Explainable LLM

05/02/2025

Context

- Project in collaboration with AI Factory of CACIB

- LLMs used by professionals in investment banking (chatbots, text generation,...)

- Problem : How to ensure that an LLM's response is coherent and meets certain criteria ?

Stakes of the project

To facilitate the tasks of CACIB's market finance professionals by providing LLMs that can be used as financial assistants, capable of delivering quick, reliable, and coherent responses.

Evaluate and quantify the model's output (define evaluation metrics)

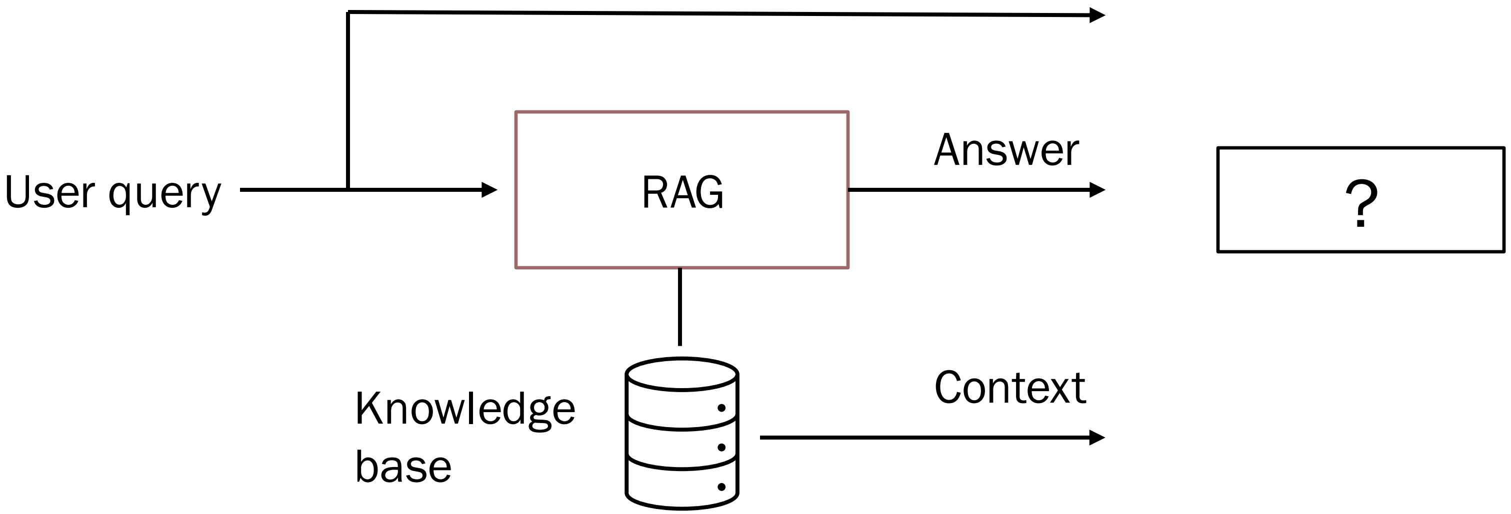
→

Understand, detect and mitigate hallucinations (XAI)

→

Assess an answer using multiple criteria to guarantee « quality » of answer

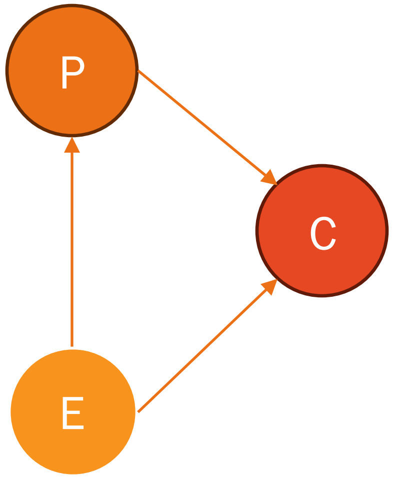
Problem



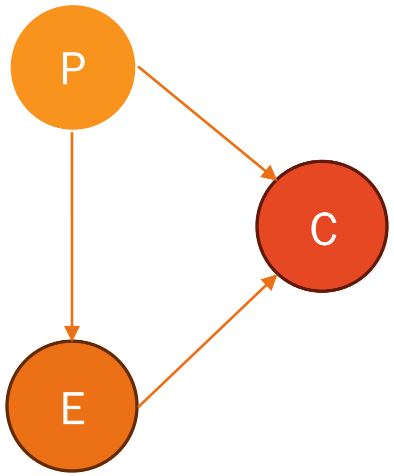
- **Goal:** Obtain explanations over a given RAG answer.

Causal construction

E: External context or knowledge, P: parametric knowledge, C: explaination criteria

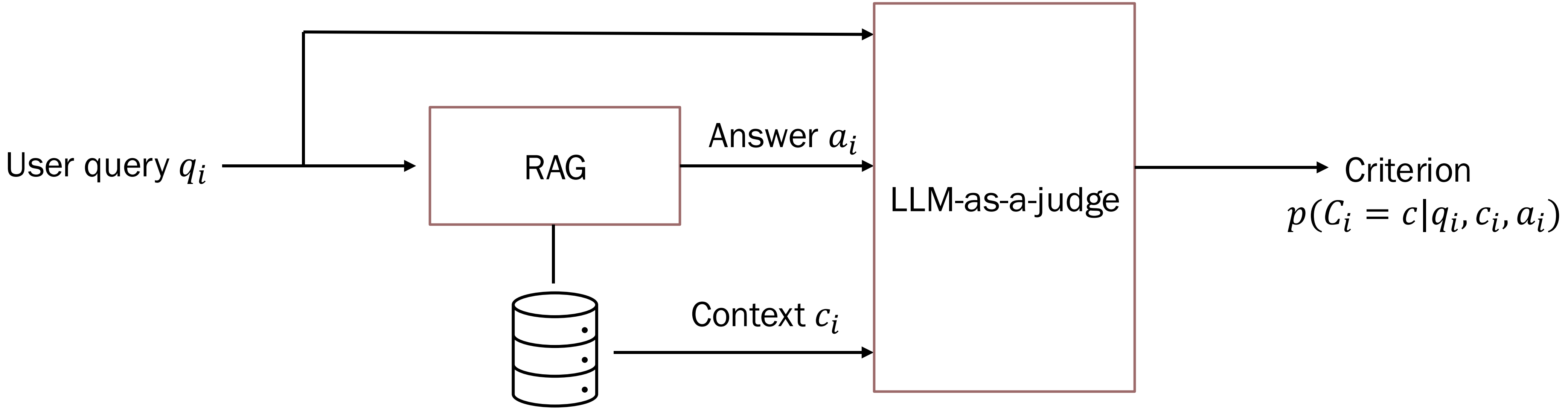


P confounded by E:
which relies on the LLM's hidden states for some criterion



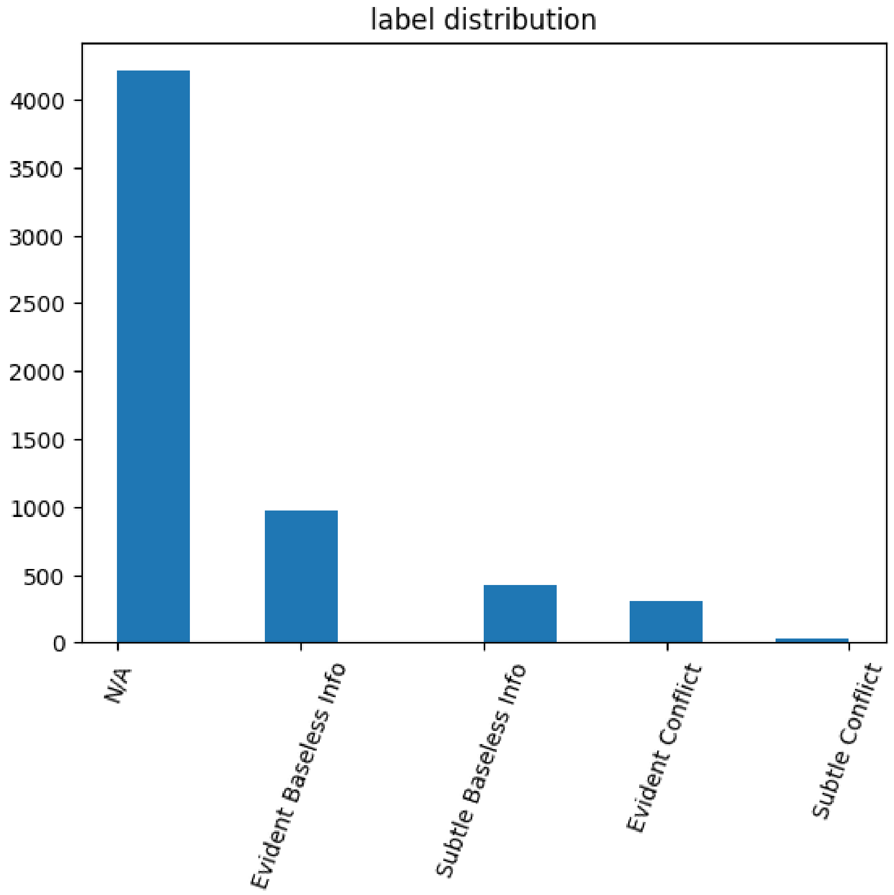
E confounded by P :
by leveraging external context and model responses

Solution



- **Goal:** Obtain explanations over a given RAG answer using a criterion C_i
- **Assumption**: the RAG is black box model i.e. we look for model-agnostic method

RAGTruth dataset

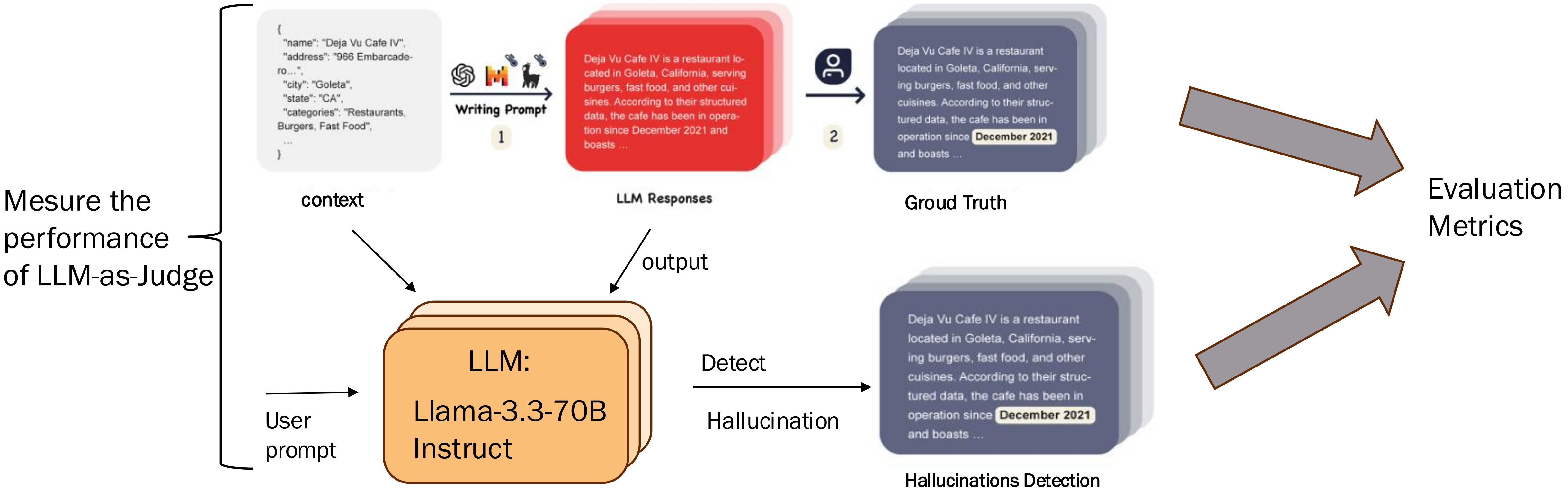


- Examples of RAG outputs

- Annotated data for hallucination

- Hallucination tasks:
  - QuestionAnswering
  - Data-to-textWriting
  - Summarization

Hallucination detection



Hallucination detection

- Evaluation Metrics:
  - <u>Faithfullness</u> = $\frac{\# \, true \, statements}{\# statements}$ --> concentrate on true statements != paper approach (detect halluciantion)
  - <u>Answer relevancy</u>
    - Generate questions q_i based on the provided RAG answer.
    - $AR = \frac{1}{n}\sum_{i=1}^{n} \text{cosine} \_sim(q_i, q)$

  --> not at all a good metric (no consideration of context)

  - <u>Response-level Detection</u>

    Accuracy, **precision**, **recall**, **F1 score** for each detection algorithm and its variants across different tasks

    --> Sample based approach (detect if the overall sentence contains halucination)

  - <u>Span-level Detection</u>

    overlap between detected span and human-labeled span and report the precision,recall,and f1score

  --> Word level evaluation

Results: Llama-3.3-70B

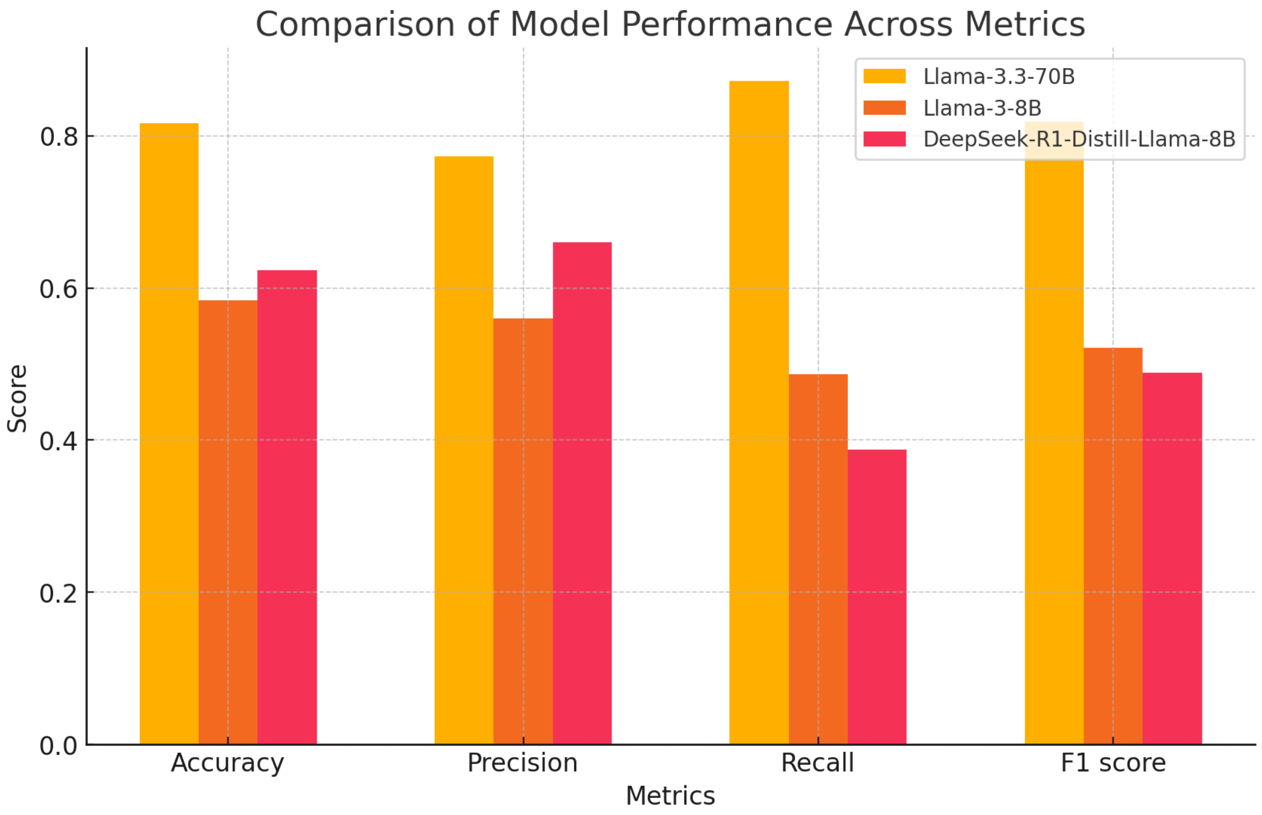
task_type	Hallucinated output	count
Data2txt	0	90
Data2txt	1	210
QA	0	196
QA	1	99
Summary	0	194
Summary	1	106

Task	Nb samples	Accuracy	Precision	Recall	F1 score
Overall performance	895	0.817	0.773	0.872	0.819
QA	295	0.792	0.727	0.800	0.762
Summary	300	0.815	0.786	0.846	0.815
Data2Text	300	0.839	0.789	0.938	0.857

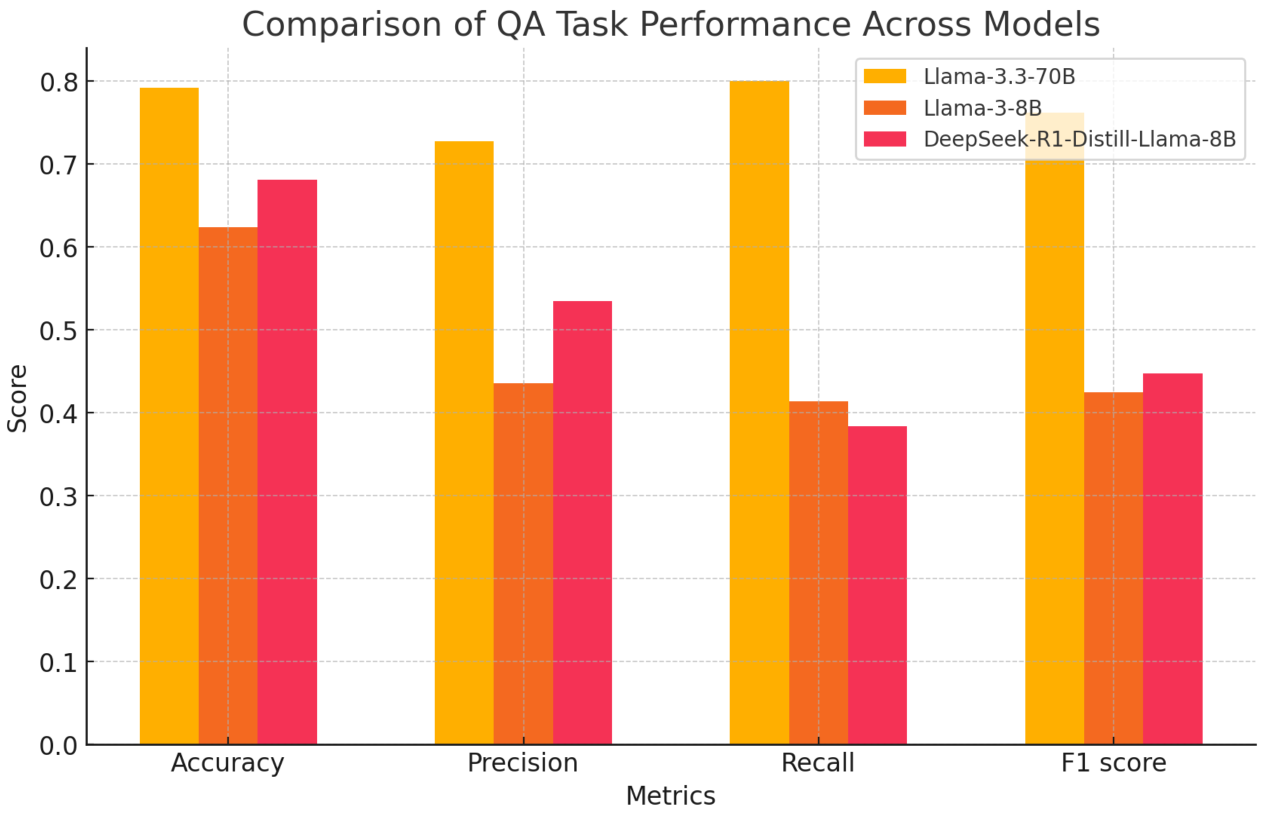
- We can achive better performance by **finetuning** the LLM as a judge

--> Halucination will be detected easily for general data

Results:

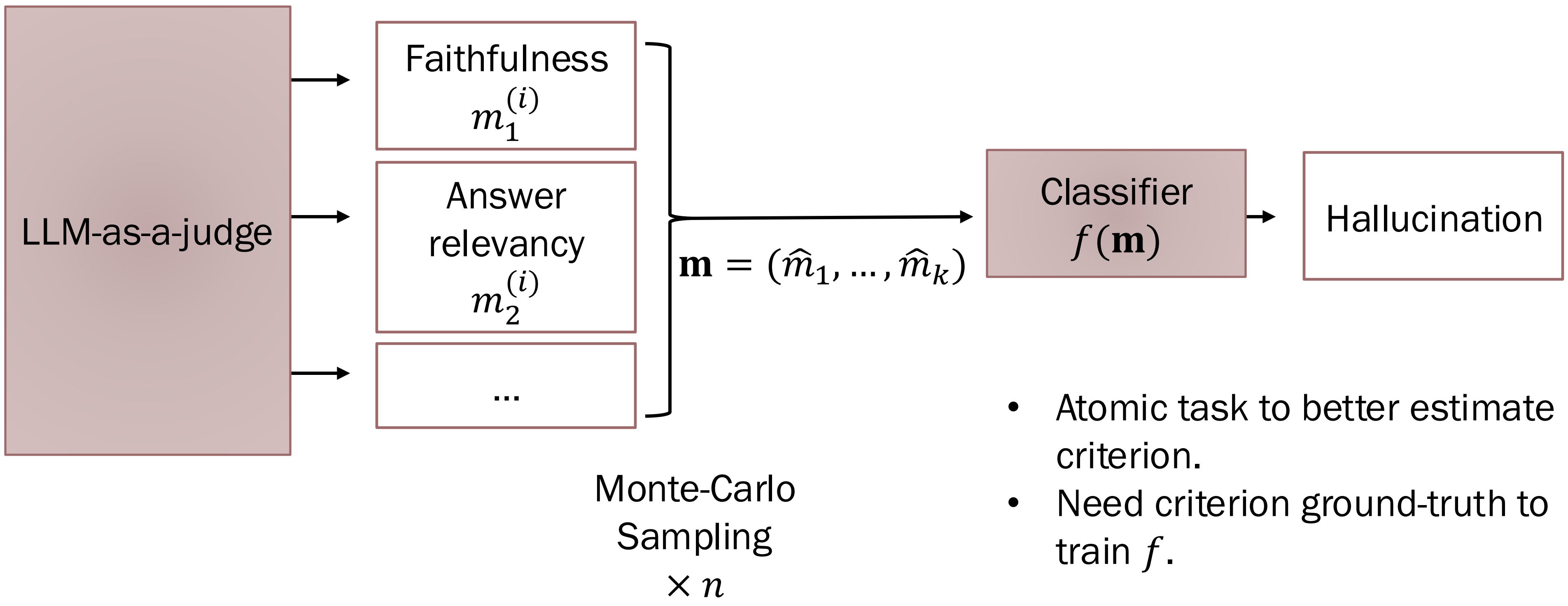


Comparison of Model Performance Across Metrics
Llama-3.3-70B
Llama-3-8B
DeepSeek-R1-Distill-Llama-8B
Score
0.0
0.2
0.4
0.6
0.8
Accuracy
Precision
Recall
F1 score
Metrics



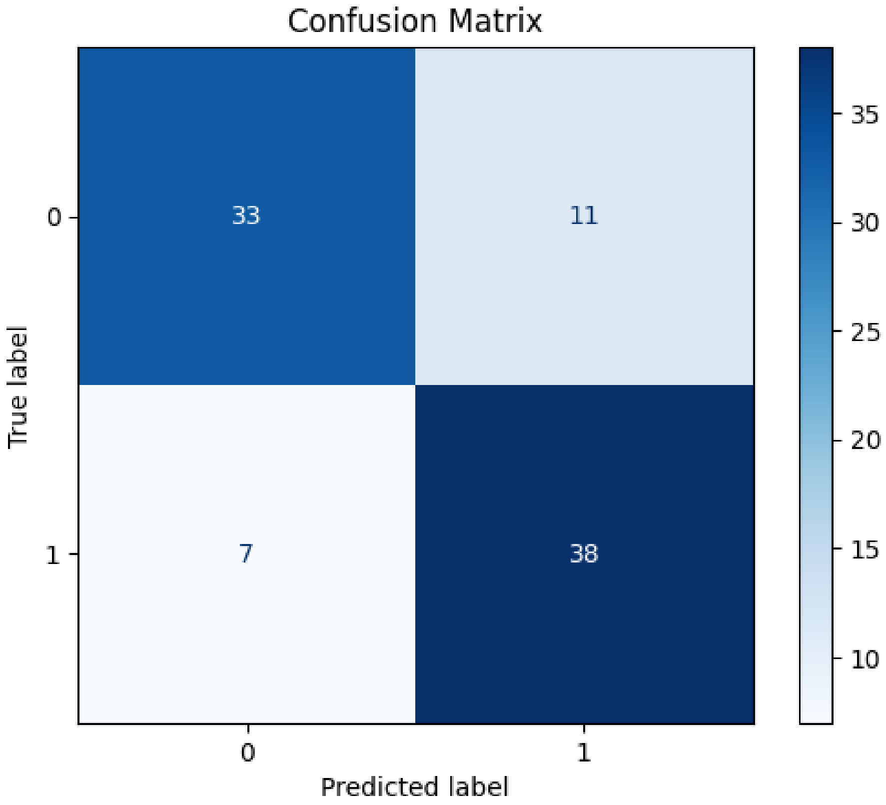
Comparison of QA Task Performance Across Models
Llama-3.3-70B
Llama-3-8B
DeepSeek-R1-Distill-Llama-8B
Score
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Accuracy
Precision
Recall
F1 score
Metrics

Metric factor model



- Atomic task to better estimate criterion.
- Need criterion ground-truth to train f.

Factor model results

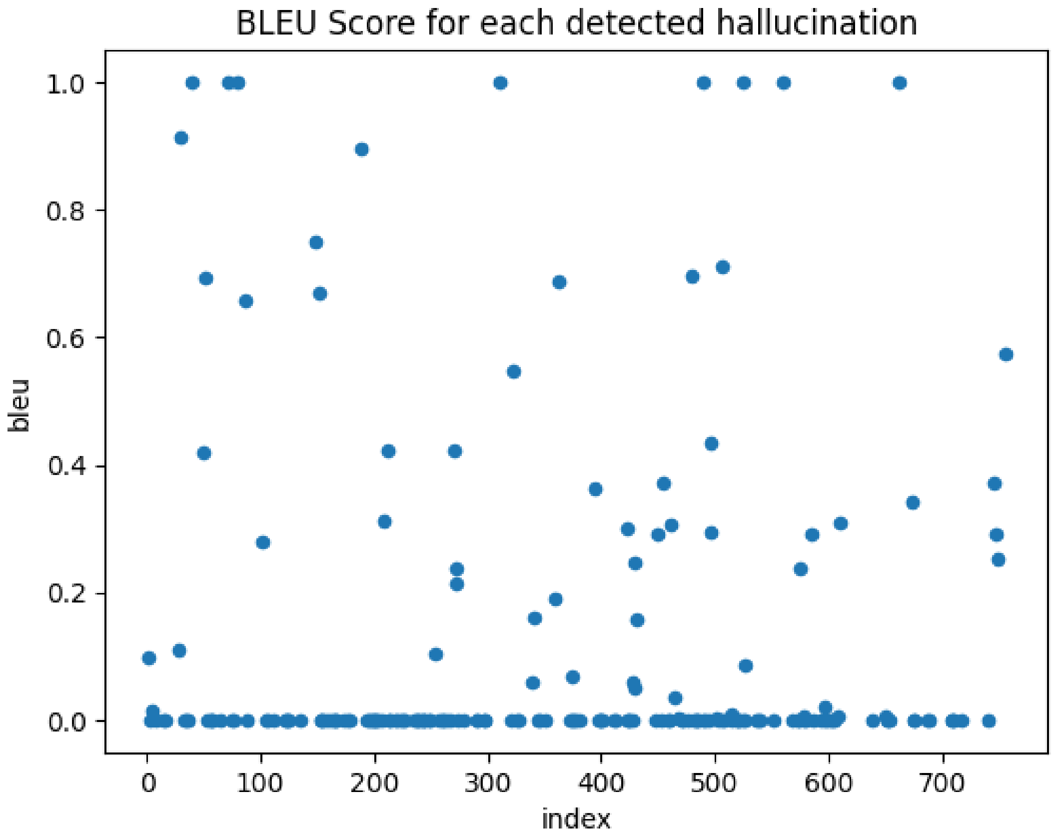


LLM Specs	f	Accuracy	Precision	Recall	F1 score	Correlation
CoT[1]/ $T = 0.1$	FFNN	0.80	0.78	0.84	0.81	0.60
$T = 1$	FFNN	0.66	0.55	0.65	0.61	0.33
Benchmark		0.79	0.73	0.80	0.76	0.58

[1]CHAIN-OF-THOUGH : (WEI ET AL.) HTTPS://ARXIV.ORG/PDF/2201.11903

More towards explainability

- LLM-as-a-judge → provide explanation.
- RAGTruth provides hallucinated segments.
- We compare the two of them using simple NLP scores such as BLEU.



Next steps

- LLM as a judge
  - Test the performance of the model on **Financial data**
  - Test the performance of the model based on **other Metrics : Completeness , Toxicity**

ANNEX

Llama-3-8B

Task	Nb samples	Accuracy	Precision	Recall	F1 score
Overall performance	895	0.584	0.560	0.487	0.521
QA	295	0.624	0.436	0.414	0.425
Summary	300	0.583	0.398	0.349	0.372
Data2 Text	300	0.547	0.713	0.590	0.646

DeepSeek-R1-Distill-Llama-8B

Task	Nb samples	Accuracy	Precision	Recall	F1 score
Overall performance	895	0.623	0.660	0.388	0.489
QA	295	0.681	0.535	0.384	0.447
Summary	300	0.593	0.367	0.208	0.265
Data2 Text	300	0.597	0.894	0.481	0.625

Examples of criteria

Binary criterions	Related papers
Hallucination	ReDEEP Sun et al.[1] Eigenscore Chen et al.[2]
Completeness	RAGAS Es et al.[3]
Toxicity	Achintalwar et al.[4]

[1]HTTP://ARXIV.ORG/ABS/2410.11414 [2]HTTP://ARXIV.ORG/ABS/2402.03744 [3]HTTP://ARXIV.ORG/ABS/2309.15217 [4]HTTP://ARXIV.ORG/ABS/2403.06009