate carvings and stunning architecture make it a must-visit destination.

Let us consider the refrence and the retrieved contexts given above.

- **Step-1**: Find entities present in the refrence.
  - Entities in ground truth (RE) - ['Taj Mahal', 'Yamuna', 'Agra', '1631', 'Shah Jahan', 'Mumtaz Mahal']
- **Step-2**: Find entities present in the retrieved contexts.
  - Entities in context (RCE1) - ['Taj Mahal', 'Agra', 'Shah Jahan', 'Mumtaz Mahal', 'India']
  - Entities in context (RCE2) - ['Taj Mahal', 'UNESCO', 'India']
- **Step-3**: Use the formula given above to calculate entity-recall

$$\text{context entity recall } 1 = \frac{|RCE1 \cap RE|}{|RE|} = 4/6 = 0.666$$

$$\text{context entity recall } 2 = \frac{|RCE2 \cap RE|}{|RE|} = 1/6$$

We can see that the first context had a high entity recall, because it has a better entity coverage given the refrence. If these two retrieved contexts were fetched by two retrieval mechanisms on same set of documents, we could say that the first mechanism was better than the other in use-cases where entities are of importance.

$$\text{Context Precision} = \frac{\#\ \text{relevant retrieved chunks}}{\#\ \text{total retrieved chunks}}$$

Context Precision

Context Precision is a metric that measures the proportion of relevant chunks in the `retrieved_contexts`. It is calculated as the mean of the precision@k for each chunk in the context. Precision@k is the ratio of the number of relevant chunks at rank k to the total number of chunks at rank k.

$$\text{Context Precision@K} = \frac{\sum_{k=1}^{K} (\text{Precision@k} \times v_k)}{\text{Total number of relevant items in the top } K \text{ results}}$$

$$\text{Precision@k} = \frac{\text{true positives@k}}{(\text{true positives@k} + \text{false positives@k})}$$

Where K is the total number of chunks in `retrieved_contexts` and $v_k \in \{0, 1\}$ is the relevance indicator at rank k.

Response Relevancy

The `ResponseRelevancy` metric measures how relevant a response is to the user input. Higher scores indicate better alignment with the user input, while lower scores are given if the response is incomplete or includes redundant information.

This metric is calculated using the `user_input` and the `response` as follows:

1. Generate a set of artificial questions (default is 3) based on the response. These questions are designed to reflect the content of the response.
2. Compute the cosine similarity between the embedding of the user input (E_o) and the embedding of each generated question (E_{g_i}).
3. Take the average of these cosine similarity scores to get the **Answer Relevancy**:

$$\text{Answer Relevancy} = \frac{1}{N} \sum_{i=1}^{N} \text{cosine similarity}(E_{g_i}, E_o)$$

$$\text{Answer Relevancy} = \frac{1}{N} \sum_{i=1}^{N} \frac{E_{g_i} \cdot E_o}{\|E_{g_i}\| \|E_o\|}$$

Where:
- E_{g_i}: Embedding of the i^{th} generated question.
- E_o: Embedding of the user input.
- N: Number of generated questions (default is 3).

How It's Calculated

Example

Question: Where is France and what is it's capital?

Low relevance answer: France is in western Europe.

High relevance answer: France is in western Europe and Paris is its capital.

To calculate the relevance of the answer to the given question, we follow two steps:

- **Step 1:** Reverse-engineer 'n' variants of the question from the generated answer using a Large Language Model (LLM). For instance, for the first answer, the LLM might generate the following possible questions:
  - *Question 1:* "In which part of Europe is France located?"
  - *Question 2:* "What is the geographical location of France within Europe?"
  - *Question 3:* "Can you identify the region of Europe where France is situated?"
- **Step 2:** Calculate the mean cosine similarity between the generated questions and the actual question.

The underlying concept is that if the answer correctly addresses the question, it is highly probable that the original question can be reconstructed solely from the answer.

Faithfulness

The **Faithfulness** metric measures how factually consistent a `response` is with the `retrieved context`. It ranges from 0 to 1, with higher scores indicating better consistency.

A response is considered **faithful** if all its claims can be supported by the retrieved context.

To calculate this:
1. Identify all the claims in the response.
2. Check each claim to see if it can be inferred from the retrieved context.
3. Compute the faithfulness score using the formula:

$$\text{Faithfulness Score} = \frac{\text{Number of claims in the response supported by the retrieved context}}{\text{Total number of claims in the response}}$$

Example

Question: Where and when was Einstein born?

Context: Albert Einstein (born 14 March 1879) was a German-born theoretical physicist, widely held to be one of the greatest and most influential scientists of all time

High faithfulness answer: Einstein was born in Germany on 14th March 1879.

Low faithfulness answer: Einstein was born in Germany on 20th March 1879.

Let's examine how faithfulness was calculated using the low faithfulness answer:

- **Step 1:** Break the generated answer into individual statements.
  - Statements:
    - Statement 1: "Einstein was born in Germany."
    - Statement 2: "Einstein was born on 20th March 1879."
- **Step 2:** For each of the generated statements, verify if it can be inferred from the given context.
  - Statement 1: Yes
  - Statement 2: No
- **Step 3:** Use the formula depicted above to calculate faithfulness.

$$\text{Faithfulness} = \frac{1}{2} = 0.5$$

`NoiseSensitivity` measures how often a system makes errors by providing incorrect responses when utilizing either relevant or irrelevant retrieved documents. The score ranges from 0 to 1, with lower values indicating better performance. Noise sensitivity is computed using the `user_input`, `reference`, `response`, and the `retrieved_contexts`.

To estimate noise sensitivity, each claim in the generated response is examined to determine whether it is correct based on the ground truth and whether it can be attributed to the relevant (or irrelevant) retrieved context. Ideally, all claims in the answer should be supported by the relevant retrieved context.

$$\text{noise sensitivity (relevant)} = \frac{|\text{Total number of incorrect claims in response}|}{|\text{Total number of claims in the response}|}$$

Example

Question: What is the Life Insurance Corporation of India (LIC) known for?

Ground truth: The Life Insurance Corporation of India (LIC) is the largest insurance company in India, established in 1956 through the nationalization of the insurance industry. It is known for managing a large portfolio of investments.

Relevant Retrieval: - The Life Insurance Corporation of India (LIC) was established in 1956 following the nationalization of the insurance industry in India. - LIC is the largest insurance company in India, with a vast network of policyholders and a significant role in the financial sector. - As the largest institutional investor in India, LIC manages a substantial life fund, contributing to the financial stability of the country.

Irrelevant Retrieval: - The Indian economy is one of the fastest-growing major economies in the world, thanks to the sectors like finance, technology, manufacturing etc.

Let's examine how noise sensitivity in relevant context was calculated:

- **Step 1:** Identify the relevant contexts from which the ground truth can be inferred.
  - Ground Truth: The Life Insurance Corporation of India (LIC) is the largest insurance company in India, established in 1956 through the nationalization of the insurance industry. It is known for managing a large portfolio of investments.
  - Contexts:
    - Context 1: The Life Insurance Corporation of India (LIC) was established in 1956 following the nationalization of the insurance industry in India.
    - Context 2: LIC is the largest insurance company in India, with a vast network of policyholders and a significant role in the financial sector.
    - Context 3: As the largest institutional investor in India, LIC manages a substantial funds, contributing to the financial stability of the country.
- **Step 2:** Verify if the claims in the generated answer can be inferred from the relevant conte
  - Answer: The Life Insurance Corporation of India (LIC) is the largest insurance company India, known for its vast portfolio of investments. LIC contributs to the financial stability the country.
  - Contexts:
    - Context 1: The Life Insurance Corporation of India (LIC) was established in 1956 following the nationalization of the insurance industry in India.
    - Context 2: LIC is the largest insurance company in India, with a vast network of policyholders and a significant role in the financial sector.
    - Context 3: As the largest institutional investor in India, LIC manages a substantial funds, contributing to the financial stability of the country.
- **Step 3:** Identify any incorrect claims in the answer (i.e., answer statements that are not supported by the ground truth).
  - Ground Truth: The Life Insurance Corporation of India (LIC) is the largest insurance company in India, established in 1956 through the nationalization of the insurance industry. It is known for managing a large portfolio of investments.
  - Answer: The Life Insurance Corporation of India (LIC) is the largest insurance company India, known for its vast portfolio of investments. LIC contributs to the financial stability the country.

  Explanation: The ground truth does not mention anything about LIC contributing to the financial stability of the country. Therefore, this statement in the answer is incorrect.

  Incorrect Statement: 1 Total claims: 3
- **Step 4:** Calculate noise sensitivity using the formula:

$$\text{noise sensitivity} = \frac{1}{3} = 0.333$$

HW 8 Prelab

RAG Metrics and Evaluation

TIM 175 WEEK 8 PRELAB

Brief Task Overview:

- Brainstorming Metrics for RAG
- Understanding existing Metrics for RAG
- Manually Evaluating with RAG metrics
- Using RAGAS for Evaluating with RAG metrics