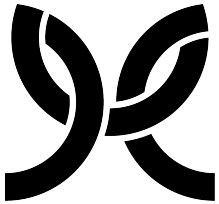
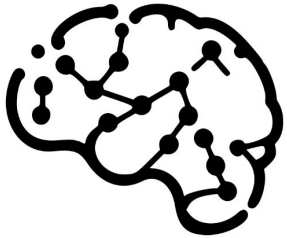
MODUS CREATE

Evaluating the evaluator: RAG eval libraries under the loop

Nour El Mawass, Maria Knorps

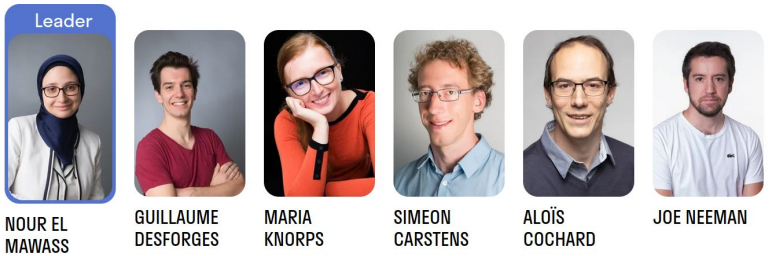
25-09-2024





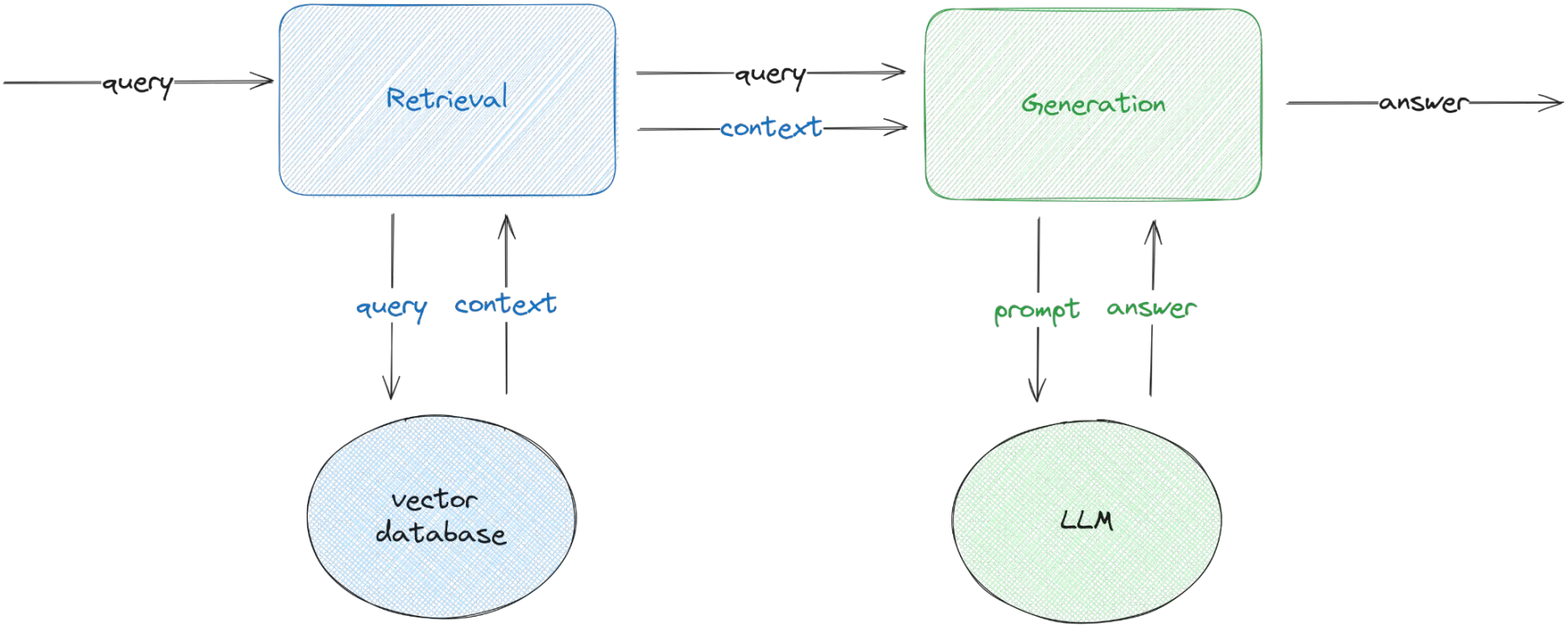
GENERATIVE AI

MEMBERS

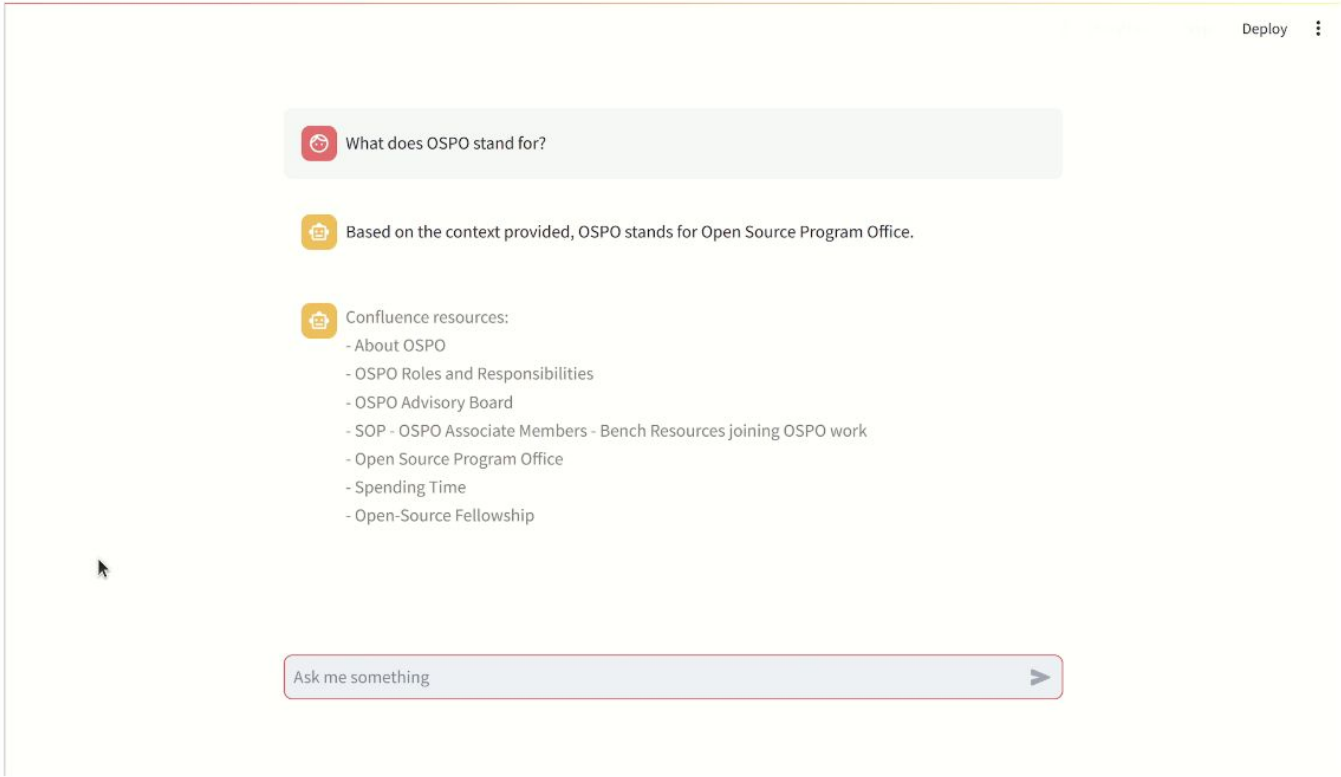


www.tweag.io/group/genai

Retrieval Augmented Generation

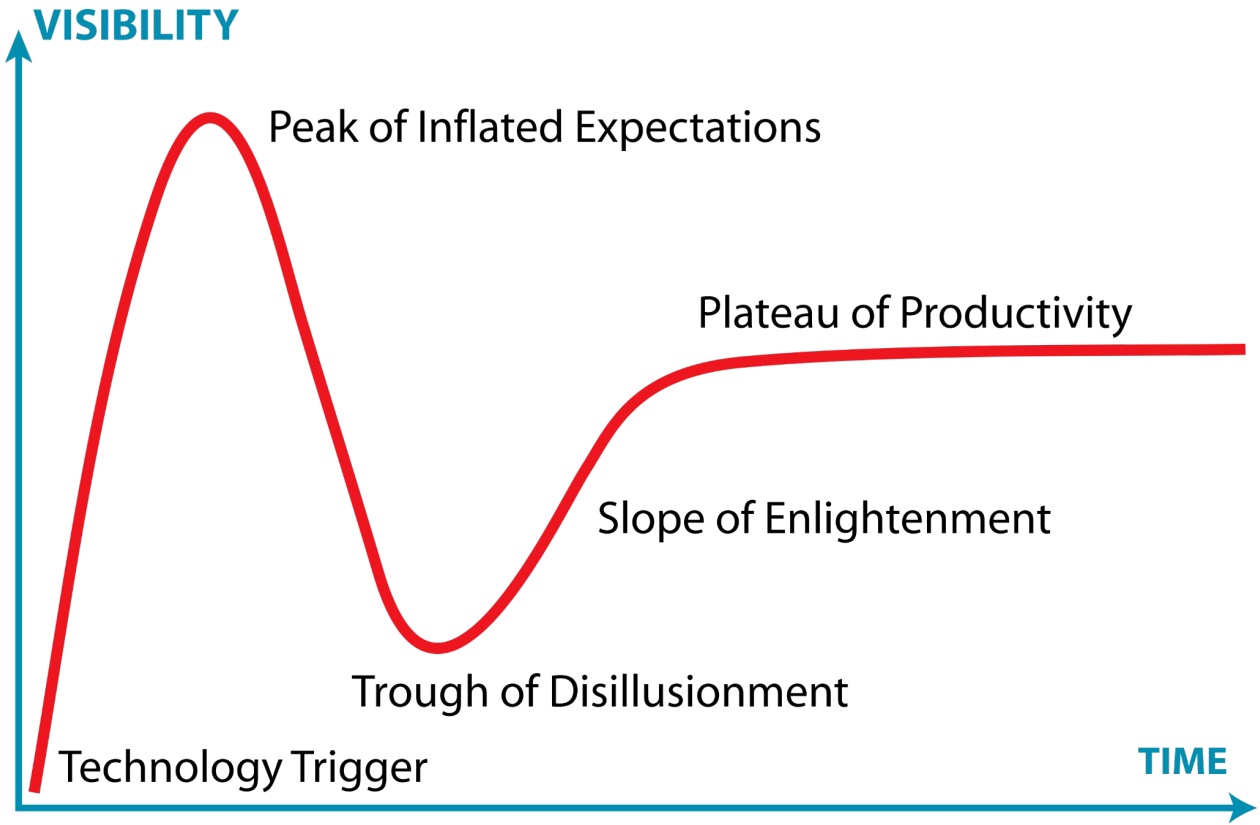


RAG in action

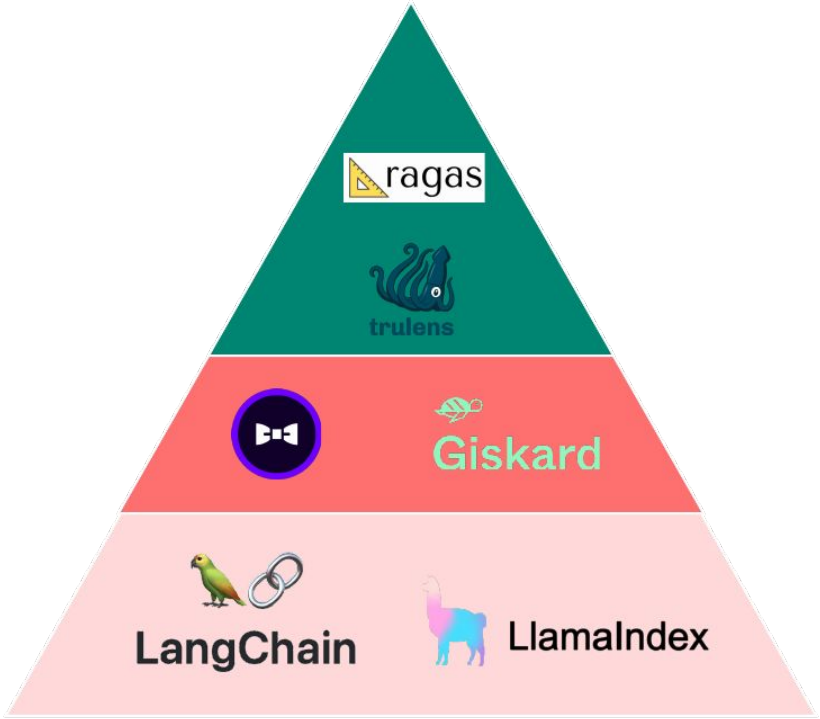


Why do we need to evaluate RAGs?

Why do we need to evaluate RAGs?

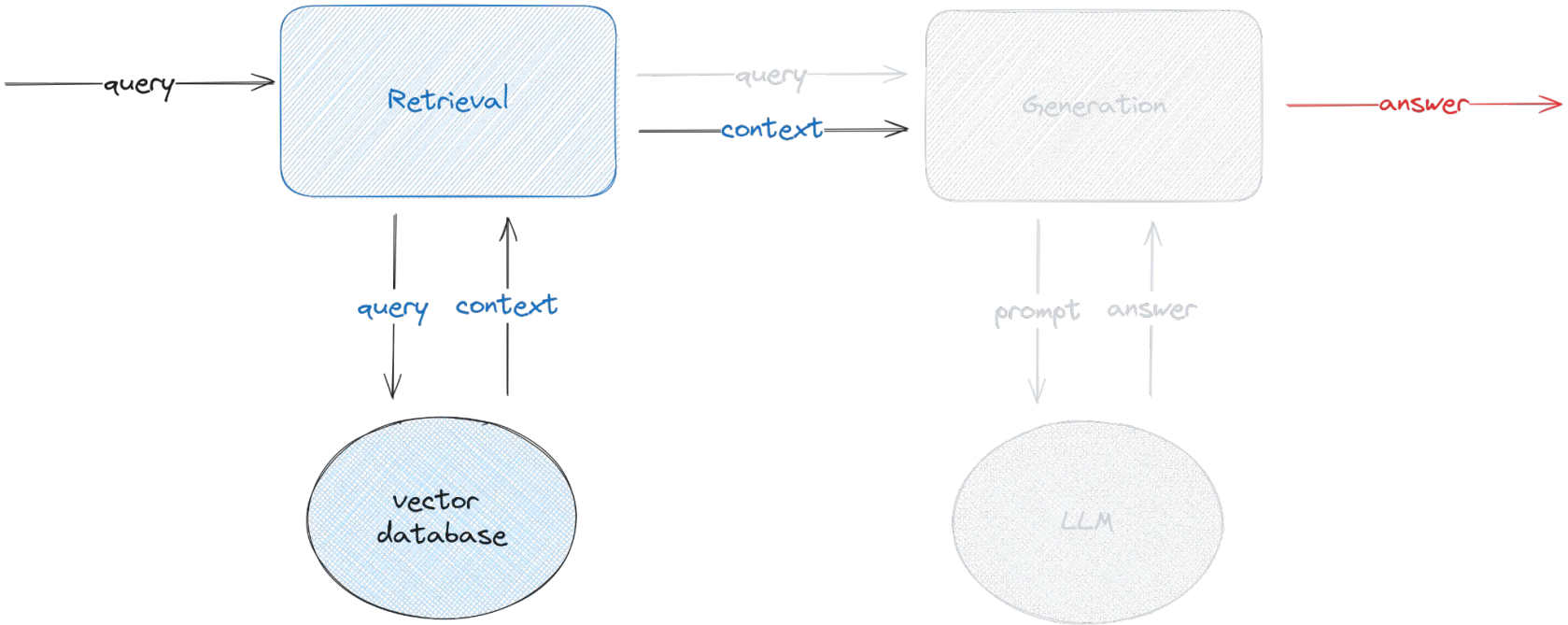


Libraries for RAG evaluation



Metrics for RAG evaluation

context precision, context recall



Context recall

Question: Where is France and what is its capital?

Ground truth: France is in Western Europe and its capital is Paris.

High context recall: France, in Western Europe, encompasses medieval cities, alpine villages and Mediterranean beaches. Paris, its capital, is famed for its fashion houses, classical art museums including the Louvre and monuments like the Eiffel Tower.

Low context recall: France, in Western Europe, encompasses medieval cities, alpine villages and Mediterranean beaches. The country is also renowned for its wines and sophisticated cuisine. Lascaux's ancient cave drawings, Lyon's Roman theater and the vast Palace of Versailles attest to its rich history.

https://docs.ragas.io/en/stable/concepts/metrics/context_recall.html

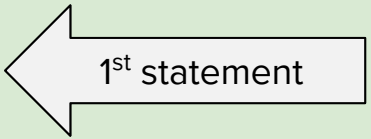
Context recall

$$\text{context recall} = \frac{|\text{GT claims that can be attributed to context}|}{|\text{Number of claims in GT}|}$$

Step 1: find statements

Question: Where is France and what is its capital?

Ground truth: France is in Western Europe and its capital is Paris.



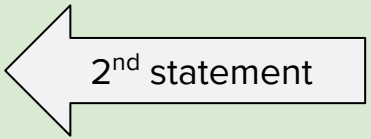
High context recall: France, in Western Europe, encompasses medieval cities, alpine villages and Mediterranean beaches. Paris, its capital, is famed for its fashion houses, classical art museums including the Louvre and monuments like the Eiffel Tower.

Low context recall: France, in Western Europe, encompasses medieval cities, alpine villages and Mediterranean beaches. The country is also renowned for its wines and sophisticated cuisine. Lascaux's ancient cave drawings, Lyon's Roman theater and the vast Palace of Versailles attest to its rich history.

Step 1: find statements

Question: Where is France and what is its capital?

Ground truth: France is in Western Europe and its capital is Paris.



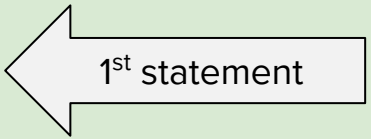
High context recall: France, in Western Europe, encompasses medieval cities, alpine villages and Mediterranean beaches. Paris, its capital, is famed for its fashion houses, classical art museums including the Louvre and monuments like the Eiffel Tower.

Low context recall: France, in Western Europe, encompasses medieval cities, alpine villages and Mediterranean beaches. The country is also renowned for its wines and sophisticated cuisine. Lascaux's ancient cave drawings, Lyon's Roman theater and the vast Palace of Versailles attest to its rich history.

Step2: is statement supported?

Question: Where is France and what is its capital?

Ground truth: France is in Western Europe and its capital is Paris.



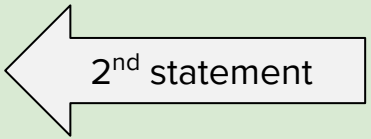
High context recall: France, in Western Europe, encompasses medieval cities, alpine villages and Mediterranean beaches. Paris, its capital, is famed for its fashion houses, classical art museums including the Louvre and monuments like the Eiffel Tower.

Low context recall: France, in Western Europe, encompasses medieval cities, alpine villages and Mediterranean beaches. The country is also renowned for its wines and sophisticated cuisine. Lascaux's ancient cave drawings, Lyon's Roman theater and the vast Palace of Versailles attest to its rich history.

Step2: is statement supported?

Question: Where is France and what is its capital?

Ground truth: France is in Western Europe and its capital is Paris.



High context recall: France, in Western Europe, encompasses medieval cities, alpine villages and Mediterranean beaches. Paris, its capital, is famed for its fashion houses, classical art museums including the Louvre and monuments like the Eiffel Tower.

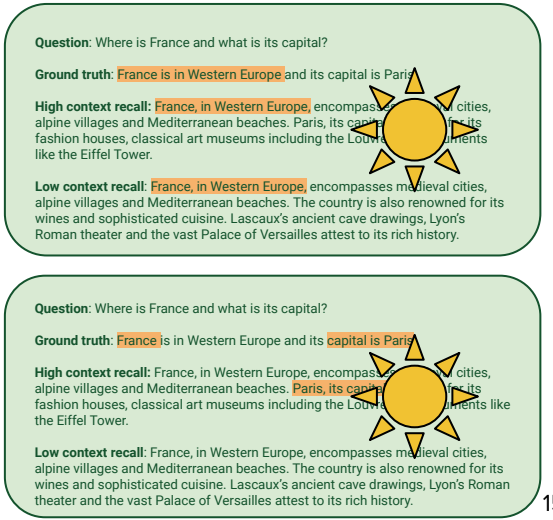
Low context recall: France, in Western Europe, encompasses medieval cities, alpine villages and Mediterranean beaches. The country is also renowned for its wines and sophisticated cuisine. Lascaux's ancient cave drawings, Lyon's Roman theater and the vast Palace of Versailles attest to its rich history.

Step 3: compute

$$\text{context recall} = \frac{|\text{GT claims that can be attributed to context}|}{|\text{Number of claims in GT}|}$$

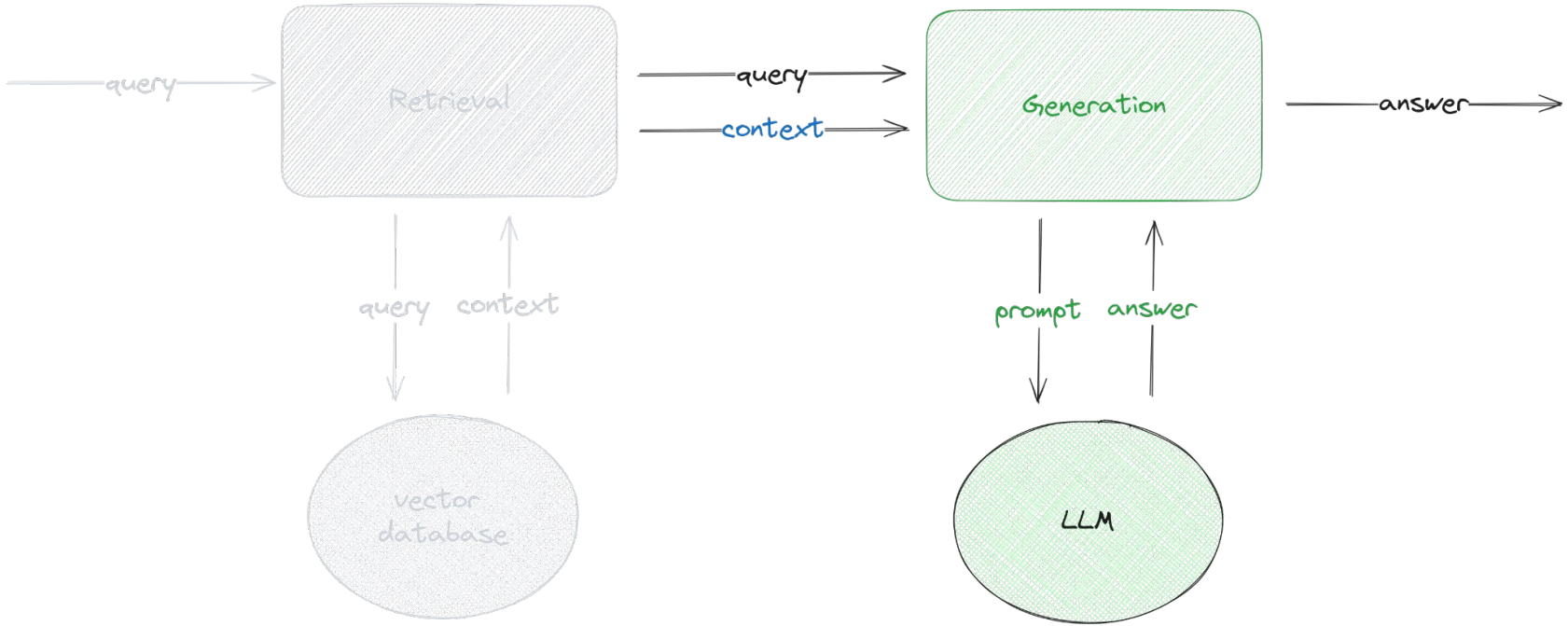
Answer 1 : $\text{context recall} = \frac{2}{2} = 1$

Answer 2 : $\text{context recall} = \frac{1}{2}$



Metrics for RAG evaluation

faithfulness, answer relevancy



Faithfulness

$$\text{faithfulness} = \frac{|\text{Claims in the generated answer that can be inferred from the context}|}{|\text{All claims in the generated answer}|}$$

Question: Where and when was Einstein born?

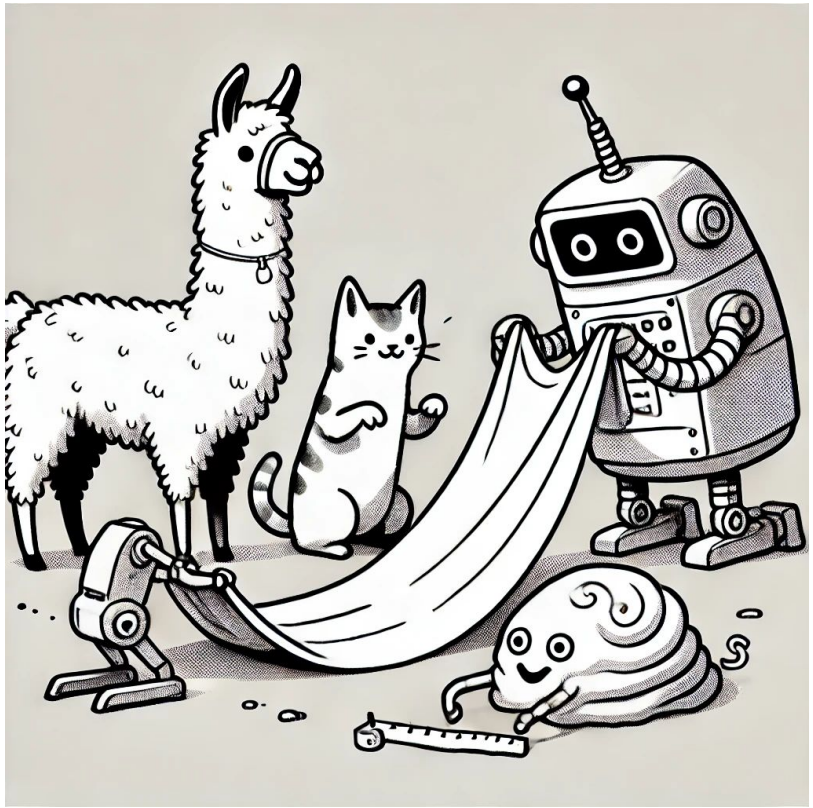
Context: Albert Einstein (born 14 March 1879) was a German-born theoretical physicist, widely held to be one of the greatest and most influential scientists of all time

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

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

https://docs.ragas.io/en/stable/concepts/metrics/faithfulness.html

Evaluating the evaluator



MEMBERS

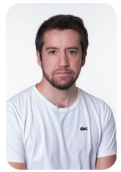
NOUR EL MAWASS

GUILLAUME DESFORGES

MARIA KNORPS

SIMEON CARSTENS

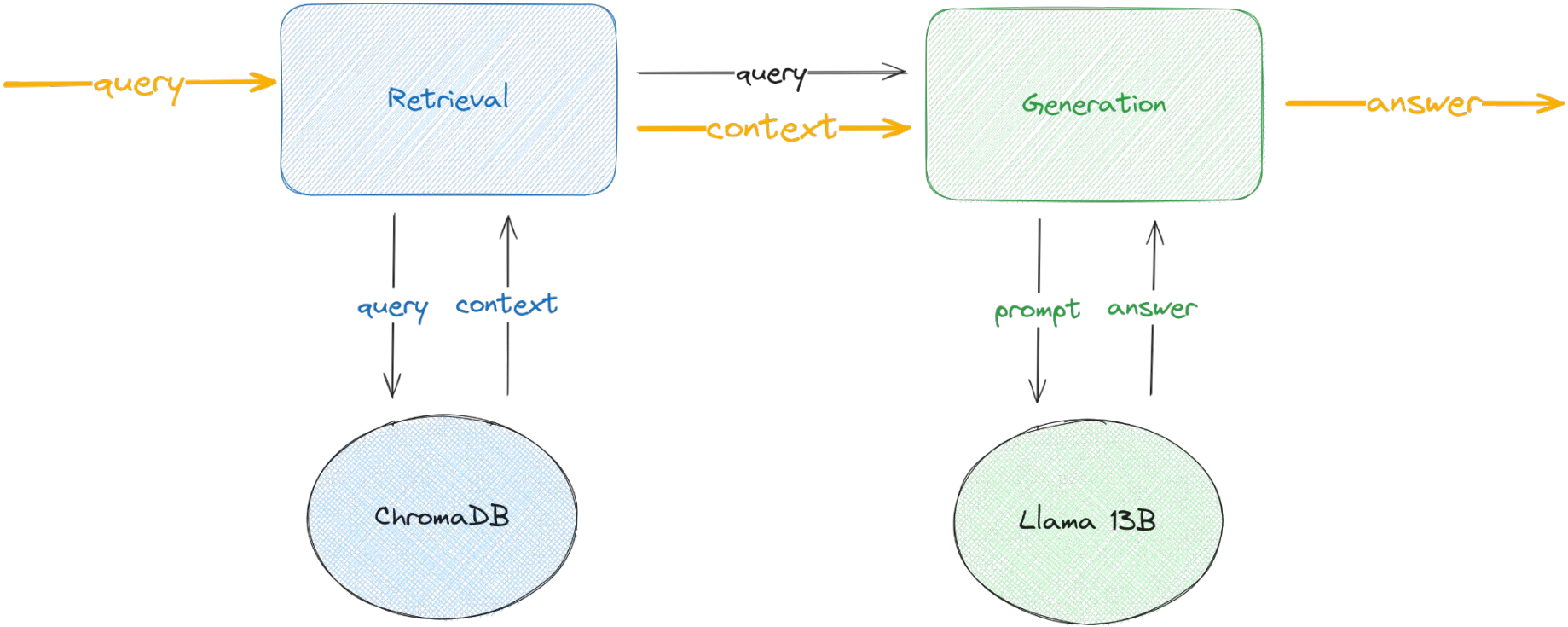
ALOÏS COCHARD



JOE NEEMAN

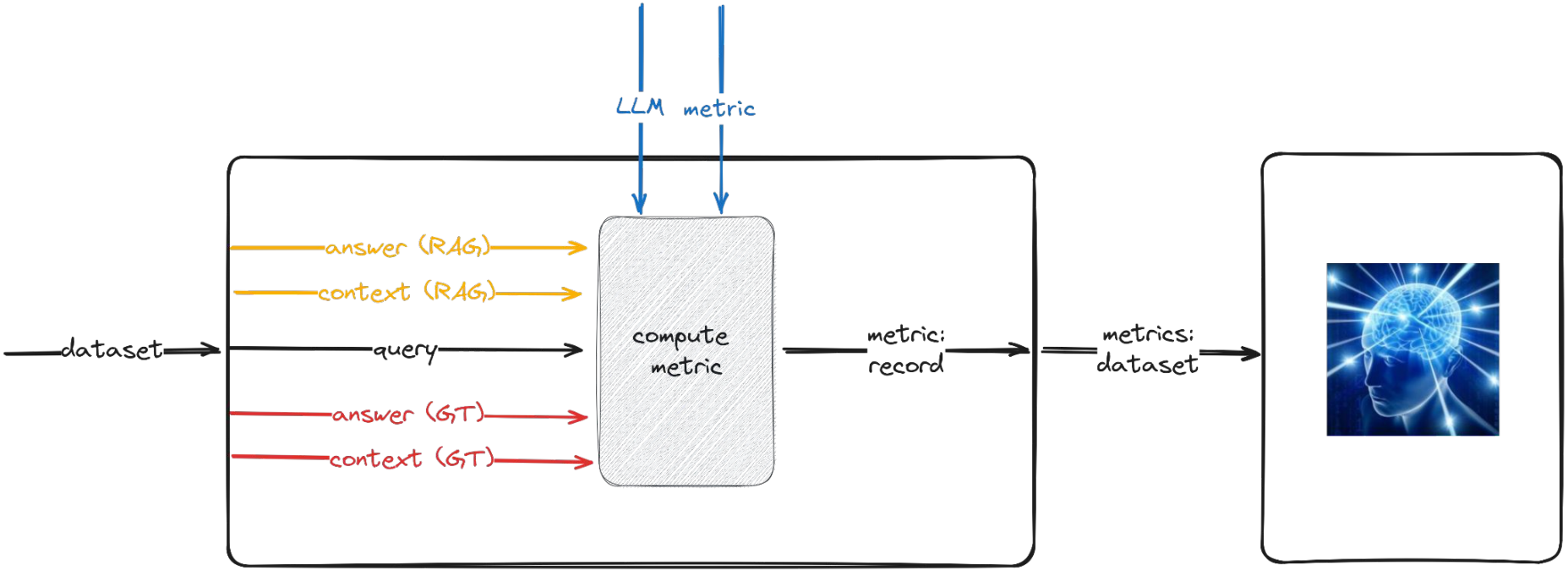
Experiment

Data: RAG results



https://huggingface.co/datasets/neural-bridge

Outline



https://huggingface.co/datasets/neural-bridge

Experiment matrix

	answer relevancy	context precision	context recall	faithfulness
claude-3-sonnet-20240229-v1:0 				
llama2-70b-chat-v1 				
llama3-70b-instruct-v1:0 				
gpt-4 				
gpt-3.5-turbo-16k 				

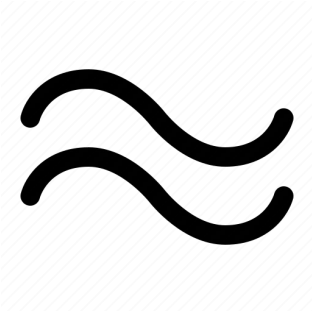
Experiment questions

Correctness

- Which LLMs evaluate which metrics correctly?

Coherence

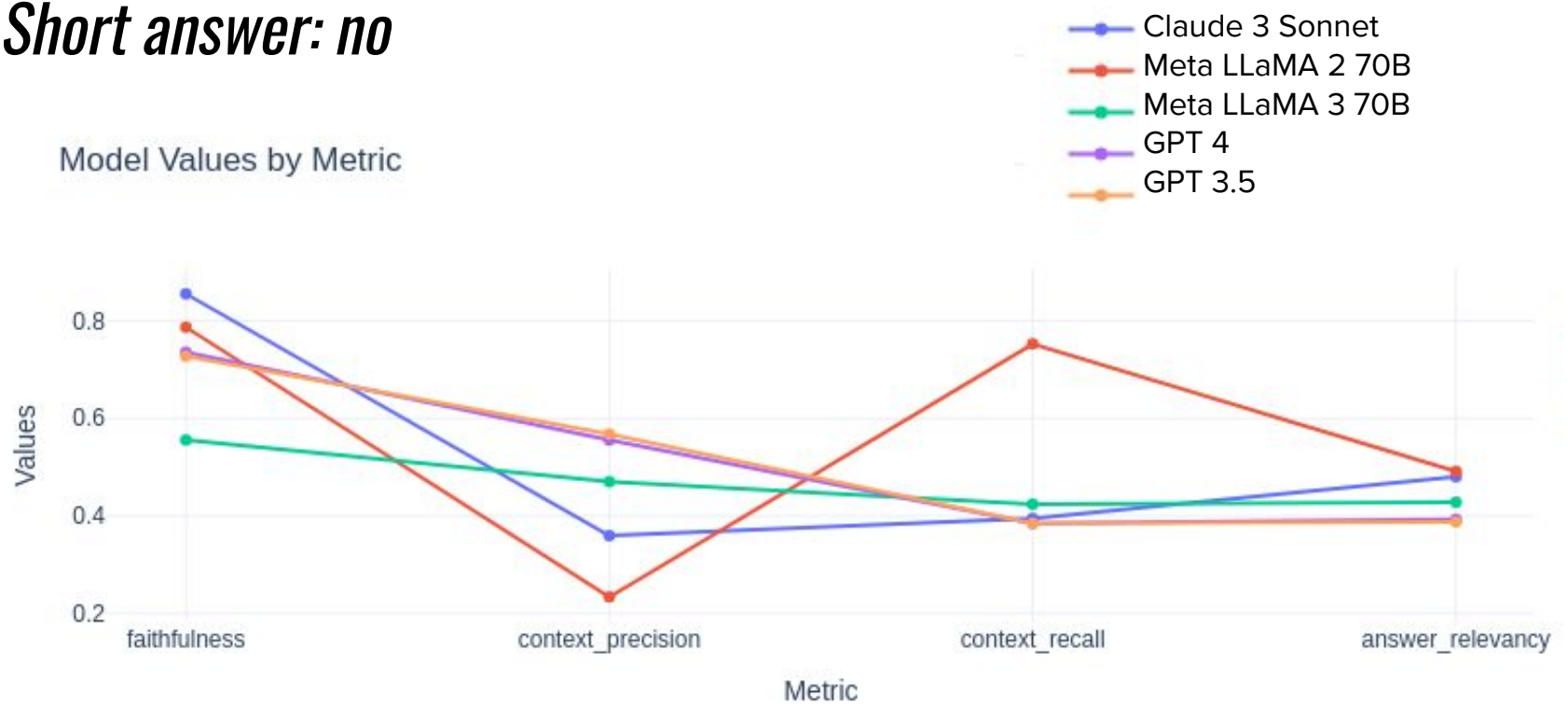
- Are LLMs in agreement?
  - Same values
  - Same understanding of metrics



Results

Are judge LLMs in agreement?

Short answer: no



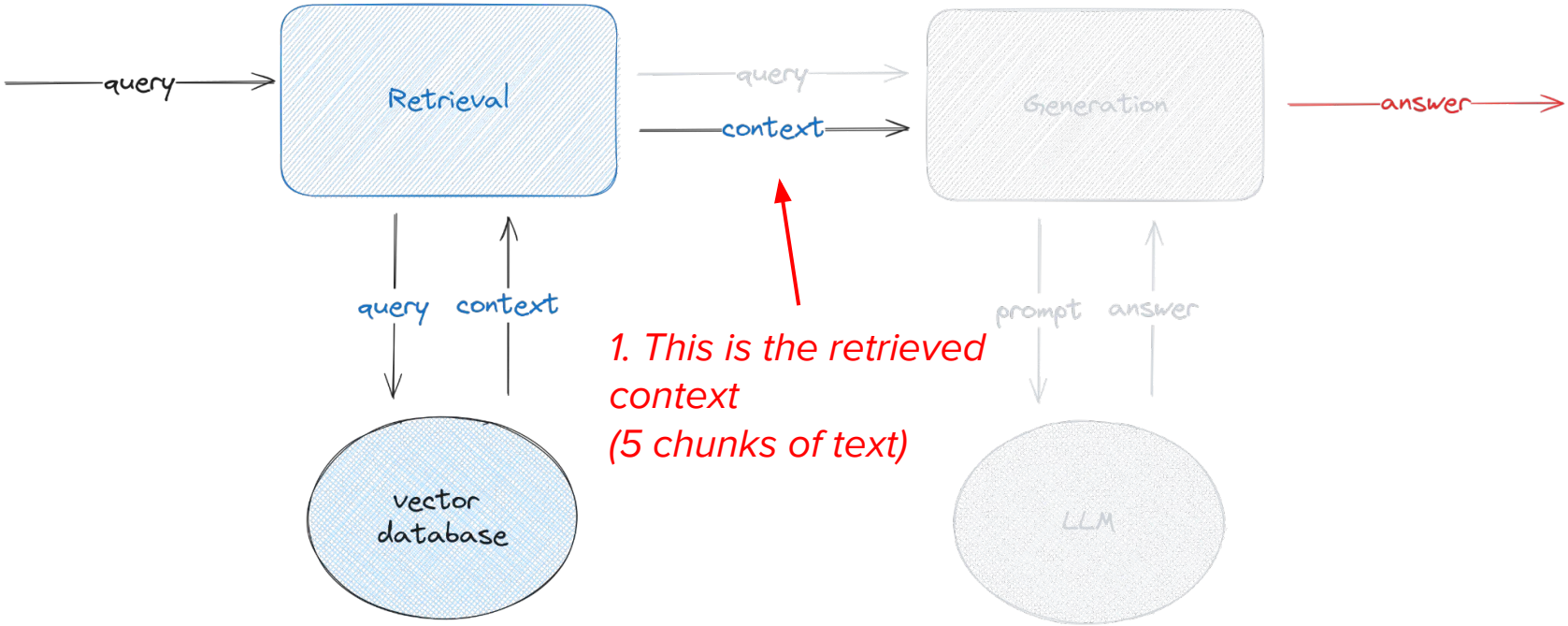
Correctness

Values for a metric are different ➜ not all LLMs are correct

Correctness: context recall & precision

3. We can compute the overlap!

2. We know the groundtruth context



1. This is the retrieved context (5 chunks of text)

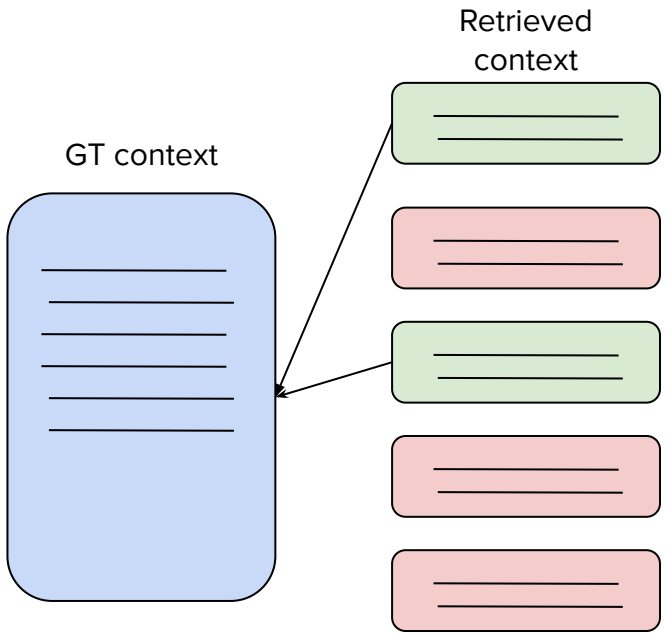
Estimating context recall and precision
(without LLMs)

$$\text{Context recall} = \frac{\text{GT context} \cap \text{retrieved context}}{\text{GT context}}$$

$$\text{Context precision} = \frac{\text{GT context} \cap \text{retrieved context}}{\text{retrieved context}}$$

Note: We can compute this because we have the ground truth context.

Estimating context recall and precision (without LLMs) - example

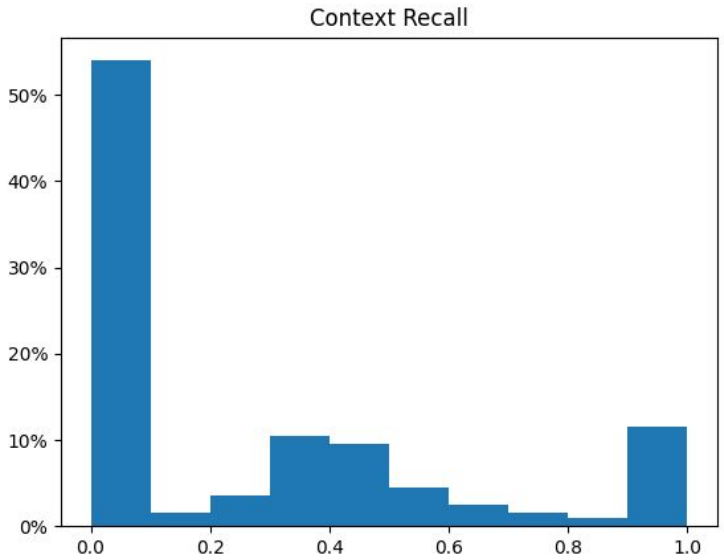


- Overlap = 2 chunks (out of 5)
- context recall
  = length of overlap / length of GT context
  = 4/6
  = 67%
- context precision
  = overlapping chunks / total of retrieved chunks
  = 2/5
  = 40%

Estimating context recall and precision
(without LLMs)

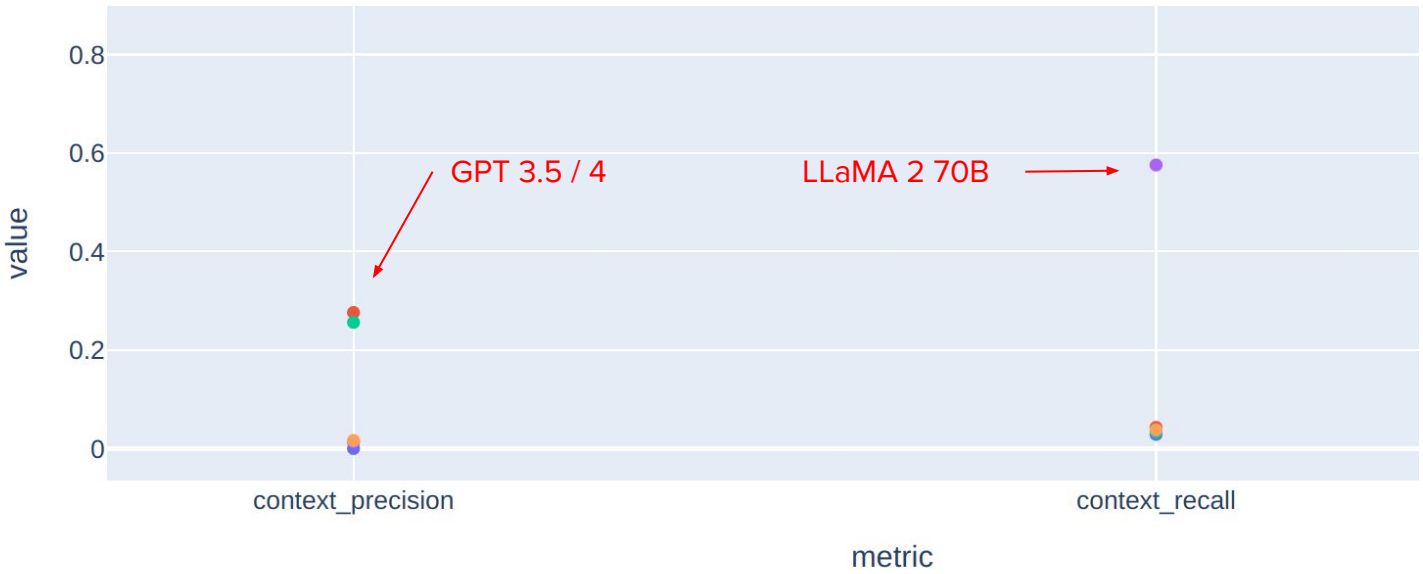
- **48%** of questions have **zero** relevant chunk in their retrieved context
  ➡ context precision and recall are also **zero**

- **52%** have only 1 relevant chunk (out of 5)
  ➡ context precision is 20%
  ➡ context recall (see plot)



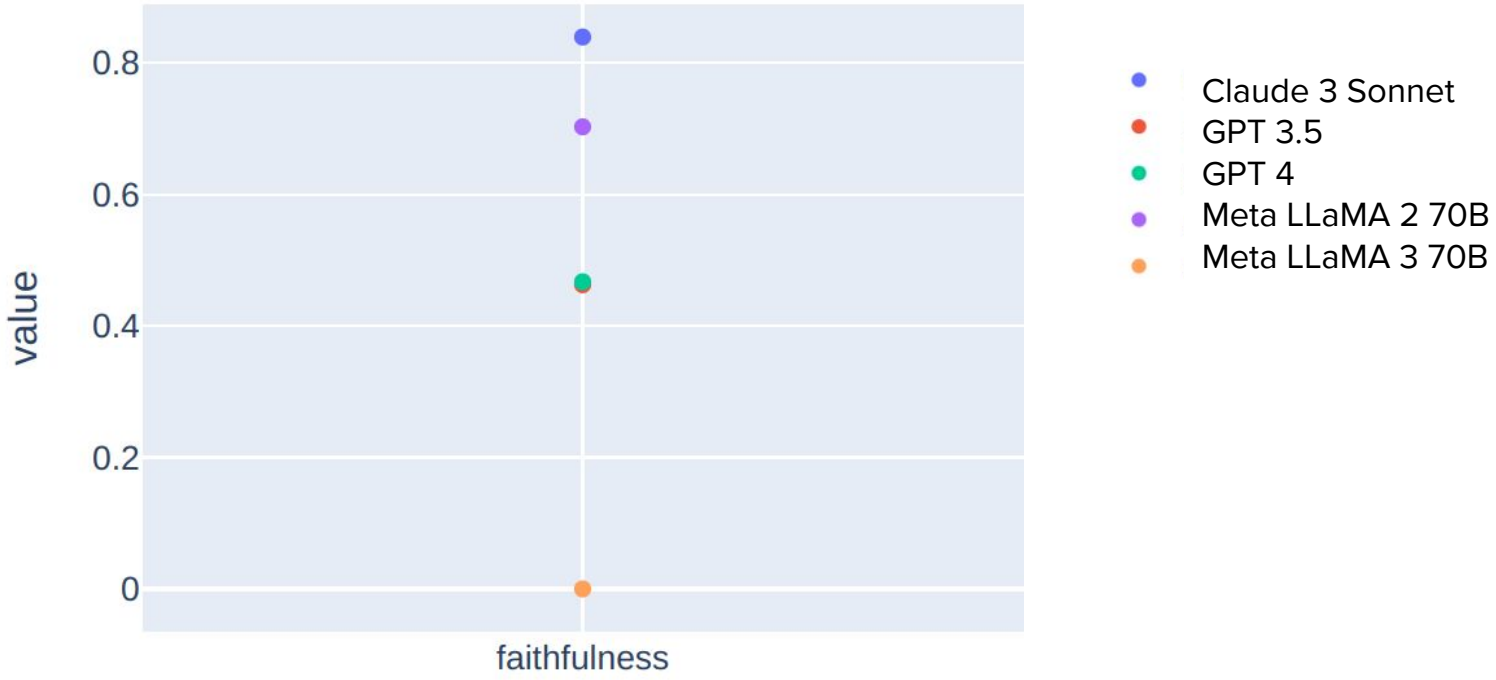
Correctness: Context recall/precision

With 0% relevant chunks in retrieved context



Correctness: faithfulness

With 0% relevant chunks in retrieved context



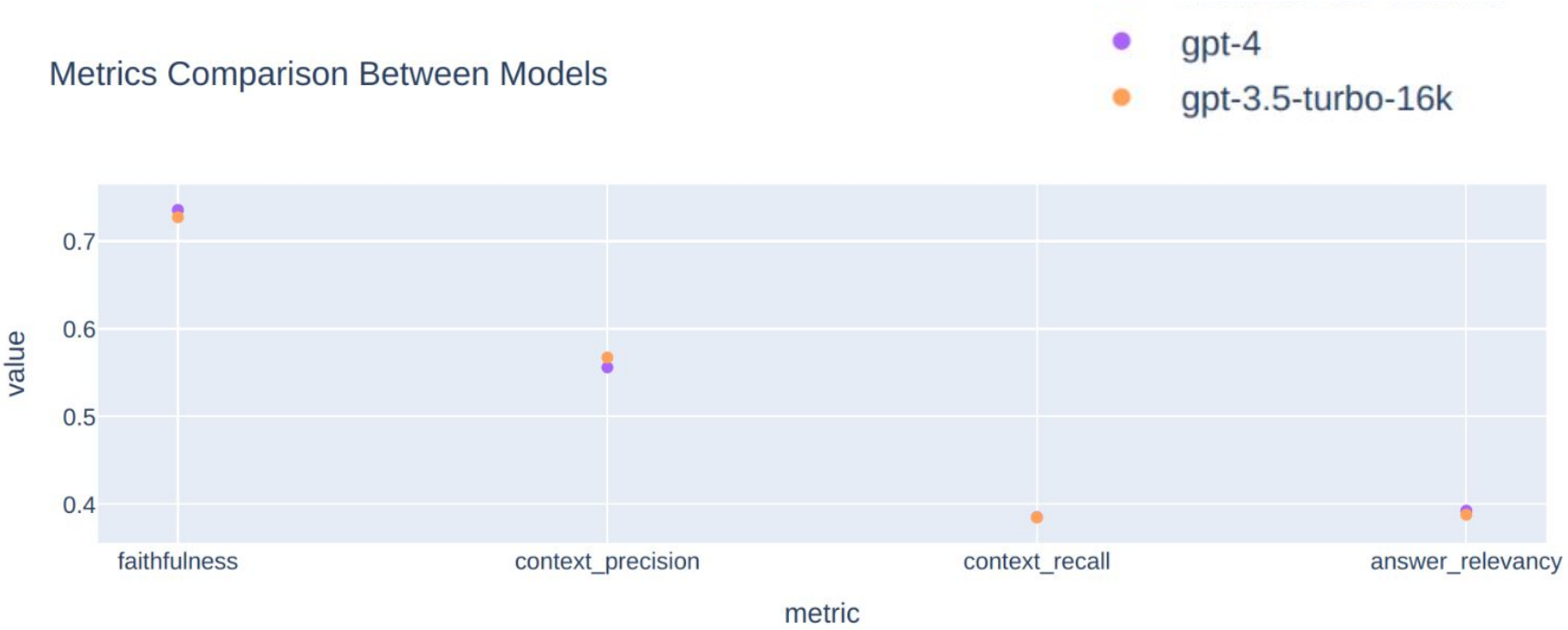
Correctness: faithfulness

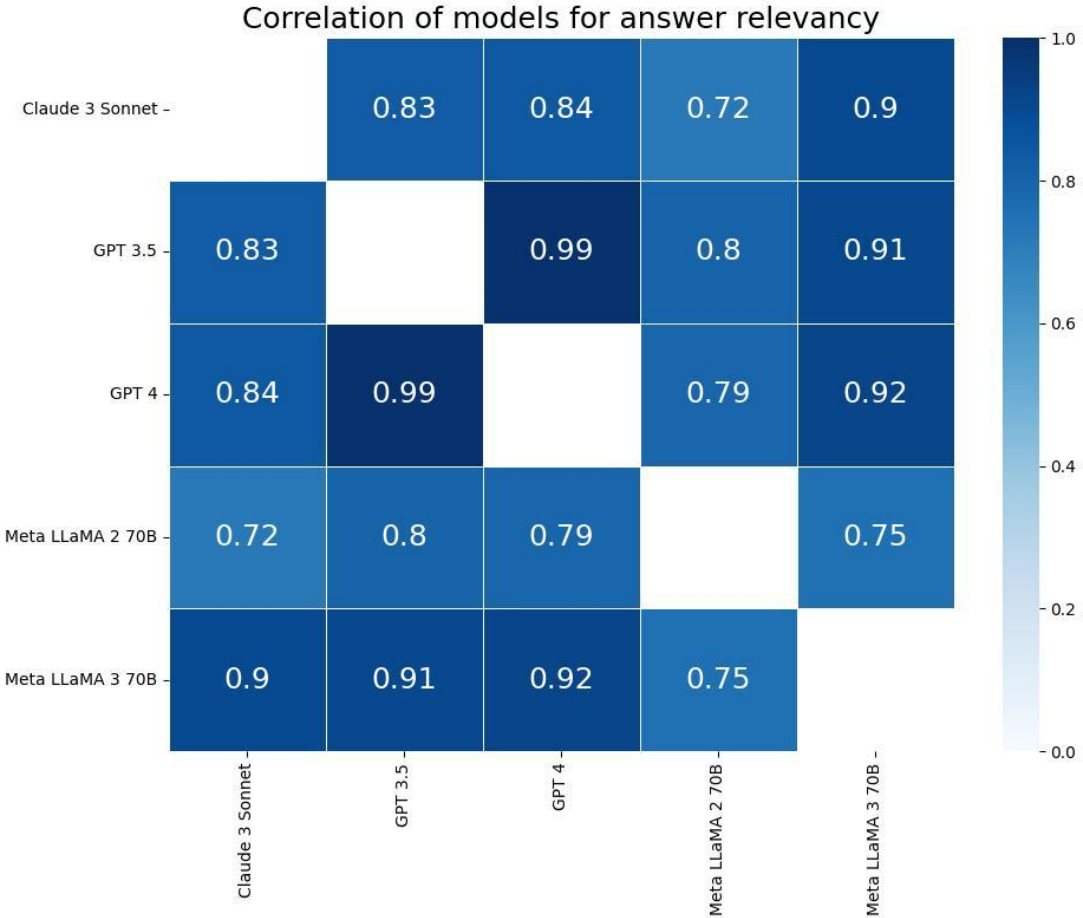
Question: Who did LeBron James decide to join for the 2010-11 NBA season?

Answer: **I don't know.** The context provided does not mention LeBron James or his decision to join any team for the 2010-11 NBA season. The context is focused on the University of Kentucky men's basketball team and their victory over Alabama, as well as the cyber attack on Sony and the company's response.

Claude 3: 1
GPT 3.5, 4, LLaMA 2 and 3: Null

Model coherence inside families
OpenAI




Correlation of models for answer relevancy
Claude 3 Sonnet
GPT 3.5
GPT 4
Meta LLaMA 2 70B
Meta LLaMA 3 70B
0.83
0.84
0.72
0.9
0.83
0.99
0.8
0.91
0.84
0.99
0.79
0.92
0.72
0.8
0.79
0.75
0.9
0.91
0.92
0.75
1.0
0.8
0.6
0.4
0.2
0.0

Model coherence inside families
Meta LLaMA

Challenges: Null metrics

- Different models might interpret the metrics prompts differently
  - [ ] **vs no output** when no claims can be extracted
- **Rate limit** issues (OpenAI)
- **Empty replies** for LLaMA3

Challenges: Implementation

- Ragas computes the **average metric** for a dataset
- We used an internal RAGAS structure to get the values for each question
  - **Batch processing does not produce results in the same order**
- **Impossible to change temperature** (set to 0 for consistency on most metrics and to 0.3 on answer relevancy)
  - But determinism is good for reproducibility

Cost considerations

Model / Metric	RAG dataset generation	Faithfulness
Claude 3 Sonnet	$0.91	$5.3
GPT 4	$1.41	$8.5
GPT 3.5	$0.14	$0.8
LLaMA 2 70B	$0.51	$3.2
LLaMA 3 70B	$0.69	$4.3

Future work

- Only one RAG setting tested
  - Study how evaluation metrics are impacted in **different settings**
- **Compare LLM-computed metrics to human-computed ones**

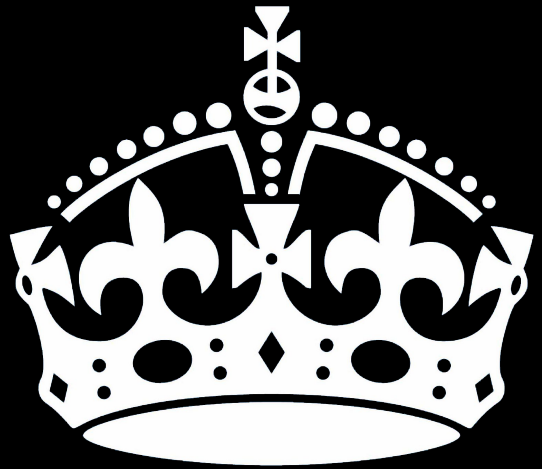
Take away lessons

- RAG eval Python libraries like RAGAS and Trulens can streamline the evaluation process

- But LLM-based metrics have issues:
  - LLMs rarely agree on the metric value
  - There is no clear LLM winner (but some LLMs are performing worse than others)
  - The same metric prompt is interpreted differently by different LLMs
  - Technical issues can result in Null values

Our technical blog:
tweag.io/blog/tags/generative-ai/

Our consultancy platform:
moduscreate.com/



KEEP
CALM
AND
EVALUATE