Large language model

PROGRESS REPORT
10/02/2025

Overview

1. LLM Evaluation
2. Retrieval Evaluation
3. Factor model results

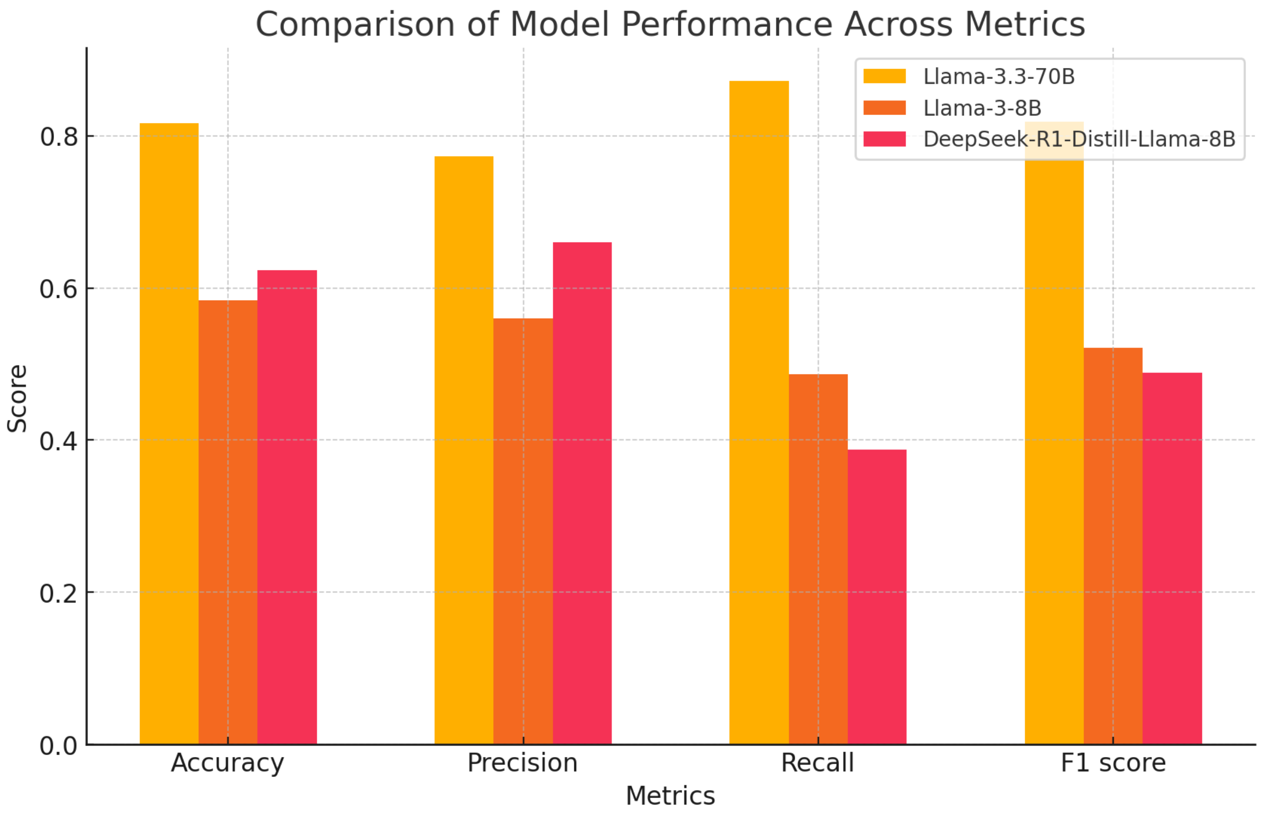
I. LLM Evaluation

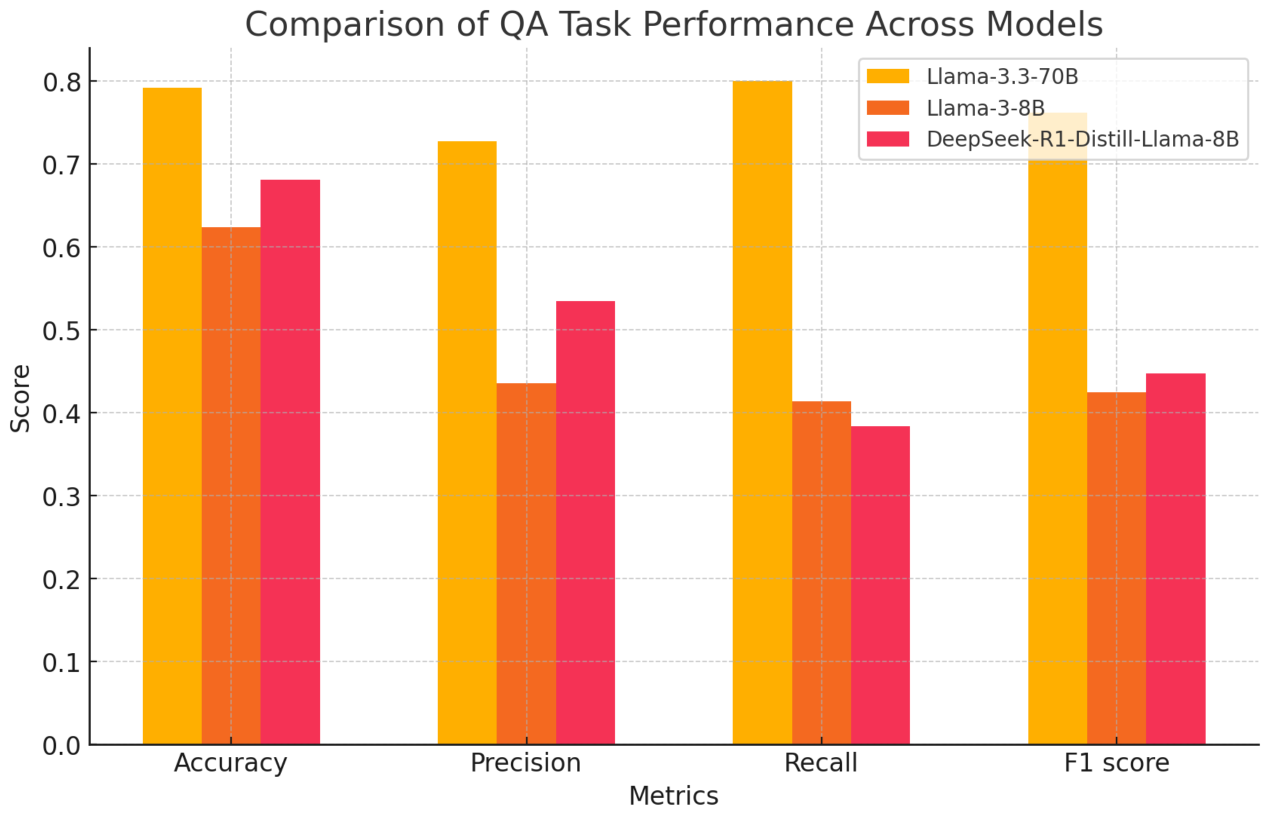
LLM Evaluation

- Llama-3.3-70B --> 70B parameters --> huge computational resources

- We tried 2 simpler models :
  - Llama-3-8B --> 8B parameters
  - DeepSeek-R1-Distill-Llama-8B

Results


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

II. Retrieval Evaluation

Dataset

Table 1: RAGBench component datasets.

Dataset	Domain	Document Source	Question Source	#docs	doc length	#Train	#Dev	#Test
PubMedQA	biomedical research	research abstracts	automated heuristics	4	99	19.5k	2.5k	2.5k
CovidQA-RAG	biomedical research	research papers	expert	4	122	2.5k	534	492
HotpotQA	general knowledge	wikipedia	crowd-sourced	4	126	3.7k	847	776
MS Marco	general knowledge	web pages	user web queries	10	94	3.7k	790	839
HAGRID	general knowl-edge	wikipedia	expert	3	153	2.0k	322	1.3k
ExpertQA	general knowl-edge	google search	expert	3	548	1.6k	202	203
CUAD	legal	legal contracts	expert	1	11k	1.5k	506	508
DelucionQA	customer support	Jeep manual	LLM	3	296	1.5k	177	182
EManual	customer support	TV manual	annotator	3	165	1k	132	132
TechQA	customer support	Technotes	tech forums	5	1.8k	1.2k	302	310
FinQA	finance	earning reports	expert	3	310	12k	1.7k	2.2k
TAT-QA	finance	financial reports	expert	5	96	26k	3.2k	3.2k
Total						78k	12k	11k

https://huggingface.co/datasets/rungalileo/ragbench

[cs.CL] 16 Jan 2025

RAGBench: Explainable Benchmark for Retrieval-Augmented Generation Systems

Robert Friel* Galileo Technologies Inc. rob@rungalileo.io
Masha Belyi* Galileo Technologies Inc. masha@rungalileo.io
Atindriyo Sanyal Galileo Technologies Inc. atin@rungalileo.io

Abstract

Retrieval-Augmented Generation (RAG) has become a standard architectural pattern for incorporating domain-specific knowledge into user-facing chat applications powered by Large Language Models (LLMs). RAG systems are characterized by (1) a document retriever that queries a domain-specific corpus for context information relevant to an input query, and (2) an LLM that generates a response based on the provided query and context. However, comprehensive evaluation of RAG systems remains a challenge due to the lack of unified evaluation criteria and annotated datasets. In response, we introduce RAGBench: the first comprehensive, large-scale RAG benchmark dataset of 100k examples. It covers five unique industry-specific domains and various RAG task types.

/arxiv.org/abs/2407.11005

Dataset

Feature	FINQA	TATQA
Main Focus	Financial reports and reasoning	Tabular data across multiple domains
Data Type	Text + Tables (Financial context)	Tables (Semi-structured, diverse topics)
Reasoning Complexity	High financial literacy required	Strong numerical reasoning across tables
Use Case	Financial document Q&A	General tabular data understanding

Evaluation Metrics

- Precision@K

$$\text{Precision@K} = \frac{\text{Relevant Retrieved Documents in Top K}}{K}$$

- Recall@K

$$\text{Recall@K} = \frac{\text{Relevant Retrieved Documents in Top K}}{\text{Total Ground-Truth Relevant Documents}}$$

- NDCG (Normalized Discounted Cumulative Gain)

--> Giving higher score if relevant documents

appear earlier

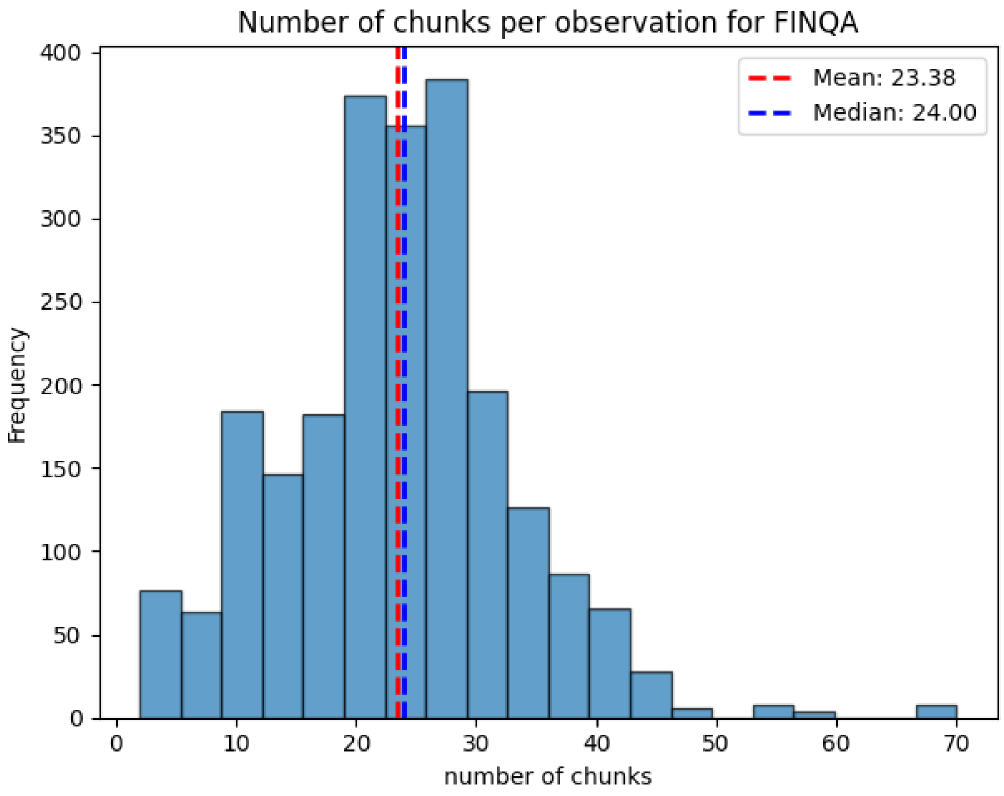
https://weaviate.io/blog/retrieval-evaluation-metrics

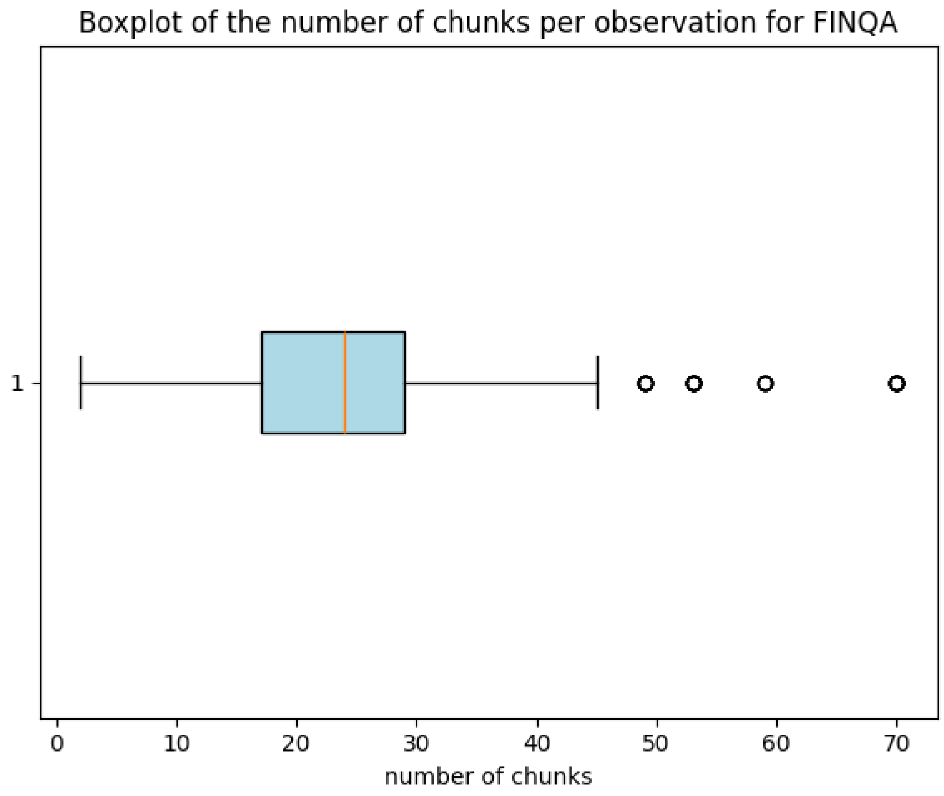
$$\text{DCG}_K = \sum_{i=1}^{K} \frac{\text{Relevance Score}_i}{\log_2(i+1)}$$

$$\text{NDCG}_K = \frac{\text{DCG}_K}{\text{Ideal DCG}_K}$$

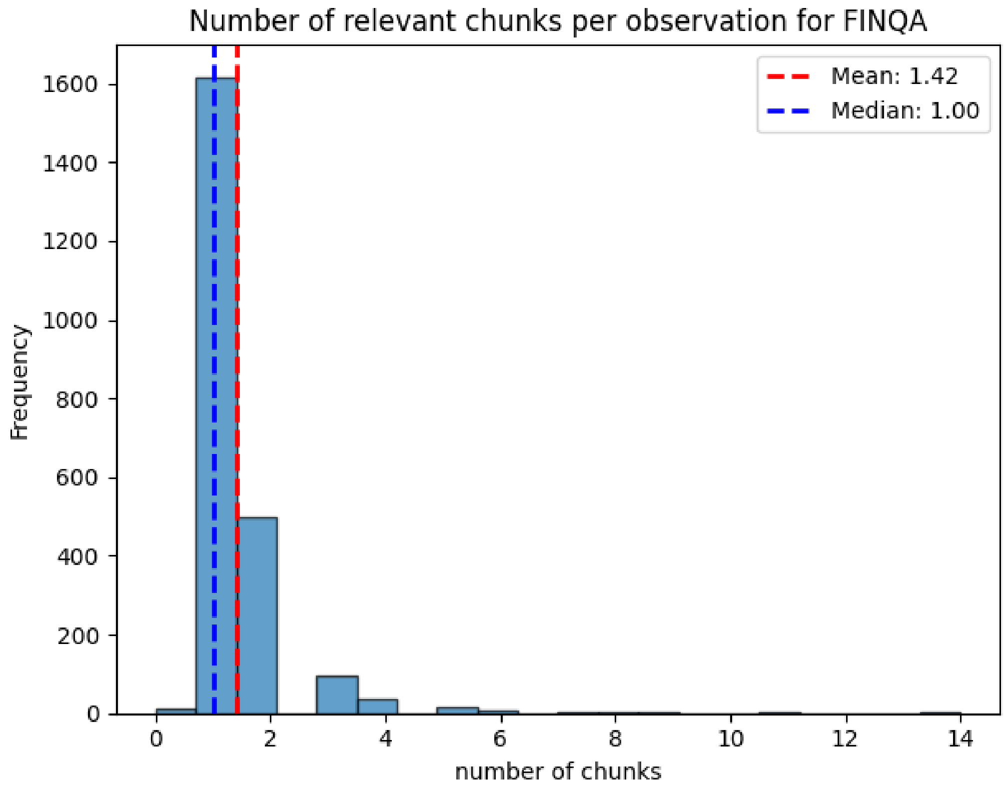
Data Distribution (FinQA)

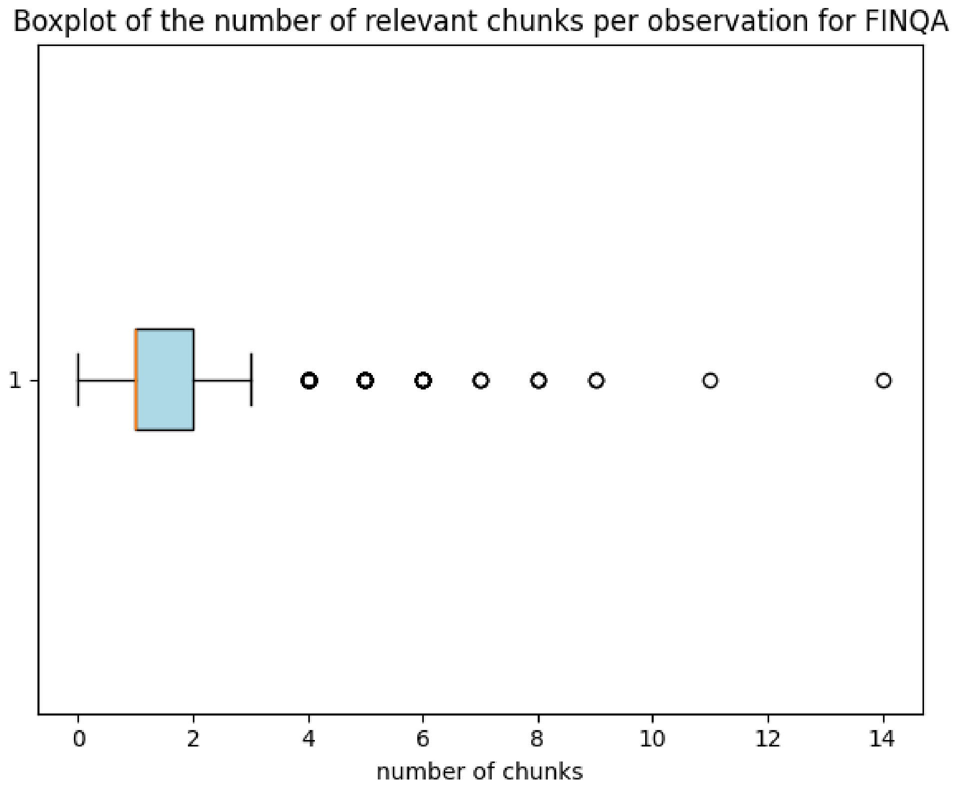
- Dataset: FinQA
- Task: QA
- Split: Test
- Nbr of samples: 2.2k





Data Distribution (FinQA)



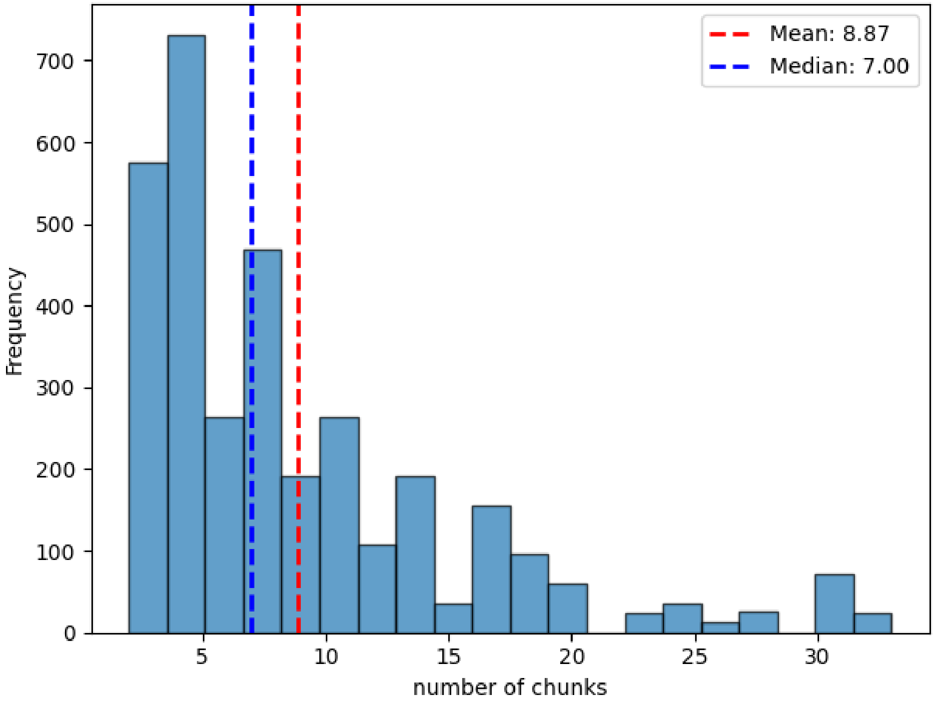


Choice of **K=5**

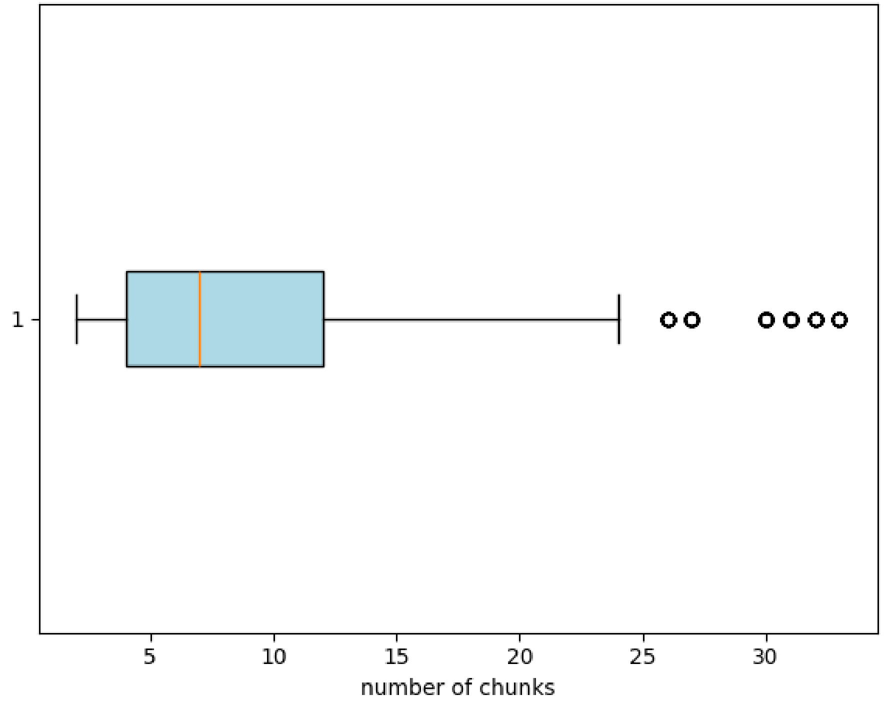
Data Distribution (TATQA)

- Dataset: TATQA
- Task: QA
- Split: Test
- Nbr of samples: 3.2k

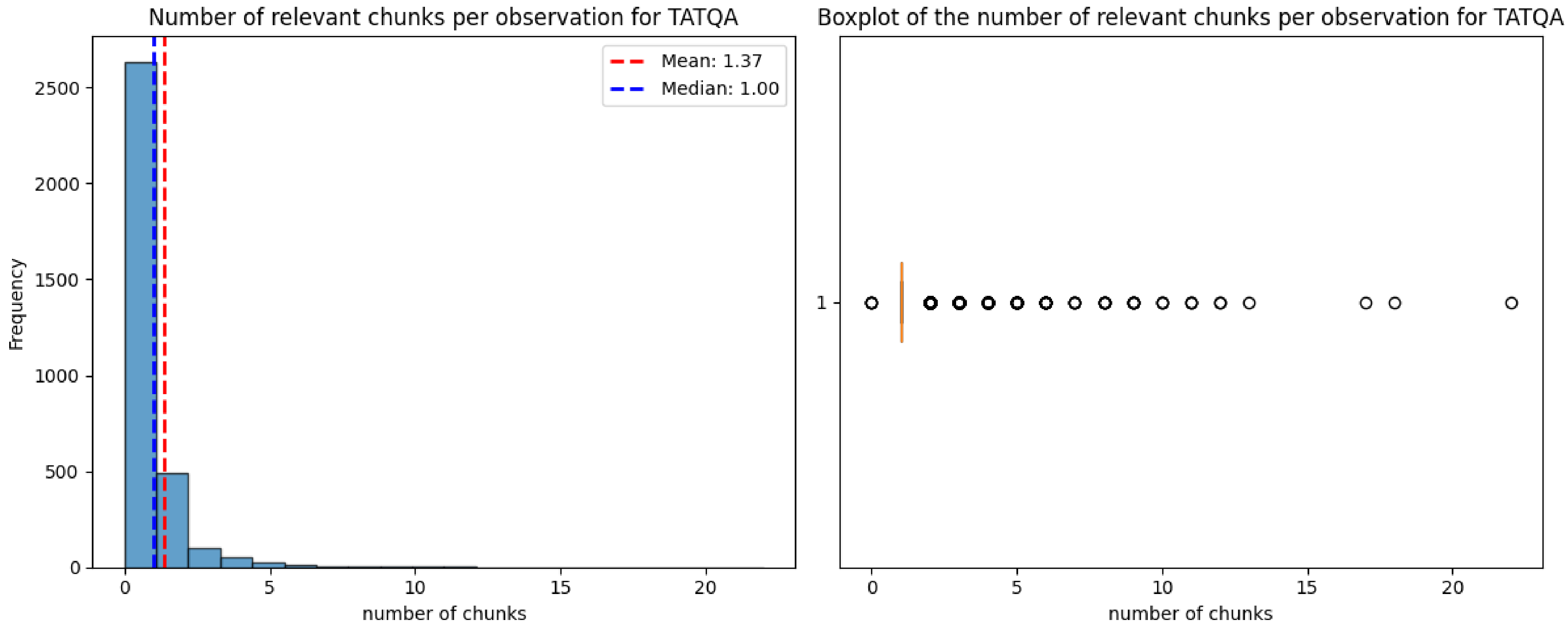
Number of chunks per observation for TATQA



Boxplot of the number of chunks per observation for TATQA



Data Distribution (TATQA)



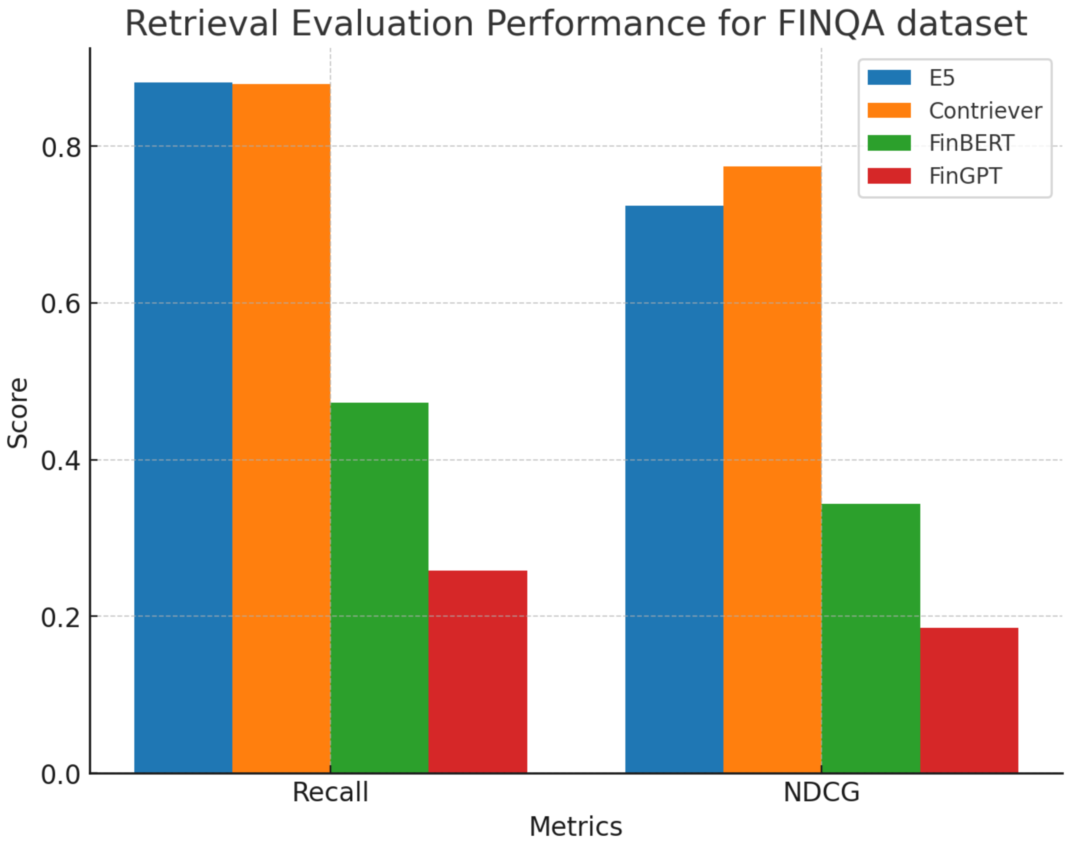
Choice of **K=5**

Models

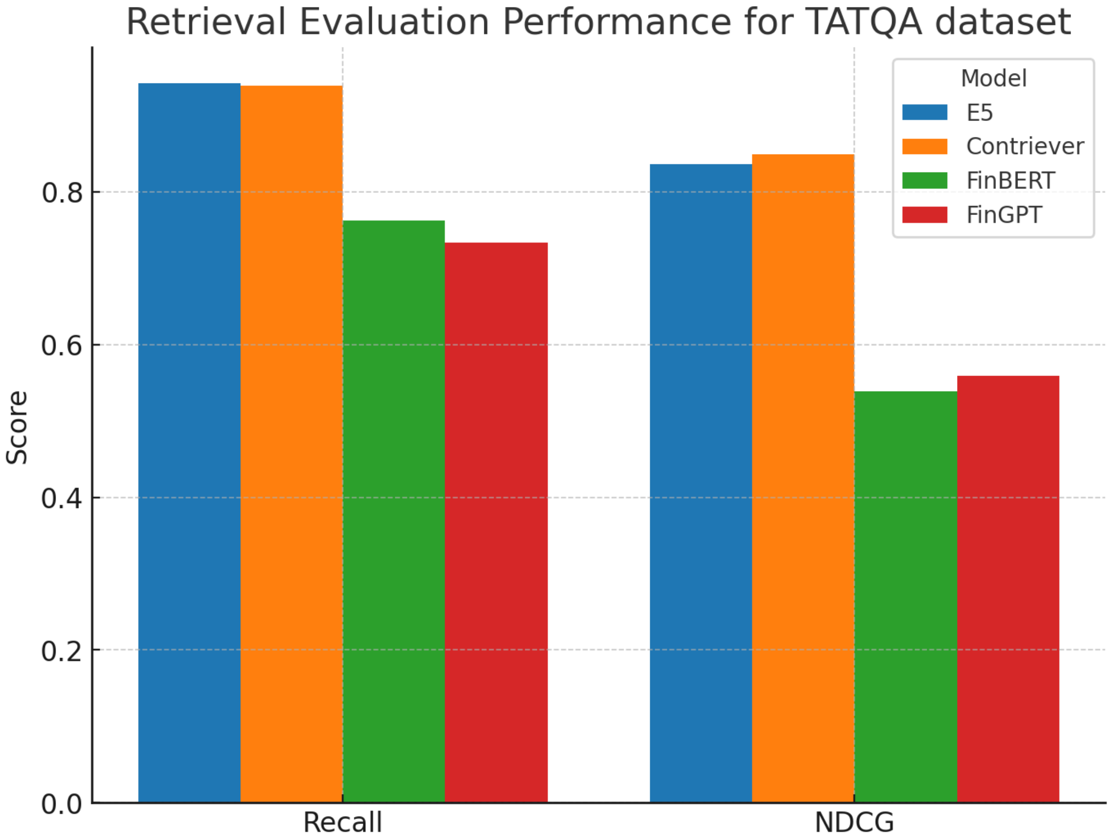
- E5 (Text Embeddings by Weakly-Supervised Contrastive Pre-training) https://arxiv.org/abs/2212.03533
  - Contrastive training: Weak supervision
  - Enable zero-shot learning
  - Positive pairs: top ranked by search engines
  - Negative pairs: low ranking
- Contriever: https://arxiv.org/abs/2112.09118
  - Contrastive learning: Unsupervised learning
  - Positive pairs: adjacent text segment
  - Negative pairs: random

- FinBert https://arxiv.org/abs/1908.10063
  - Pre-trained on large corpus on financial documents
  - Financial sentiment analysis, entity recognition, classification
- FinGPT https://arxiv.org/abs/2306.06031
  - Financial question answering, financial report summarization, retrieval-augmented generation

Results



Retrieval Evaluation Performance for FINQA dataset
E5
Contriever
FinBERT
FinGPT
Score
0.0
0.2
0.4
0.6
0.8
Recall
NDCG
Metrics



Retrieval Evaluation Performance for TATQA dataset
Model
E5
Contriever
FinBERT
FinGPT
Score
0.0
0.2
0.4
0.6
0.8
Recall
NDCG

Results

- E5: more params (560M) --> slow but more accurate

- Contriever (110M): Fast but less accurate

- FinBERT (110 M)

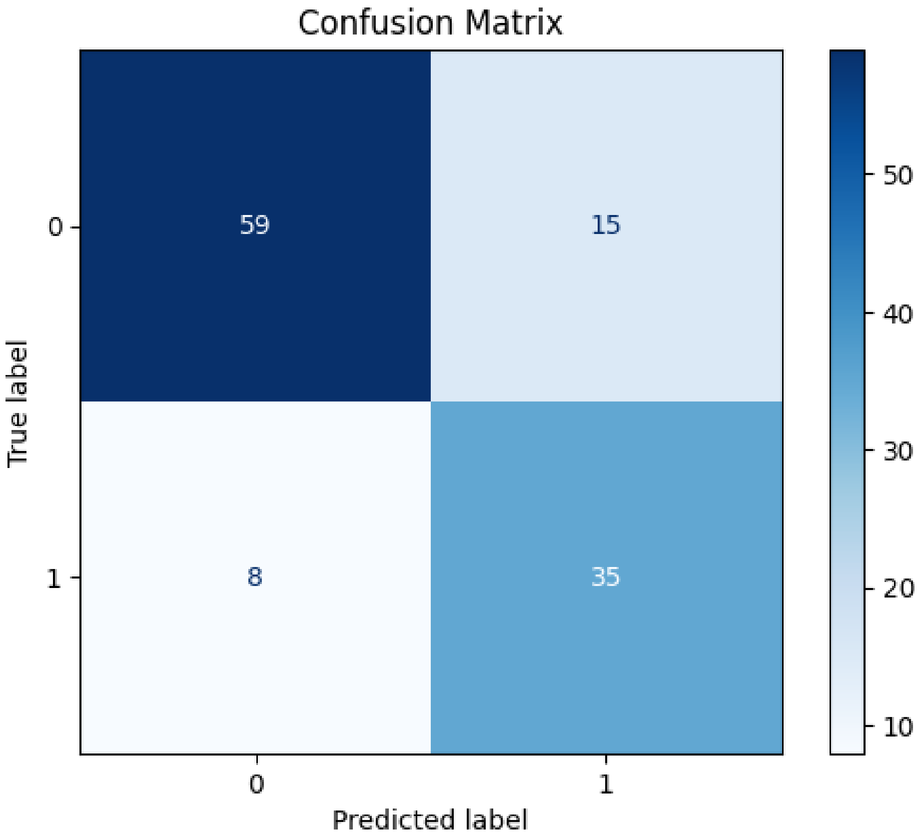
It has been specifically fine-tuned on financial texts for sentiment classification and language understanding, making it effective for retrieval-based QA

- FinGPT (7B) params

FinGPT, on the other hand, is a large generative model trained mainly for financial forecasting and language generation, not retrieval.

III. Classifier model results

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

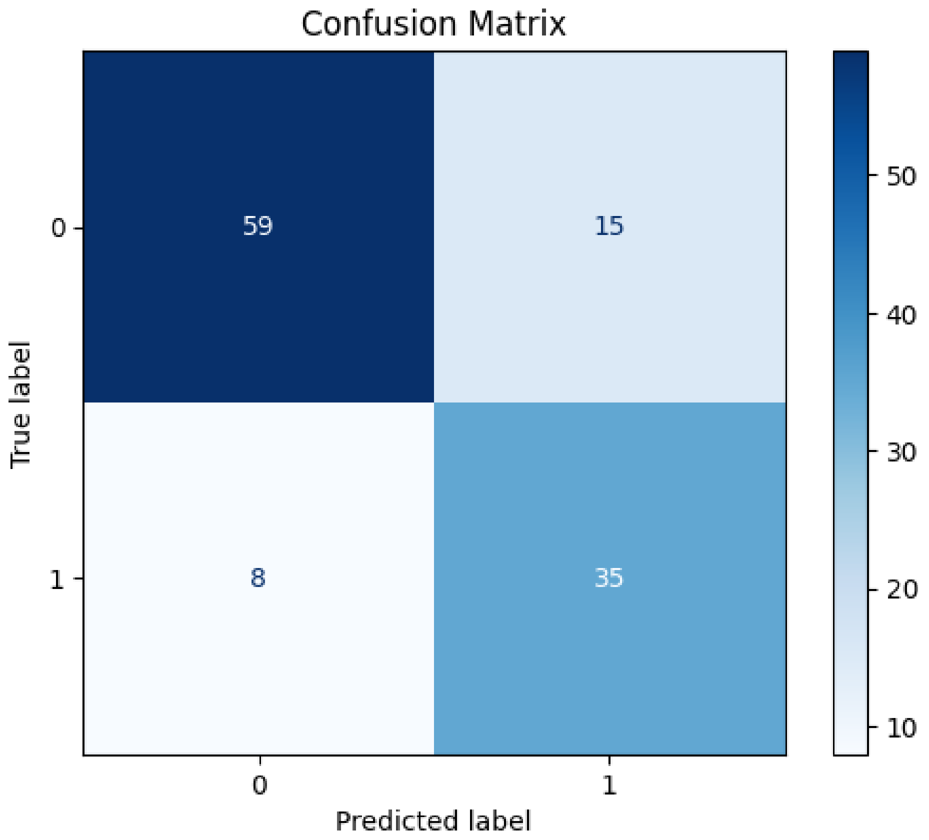
Thresholding [https://arxiv.org/pdf/2412.12148]

- Estimate density for metric and take the αth quantile $Q_{1-\alpha}$
    - Z-score $Z = \frac{X - \hat{\mu}}{\sigma}$
    - KDE

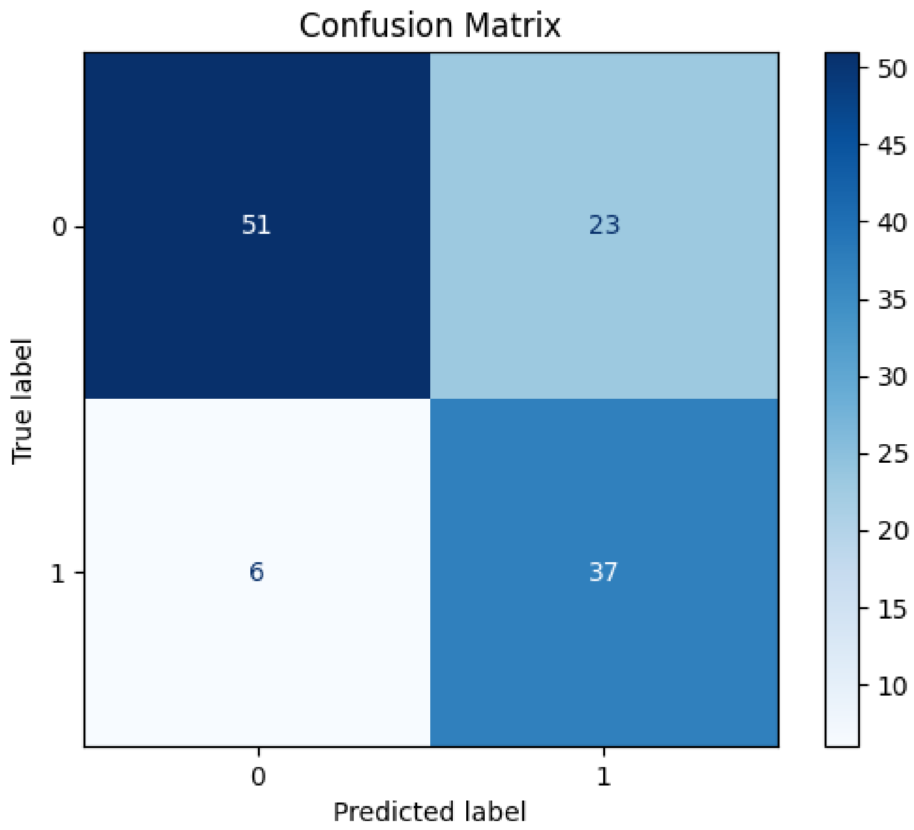
- Conformity scores $s_i = s(X_i, y_i) = \left|1 - \max_y \hat{p}(Y_i = y | X_i)\right|$
    - Train set, calibration set, evaluation set
    - Find the α-th quantile of calibration set $q_{1-\alpha} = ((1-\alpha)^{th}\ quantile((s_i)_i^n)$
    - On evaluation set $\hat{C} = 1 : \hat{p}(x_{new}) > 1 - q_{1-\alpha}$

    - Confidence interval $\hat{p}(Y_i = y_i | X_i) \pm q_{1-\alpha}$

Thresholding



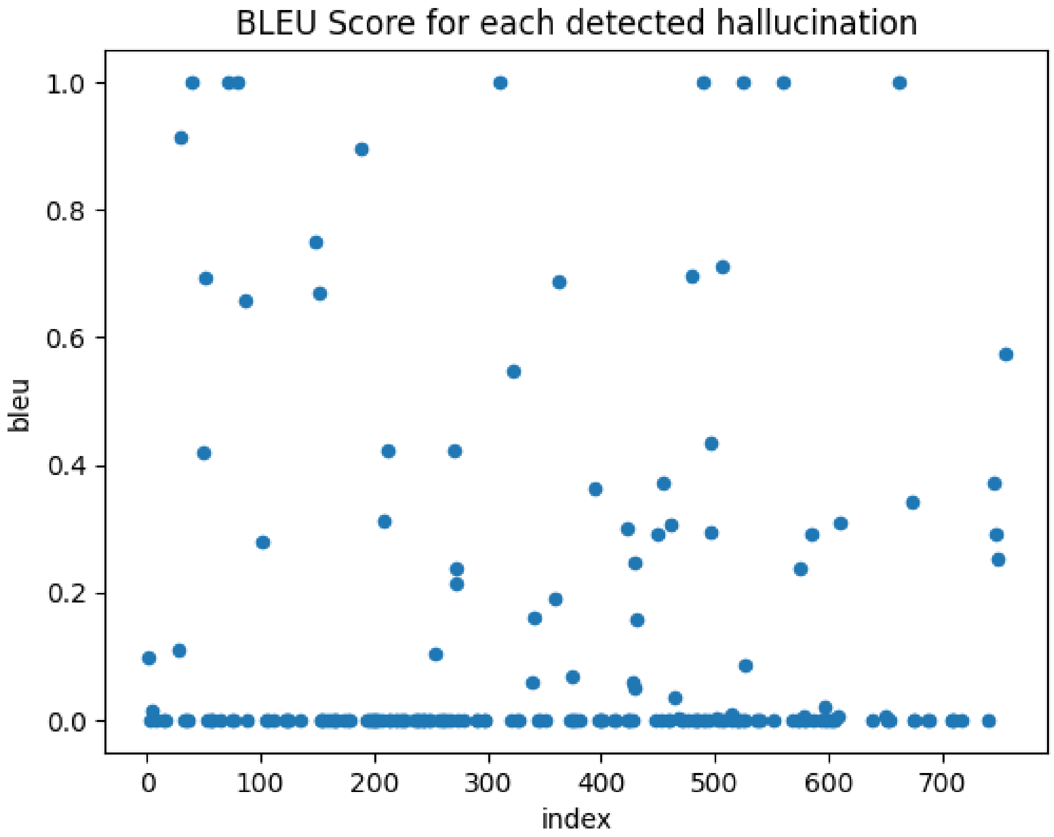
Classifier, $accuracy = 0.80$



Classifier w/ thresholding $\alpha = 0.1$, $accuracy = 0.75$

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.



ANNEX: 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

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