Open-Source RAG pipeline for the EIC

Tina J Jat
B.Sc. (Hons.) Data Science and Analytics
M. S. Ramaiah University of Applied Sciences

Supervisor : **Dr. Tapasi Ghosh**

Outline

- Objective
- Methodology
- Overview of prior work (AI4EIC Group)
- Schematic Representation of the pipeline
- Comparison of tools and frameworks
- Results
- Challenges and observations
- Conclusions and Future Work

Objective

- Develop an in-house RAG-based QA system using open-source models and frameworks.
- Comparison study with respect to proprietary model.
- To identify a cost-effective and performant alternatives, especially in resource-constrained or offline environments.

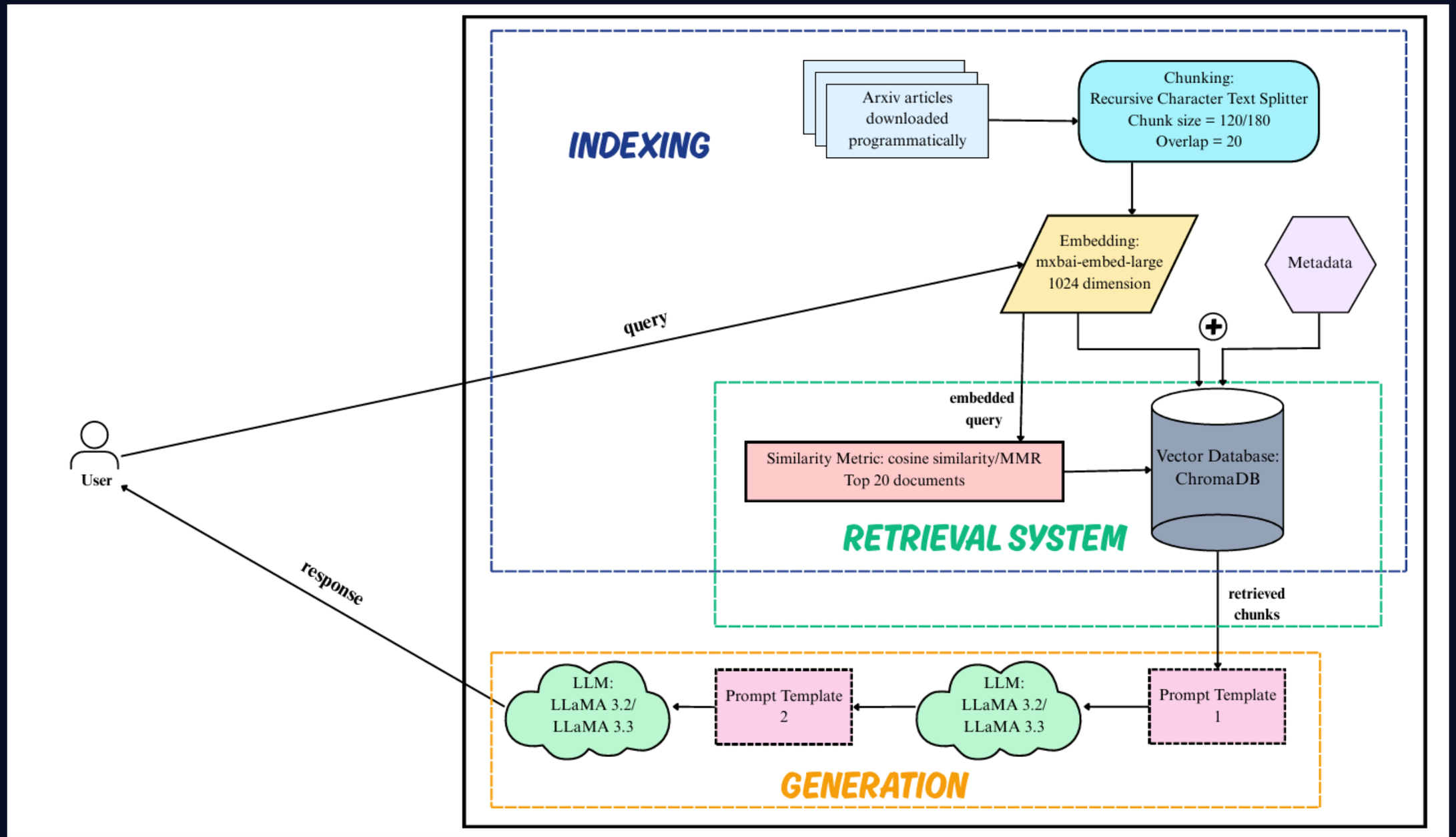
Methodology

1. Phase I: Understanding Prior Work and defining the problem statement
2. Phase II: Exploration of SOTA tools
3. Phase III: Building Knowledge Base
4. Phase IV: Implementation of QA System
5. Phase V: Evaluation and Result Analysis

Overview of the prior work (AI4EIC Group)

- RAG-based Q&A system been developed with the proprietary OpenAI models.
- Database: 178 EIC-related arxiv articles dated after 2021
- Embedding model: OpenAI's text-embedding-ada-002 (dimension – 1536)
- Vector DB: PineCone (API)
- Language Model used: GPT-3.5 (gpt-3.5-turbo-1106)

Schematic Representation of the pipeline



Comparison of Tools and Frameworks Used:

	Purpose	Prior work	Present work
Embedding Model	To convert the text into numerical representations called vector embeddings	OpenAI's text-embedding-ada-002 Dimension: 1536	Mixedbread AI's mxbai-embed-large Dimension: 1024
Vector Database	For storing the context data in the form of vector embeddings.	PineCone (Cloud)	ChromaDB (in-house Database)
Language Model	Generate output answer for a given input prompt	GPT-3.5 (gpt-3.5-turbo-1106)	Quantized LLaMA 3.2 (3B parameters) and LLaMA 3.3 (70B parameters)

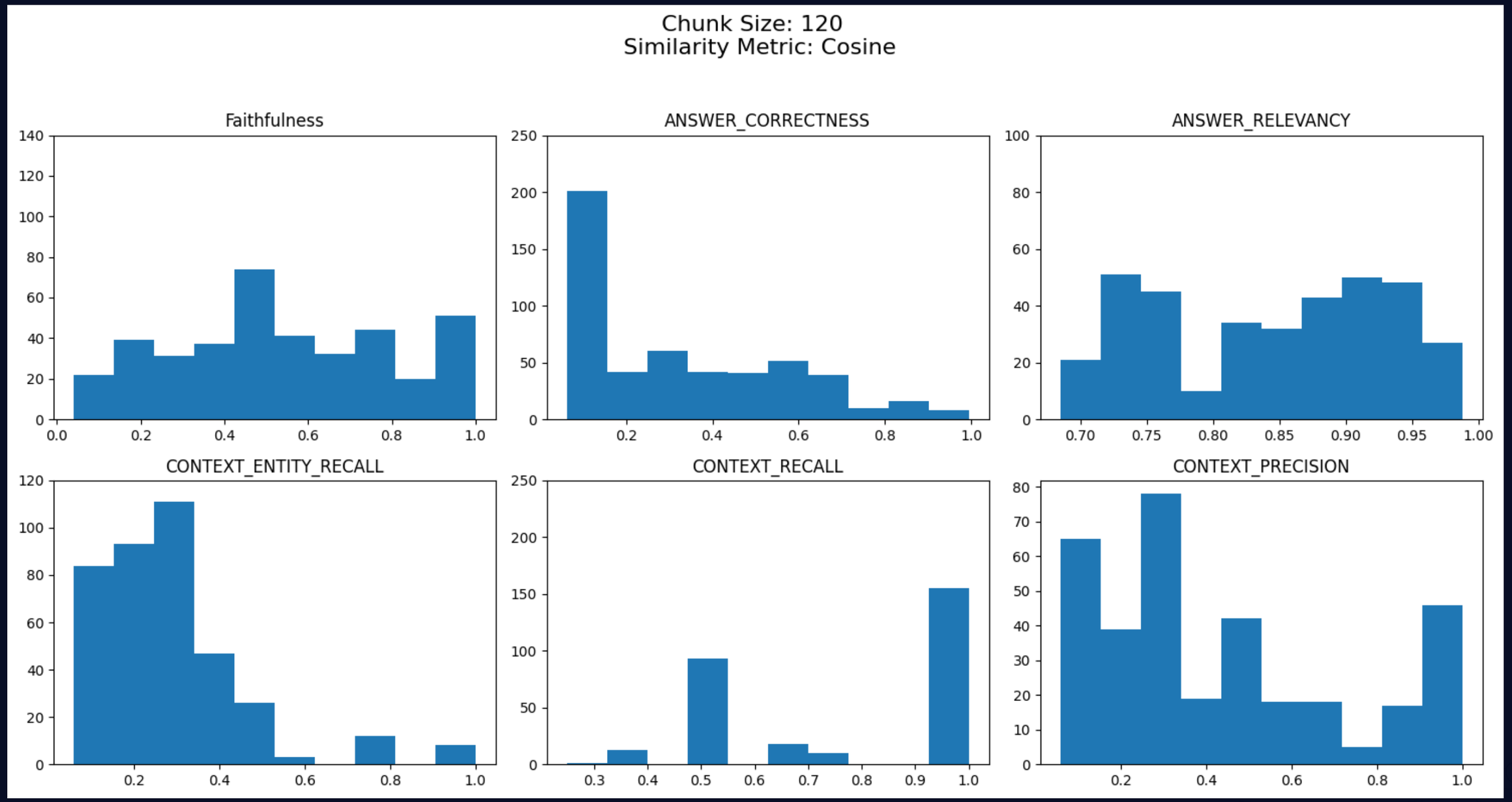
Context and strategy Details:

- Vector Database: 178 EIC-related arxiv articles in **PDF** format.
- Benchmark dataset: To validate the performance of the application
  - AI Generated and validated by domain expert
- Chunk size: 120 and 180
- Retrieval Strategy: Cosine Similarity and Maximum Marginal Relevance (MMR)
- LLMs choice: Quantized LLaMA 3.2 (3B parameters) and LLaMA 3.3 (70B parameters)
- Evaluation Metrics: Faithfulness, Answer Relevancy, Answer Correctness, Context Entity Recall, Context Recall, Context Precision.

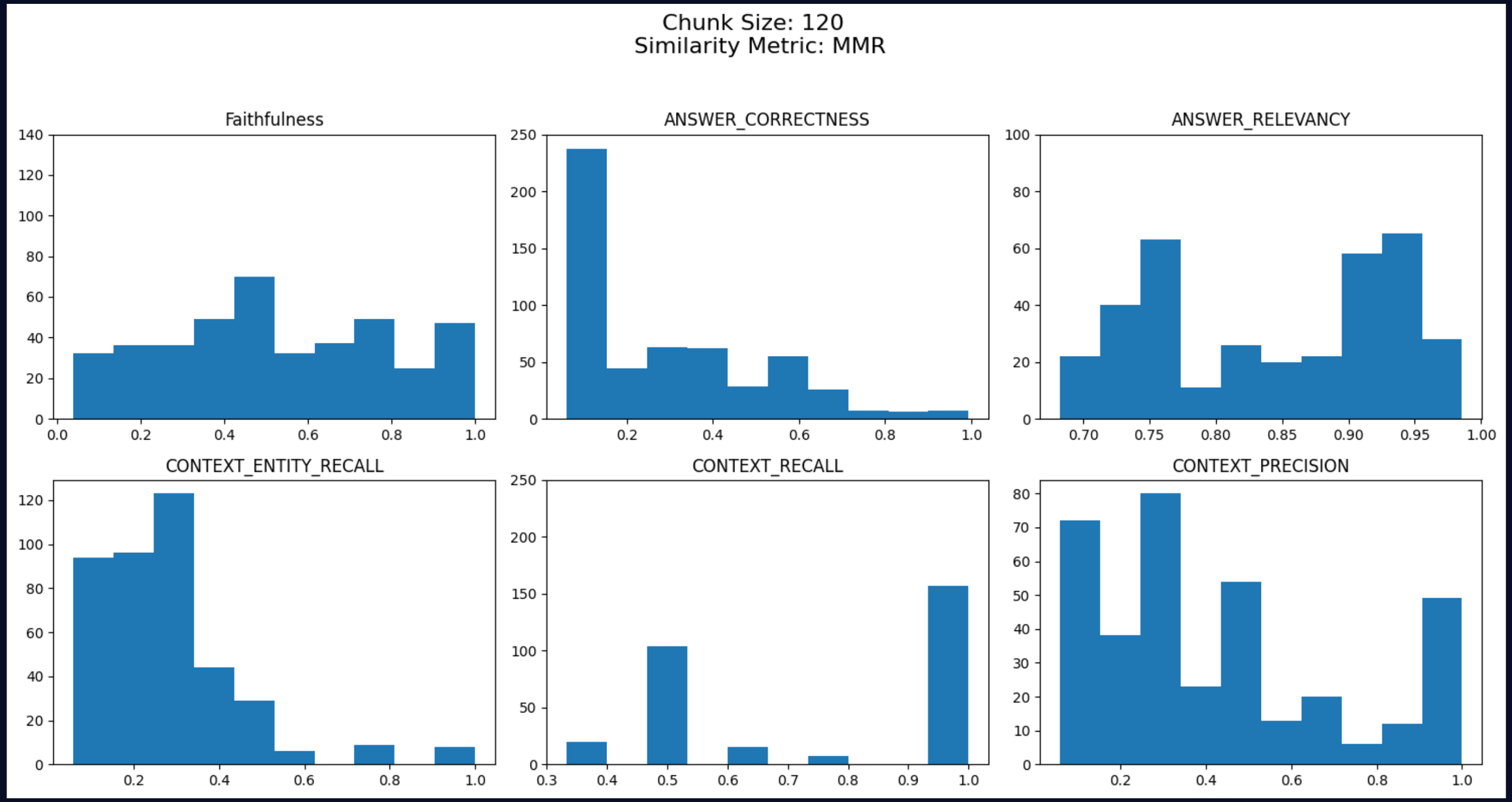
Results: Latency Comparison

Statistic	LLaMA 3.2	LLaMA 3.3
Mean	14.30	226.46
Standard Deviation	9.36	75.54
Minimum	2.95	90.91
25% (Q1)	8.30	175.14
Median	11.33	215.88
75% (Q3)	17.31	266.66
Maximum	59.78	568.20

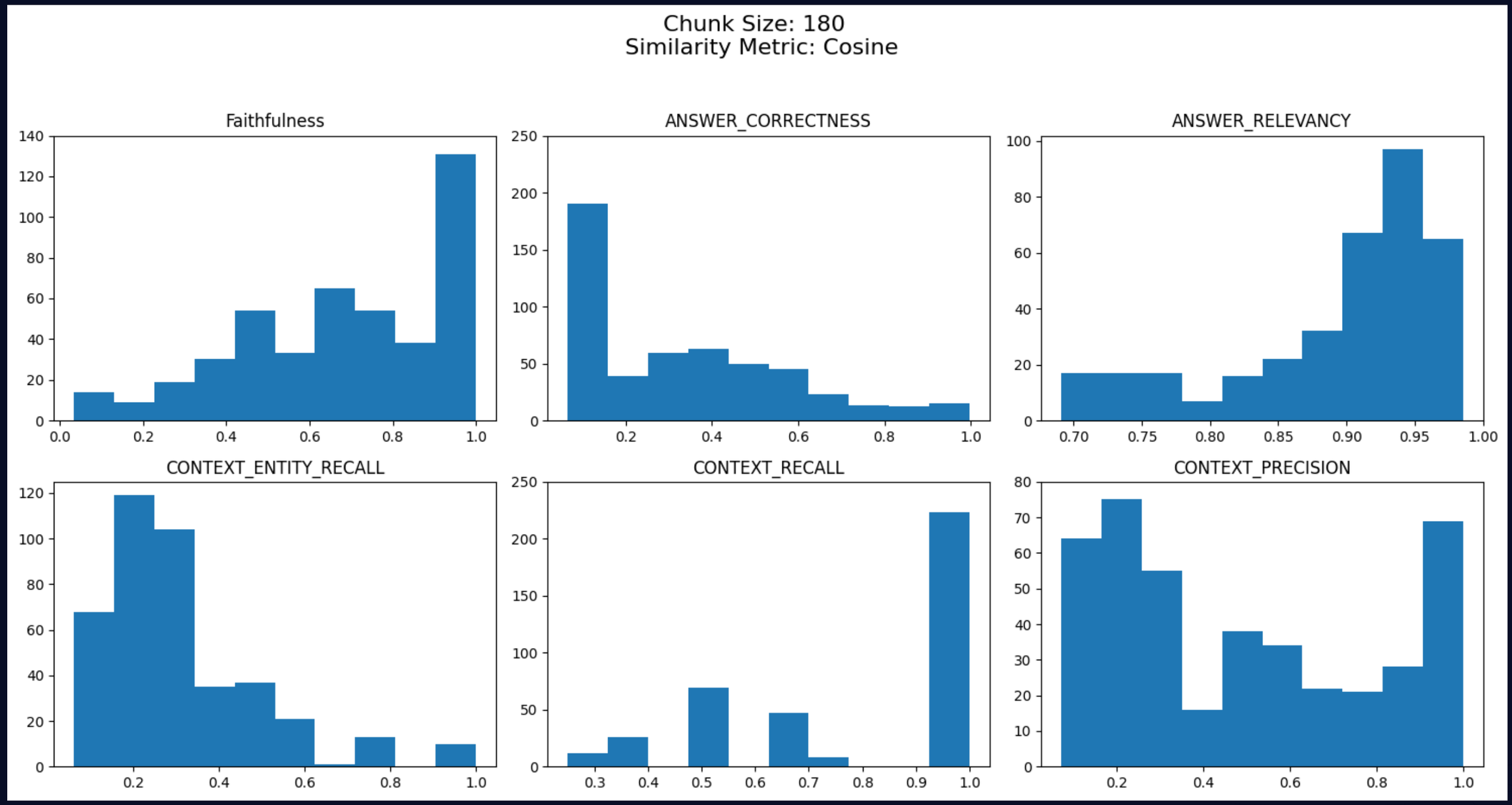
Results : RAGAS Evaluation Scores



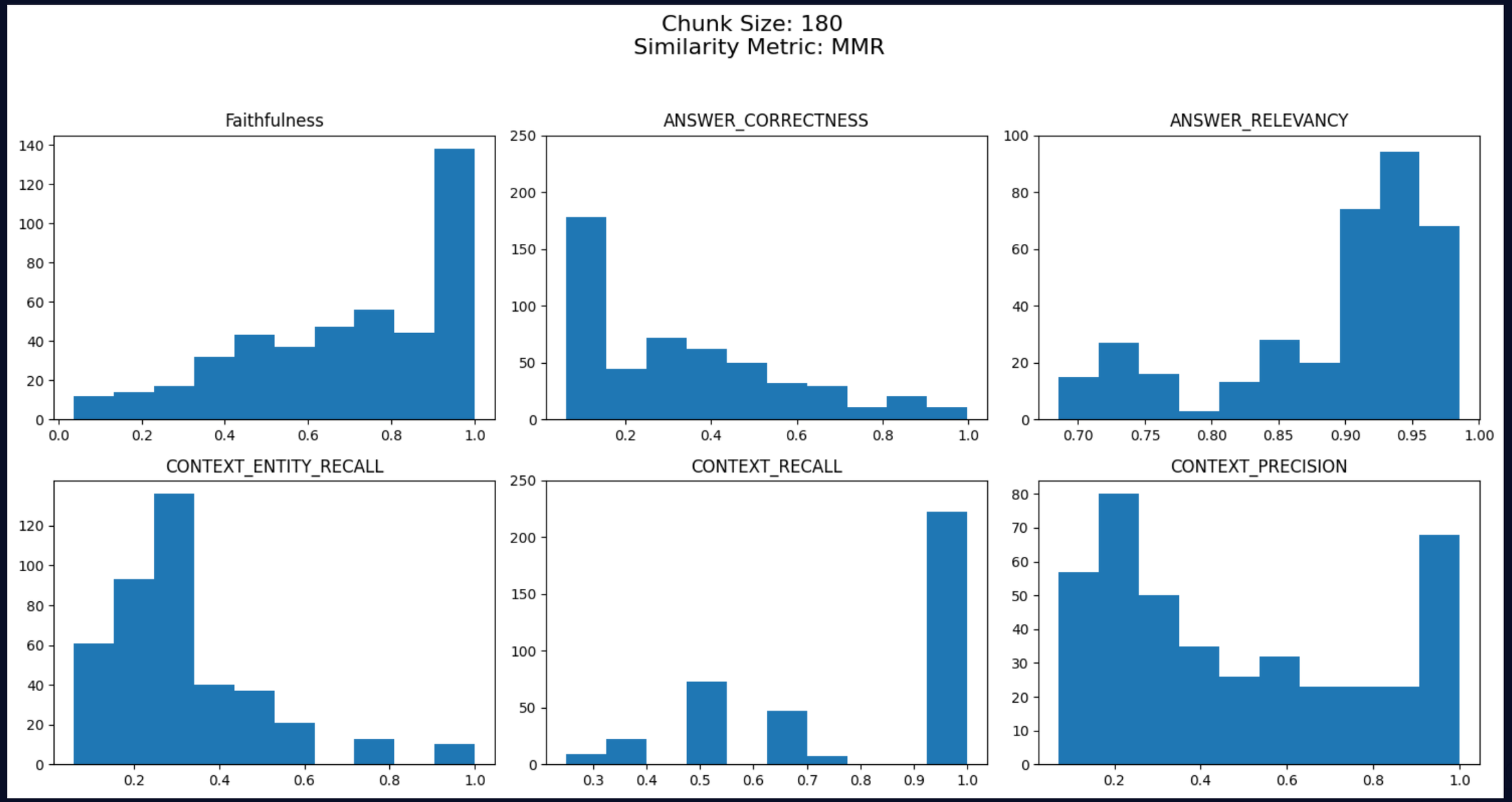
Results : RAGAS Evaluation Scores (Cont.)



Results : RAGAS Evaluation Scores (Cont.)



Results : RAGAS Evaluation Scores (Cont.)



Challenges

- The system occasionally generates answers based on the model's intrinsic parametric knowledge rather than grounding them solely in the retrieved database context.
- For some questions, the generated response contains repetition of same answer multiple times. This is also reflected with higher latency.

Observations

- The latency of context retrieval varies depending on the chosen LLM model.
- The system experiences a slight reduction in answer correctness and entity recall but performs well in terms of faithfulness and relevancy, notably with a 180-character chunk size and MMR retrieval strategy.

Conclusions

- RAG-based QA system built entirely on open-source tools can offer competitive performance for EIC-related document understanding.
- The prior work employed proprietary OpenAI GPT-3.5 models with over 100 billion parameters.
- The current system leverages significantly smaller models (3B parameters) that reduce memory footprint and latency.
- These findings highlight a practical trade-off: large-scale proprietary models offer superior accuracy, while lightweight open models can provide cost-effective and performant alternatives.

Future Work

- Extend the database with multimodal formats like images, ppt, etc.
- Build an Agentic RAG

Acknowledgements:

Dr. Karthik Suresh

Dr. Cristiano Fanelli

Thank You

Backup Slides

Results – Latency

Statistic	120	180
Mean	0.11	0.11
Std Dev	0.02	0.02
Min	0.08	0.07
25% (Q1)	0.10	0.10
Median	0.11	0.11
75% (Q3)	0.12	0.12
Max	0.25	0.14

(a) Chunk Size: 120 vs 180

Statistic	Cosine	MMR
Mean	0.11	0.12
Std Dev	0.02	0.01
Min	0.07	0.08
25% (Q1)	0.10	0.11
Median	0.10	0.12
75% (Q3)	0.11	0.12
Max	0.25	0.14

(b) Similarity Metric: Cosine vs MMR

Statistic	LLaMA 3.2	LLaMA 3.3
Mean	14.3	226.46
Std Dev	9.36	75.54
Min	2.95	90.91
25% (Q1)	8.30	175.14
Median	11.33	215.88
75% (Q3)	17.31	266.66
Max	59.78	568.20

(c) Language Model: LLaMA 3.2 vs LLaMA 3.3

Table 1: Descriptive statistics of latency across different configurations