RAG-inspired Open-source based Q&A system for scholarly articles in EIC

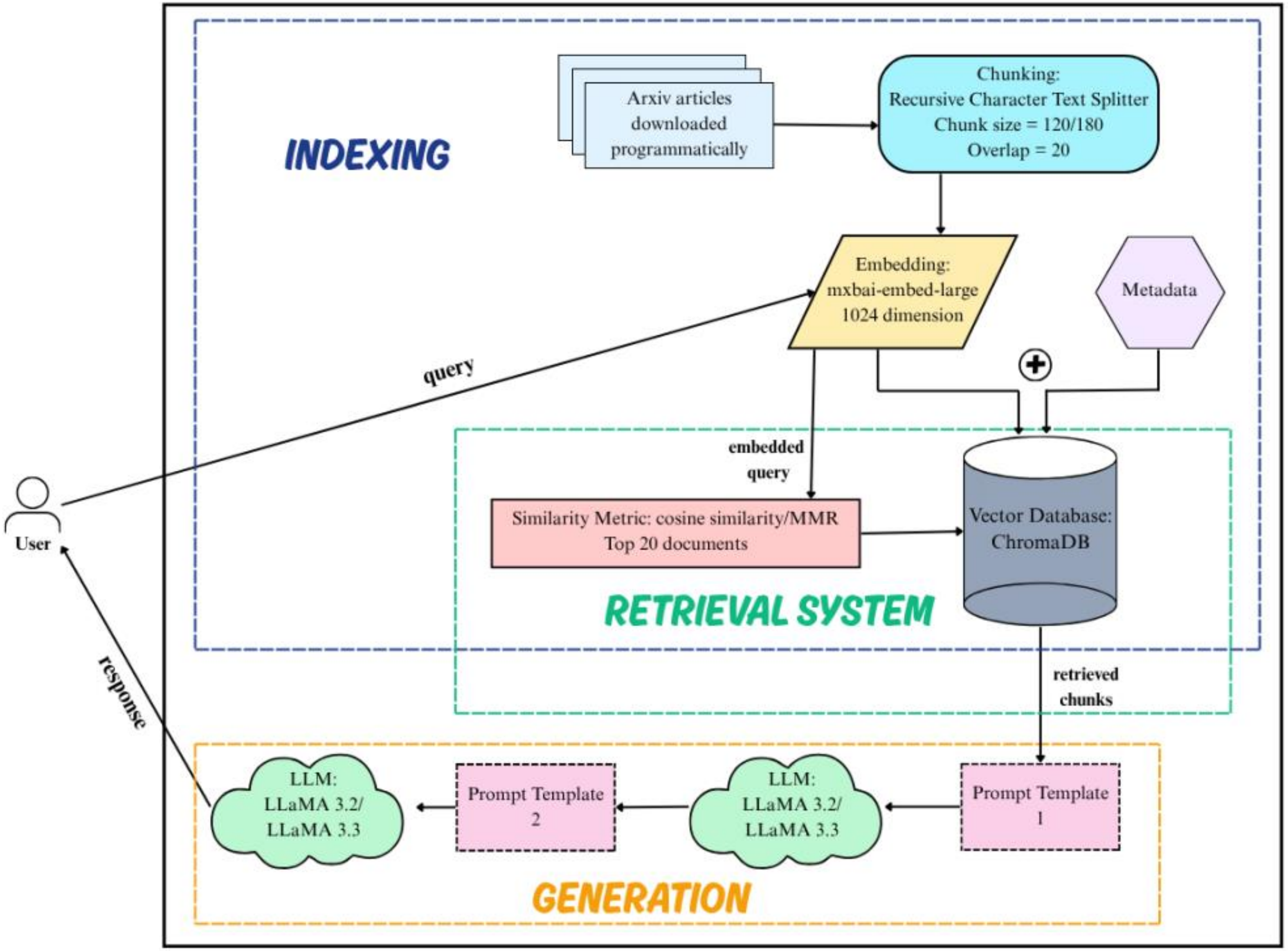
Tapasi Ghosh

Department of Data Sciences and Analytics
Ramaiah University of Applied Sciences, India

Objectives:

- Build an **in-house knowledge base** of EIC-related scholarly articles.
- Implement Retrieval Augmented Generation(RAG) to improve the factual accuracy of a Q&A system for EIC.
- Extend the previous work built on **proprietary model** such as OpenAI to **open-source models/frameworks**.

Flowchart of the end-to-end RAG system:



Indexing : Vectorized Database

- **In-house Knowledge base**: Around 200 pdf articles from arXiv
- **Chunking strategy**: Recursive Character Text Splitter model
- **Chink size**: 120 and 180 tokens with an overlap of 20 characters
- **Metadata**: indexed with arXiv ID, title, categories, primary categories, authors and publication date.
- **Embedding model**: 1024-dim vector representations; ***all-MiniLM*** and ***mxbai-embed-large*** models.
- **Database**: ***Pinecone*** (cloud-based) and ***ChromaDB*** (in-house) were explored

Retrieval:

- **Query embedding** : Same ***mxbai-embed-large*** model to convert user queries with 1024-dim vector
    - Hosted locally via Ollama.
    - This ensures semantic consistency between database embeddings and the user queries.
- **Vector Search**: The top 20 most relevant chunks using **Cosine Similarity or Maximum Marginal Relevance (MMR)** methods.

Answer Generation and Tracing:

- **Prompt Stuffing**: Retrieved context is combined with user query

- **LLM processing**: The ***Llama 3.2*** and ***Llama 3.3*** models, deployed locally via Ollama

    - *Llama3.2*: A quantized model with 3.21B parameters
    - *Llama3.3* : A quantized model with 70B parameters with 1,28,000-token context window.

- **Answer generation**: through a two-stage LLM processing mechanism.

- **Pipeline management and Tracing**:
    - ***LangChain***: Manages the flow between the query embedding, retrieval and answer generation

    - ***LangSmith***: Platform to trace the intermediate steps of the RAG-pipeline, essential for debugging

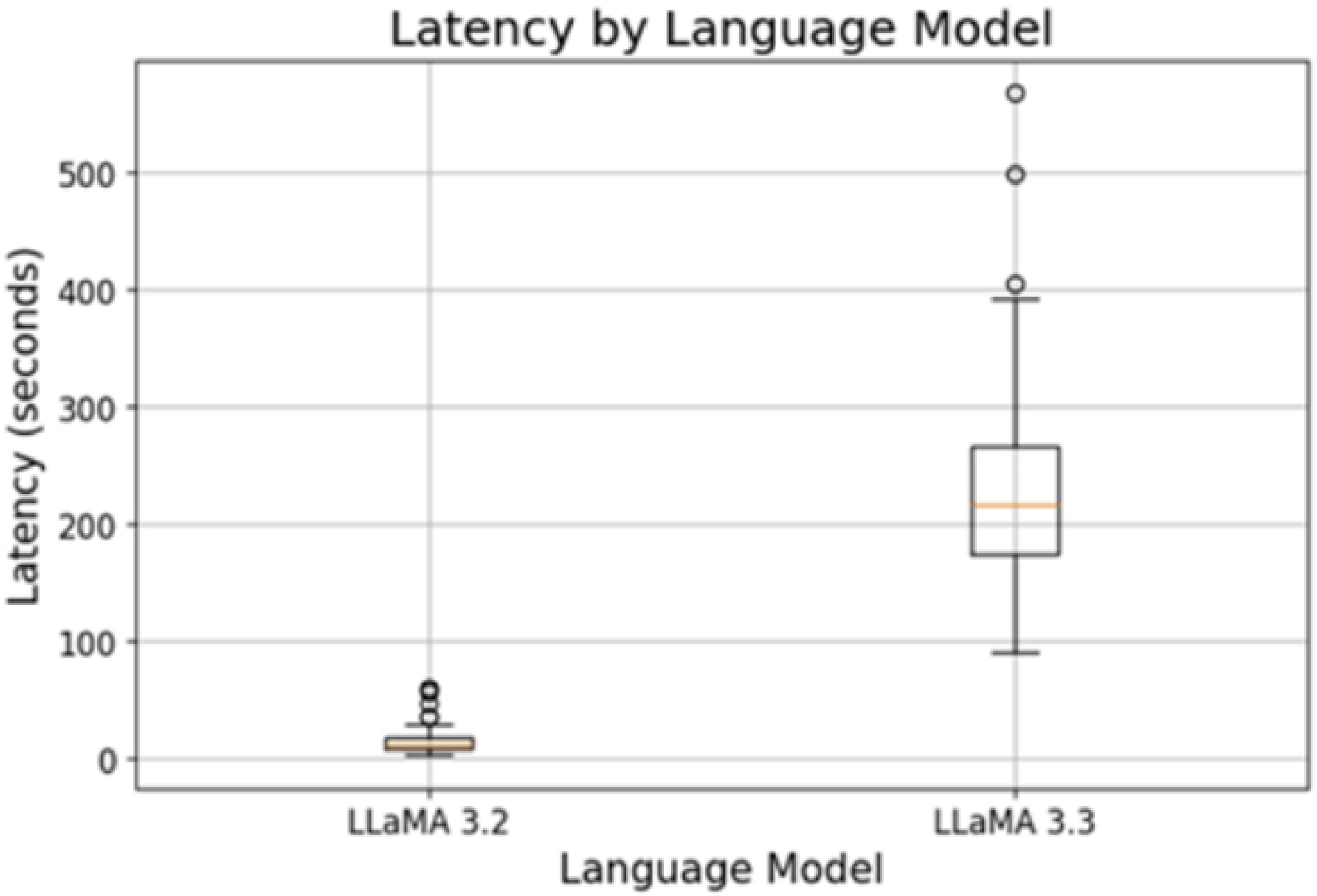
Evaluation: OpenAI RAGAS framework

- **Benchmark dataset**:
  - GPT4.0 to create a set of Q&A pairs from EIC-related arXiv articles on theory, simulation, hardware etc.
  - Scrutinized by domain expert
  - Each question in the dataset is explicitly
    - linked to a defined number of "claims"
    - corresponding answer that specifies individual claim and an ideal response
    - a comprehensive overall response
- **Retrieval Quality evaluation metrics:**
    - Context Entity Recall, Context Precision and Context Recall,
- **Generated answer evaluation metrics:**
    - Answer Relevancy, Answer Correctness and Faithfulness

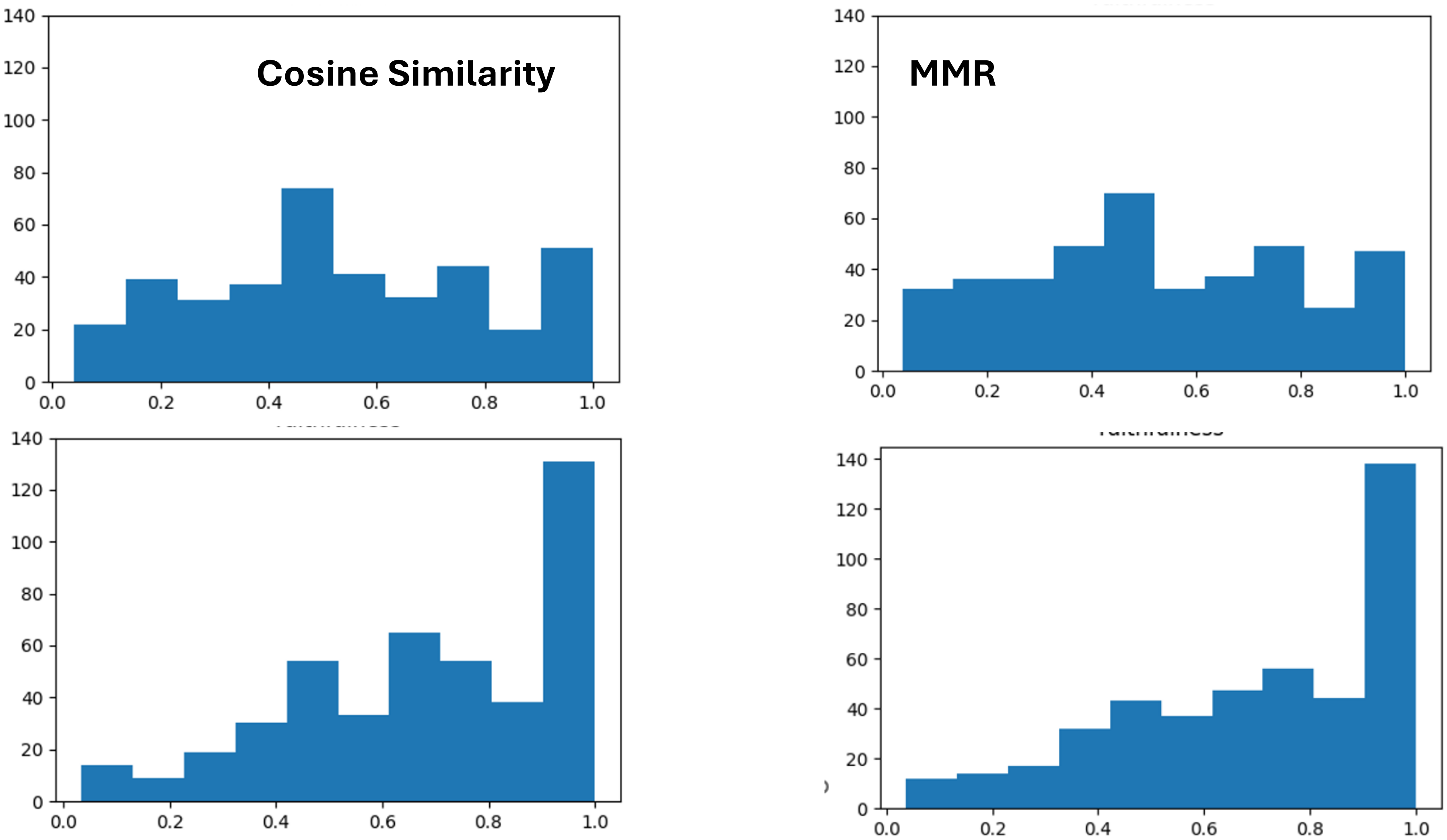
Evaluation Metrics:

- **Retrieval:**
  - **Context Precision**: Proportion of the retrieved context chunks that are relevant to user query
  - **Context Recall**: Proportion of the context that is being supported by the ground truth answer
  - **Context Entity Recall**: Determines whether the entities in the ground truth answer are successfully recalled within the retrieved context
- **Generation :**
  - **Answer Correctness**: Factual alignment and factual similarity of the generated answer w.r.t. ground truth
  - **Answer Relevancy**: Semantic alignment b/w the generated answer and the query
  - **Faithfulness**: Examines whether the generated answer is factually consistent with the retrieved context calculated based on the claims.
    - Identify instances of hallucinations or unsupported claims

Results: Latency in sec (26 Gb GPU)



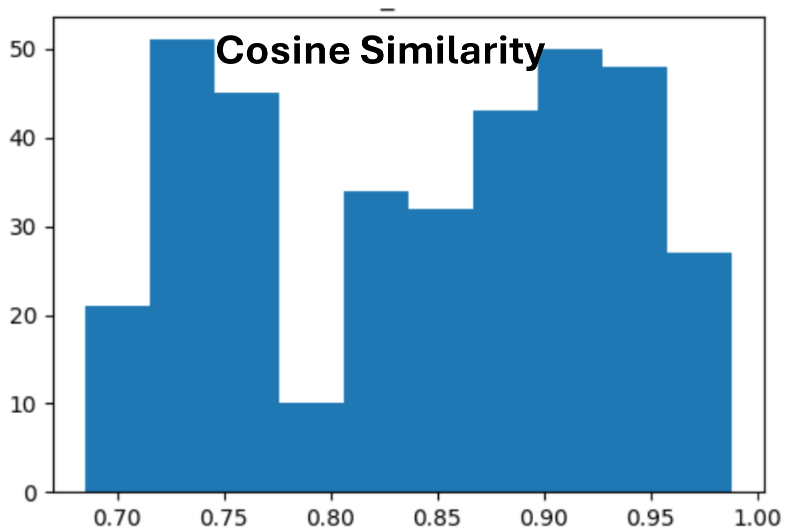
Performance: Faithfulness

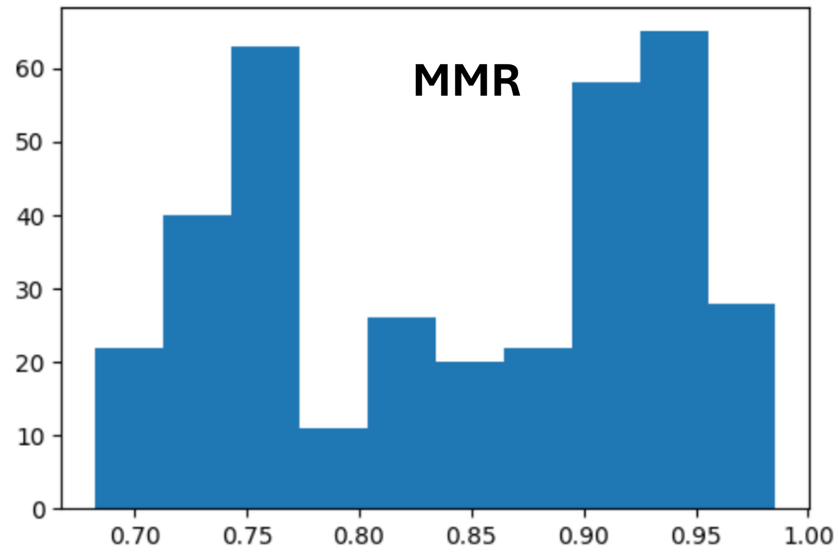


Chunk Size : 120

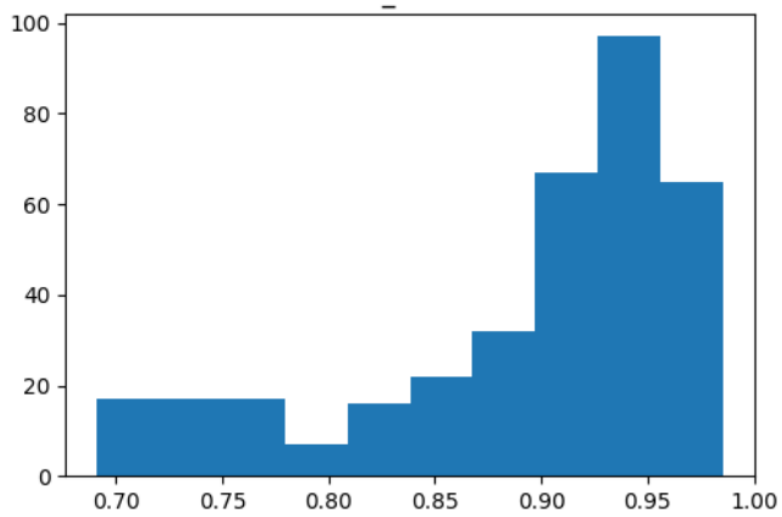
Chunk Size : 180

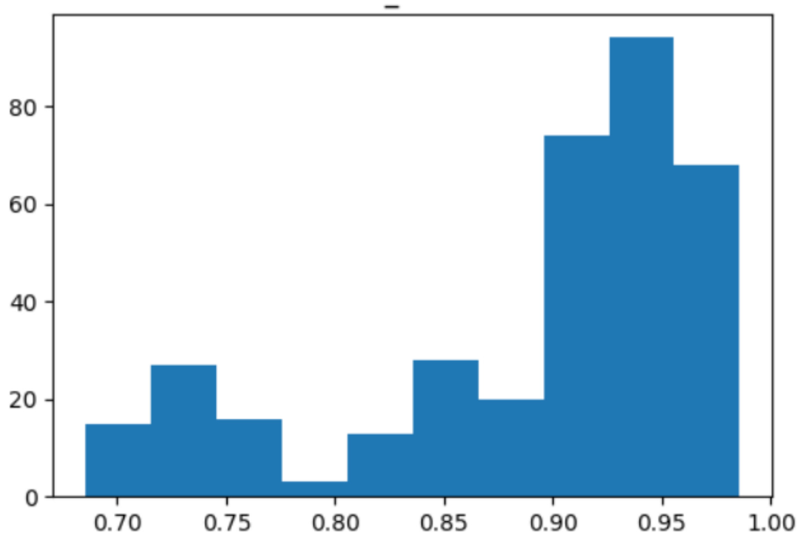
Answer Relevancy:





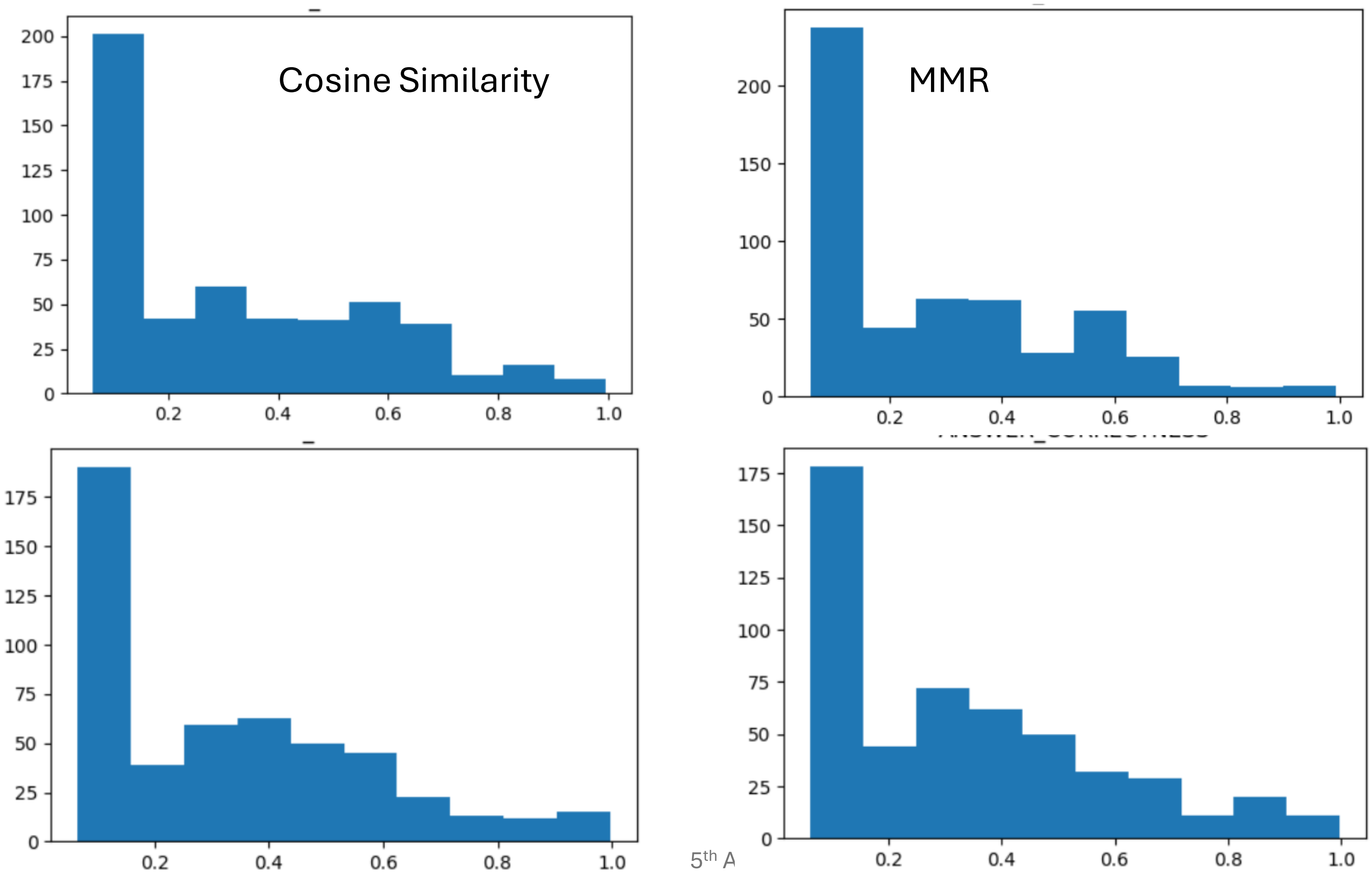
Chunk Size : 120





Chunk Size : 180

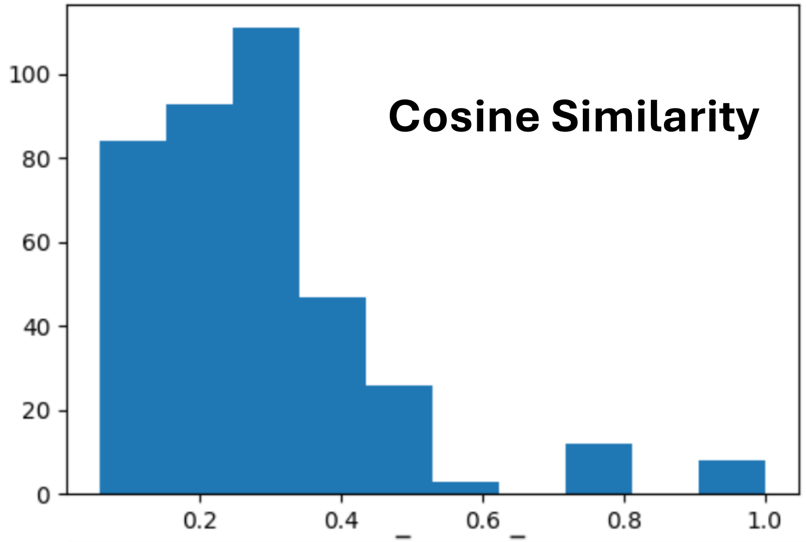
Answer Correctness:

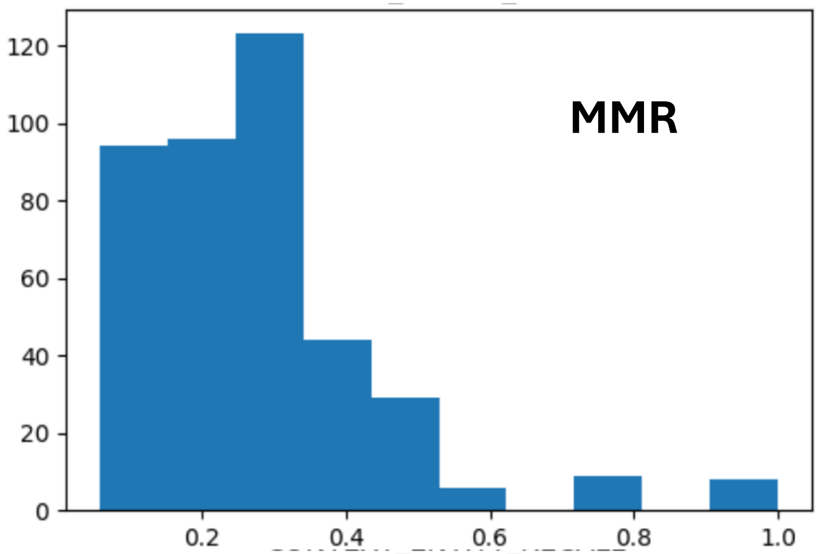


Chunk Size : 120

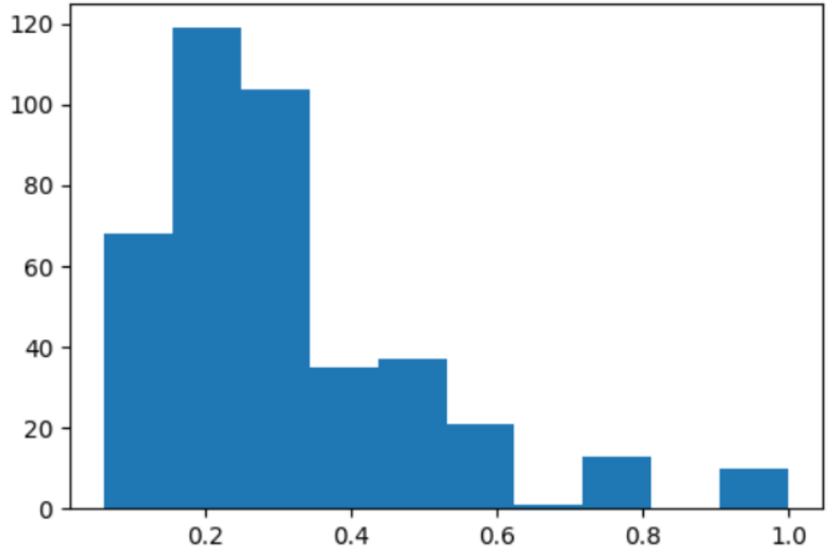
Chunk Size : 180

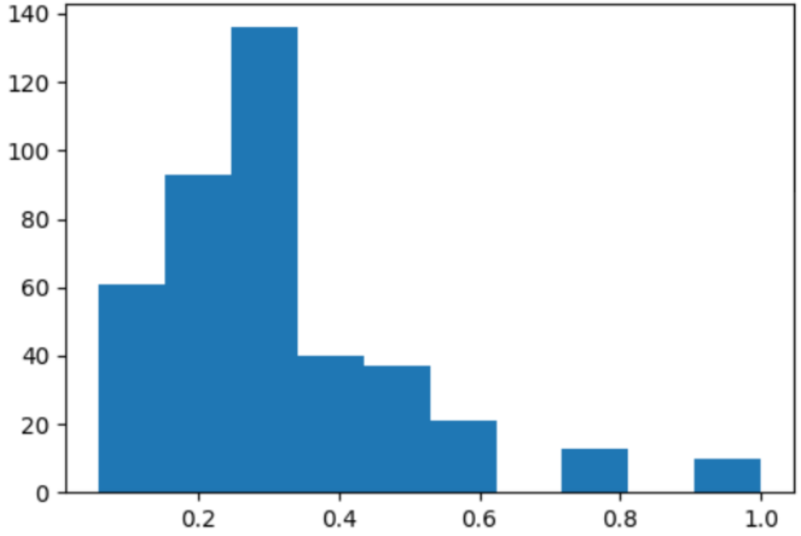
Context Entity Recall:





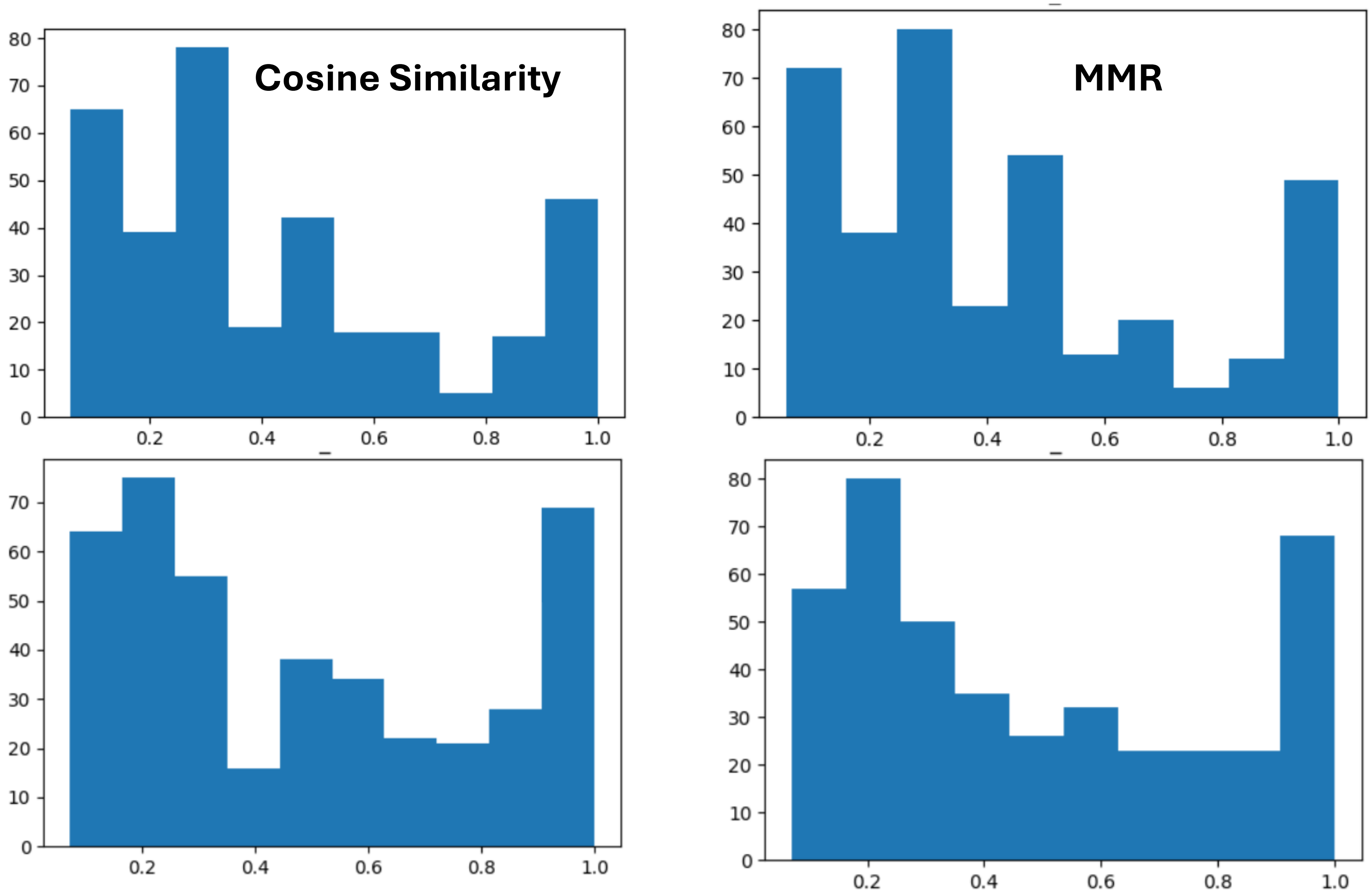
Chunk Size : 120





Chunk Size : 180

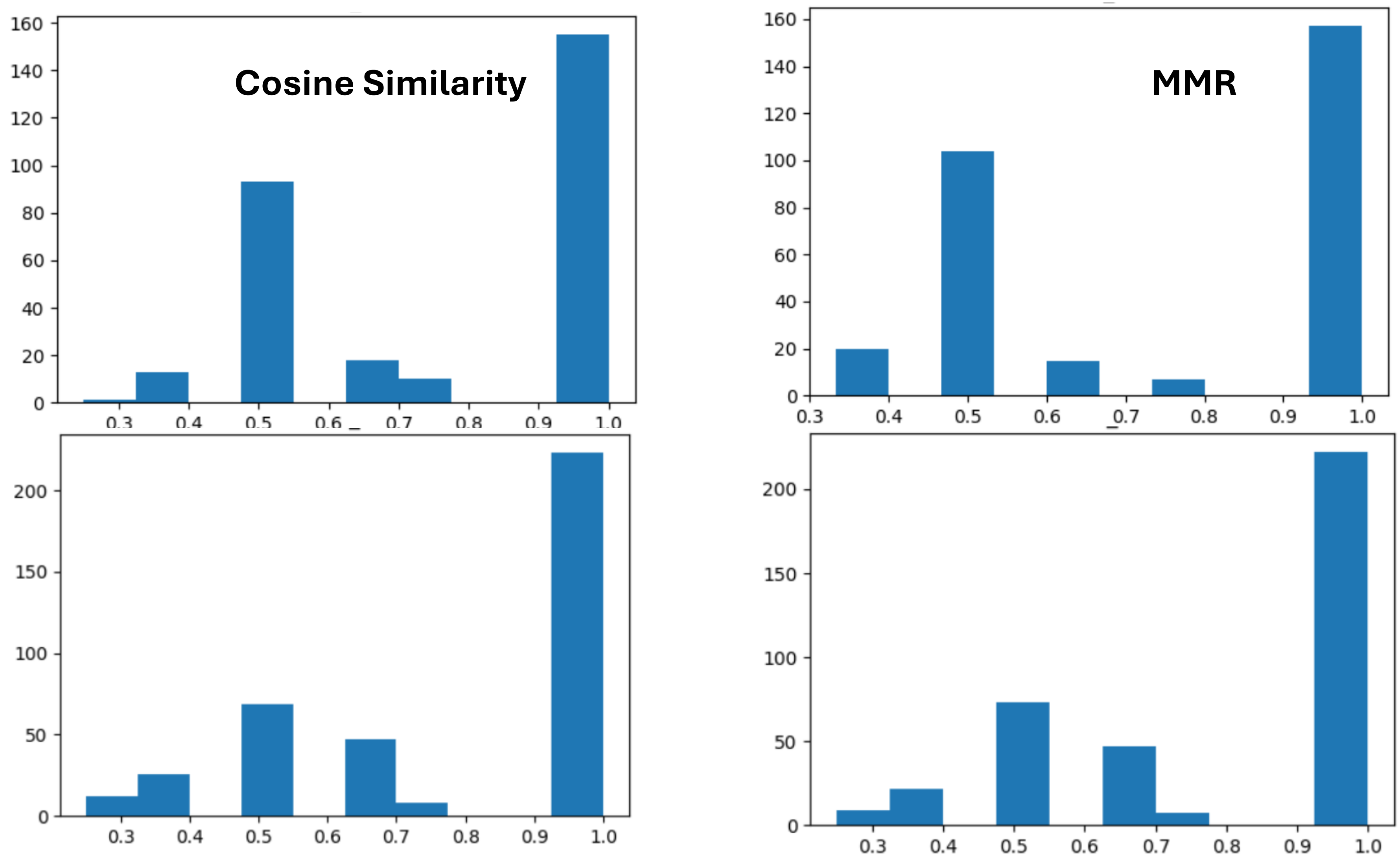
Context Precision:



Chunk Size : 120

Chunk Size : 180

Context Recall:



Chunk Size : 120

Chunk Size : 180

Result Summary:

- **Faithfulness** and **Answer Relevancy** :
  - 180 chunk size with Cosine Similarity and MMR combinations demonstrates a positive trend achieving scores around 90%
  - 120 chunk size showed a slightly bimodal distribution and higher variability in scores across the range.
- **Answer Correctness, Context Precision** and **Context Entity Recall** : All configurations demonstrated poor performance, with most answers receiving low correctness scores.
- **Context Recall**: Robust performance was evident across all configurations
- **Insight**: Increasing chunk size contributes to more factually grounded responses, regardless of the similarity metric.

Conclusions and Future outlook:

- **Conclusion**:
  - A RAG-based QA system for for EIC-related articles built entirely on open-source tools offers competitive performance and practical trade-off **compared to large proprietary model.**
  - A smaller model (~3B parameters) that reduces memory footprint and latency but offers cost effective and performant alternatives.
- **Challenges**:
  - For some questions, the generated response contains repetition of same answer multiple times. This is also reflected with higher latency.
  - Instances of hallucination observed
- **Outlook**:
  - Extending the knowledge base into multimodal format and diverse content ingestion PPT, indico page contents.
  - Improving uncertainty quantification of the generated answer.
  - Integrating agentic RAG for a comprehensive workflow.

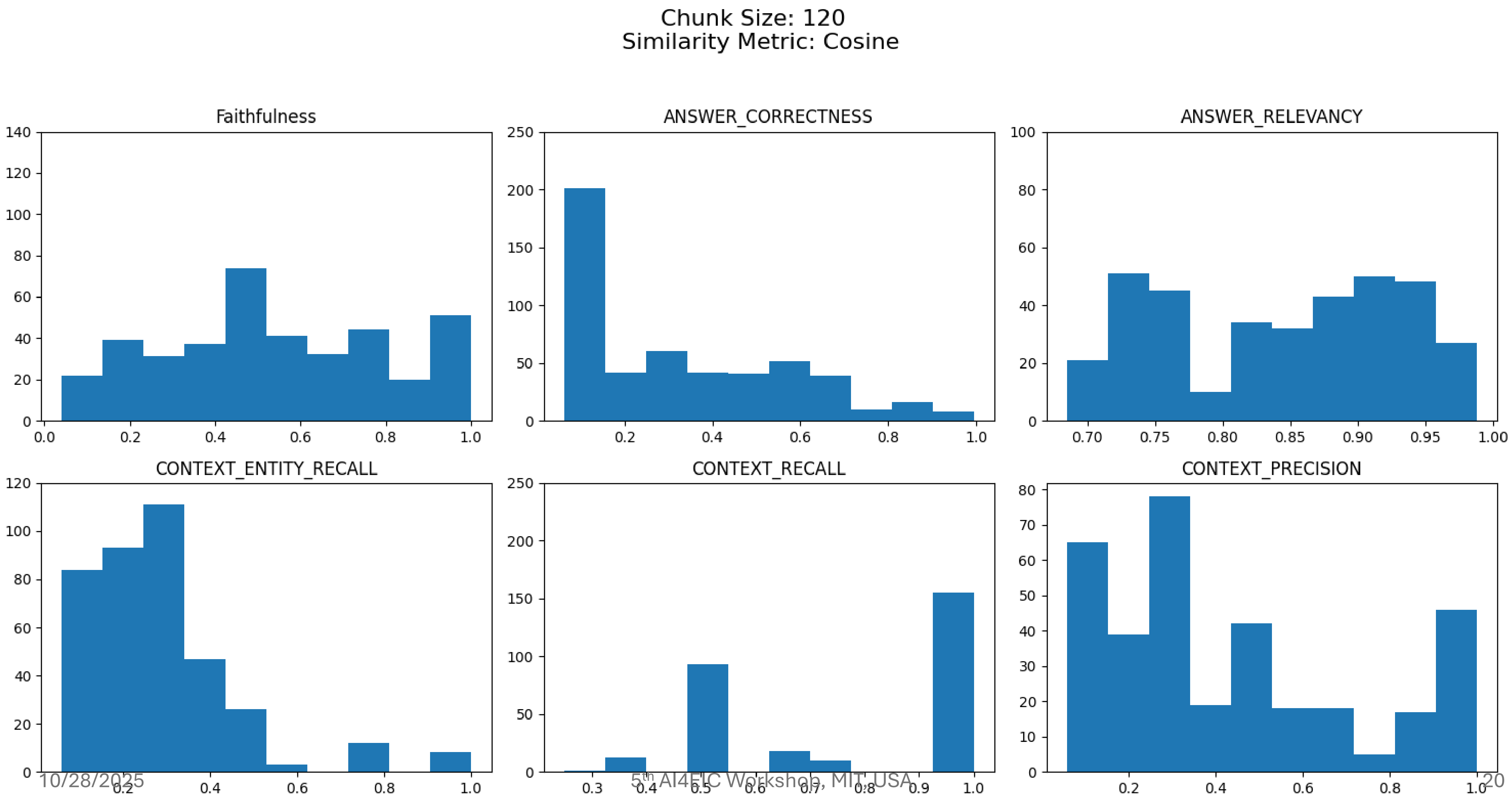
Acknowledgement:

- Tina J Jat, B.Sc. Student
- Karthik Suresh and Cristiano Fanelli, College of W&M
- The Ramaiah University of Applied Sciences
- The organizers of AI4EIC workshop

Thank You

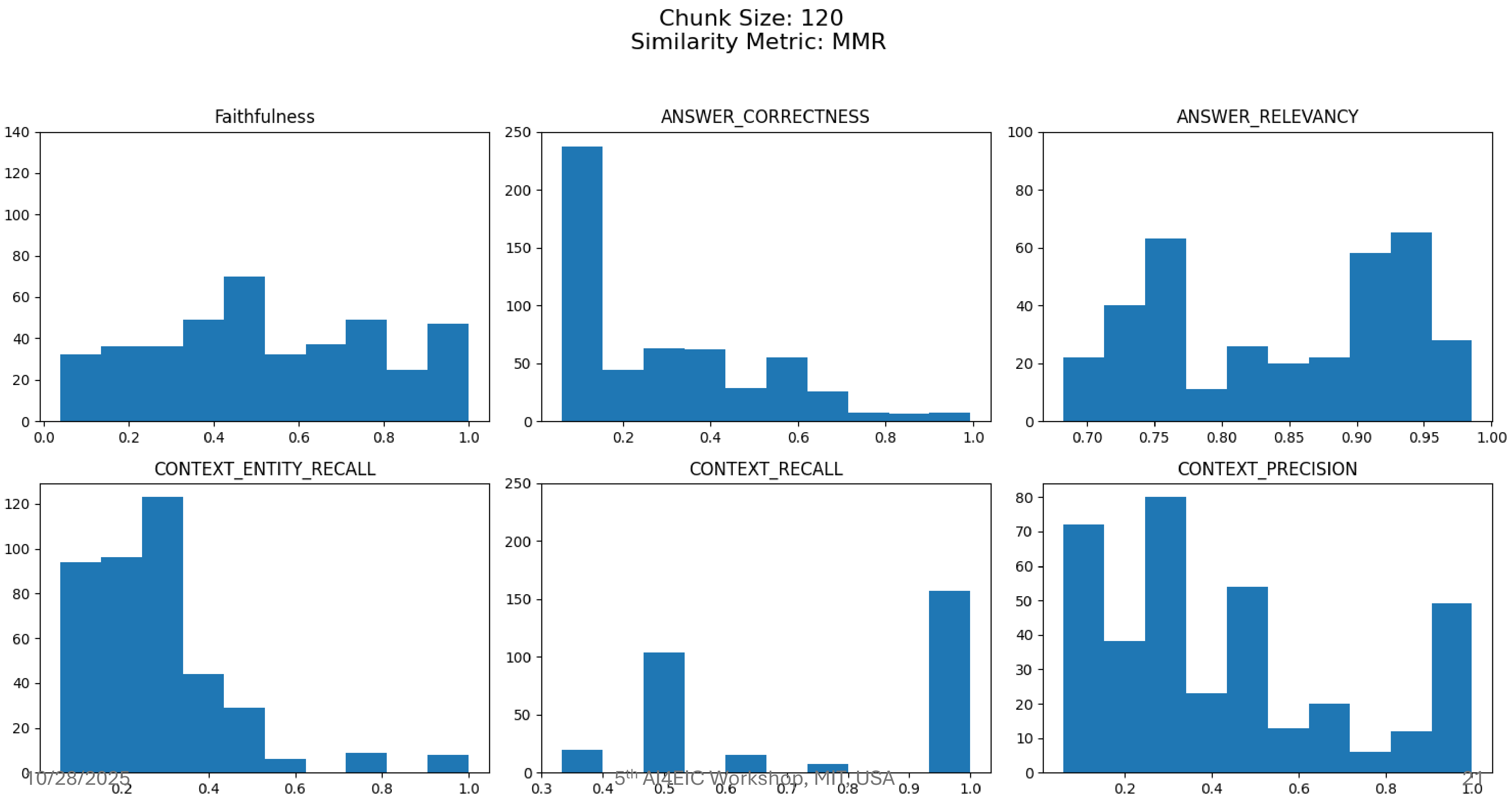
Backup

Results:



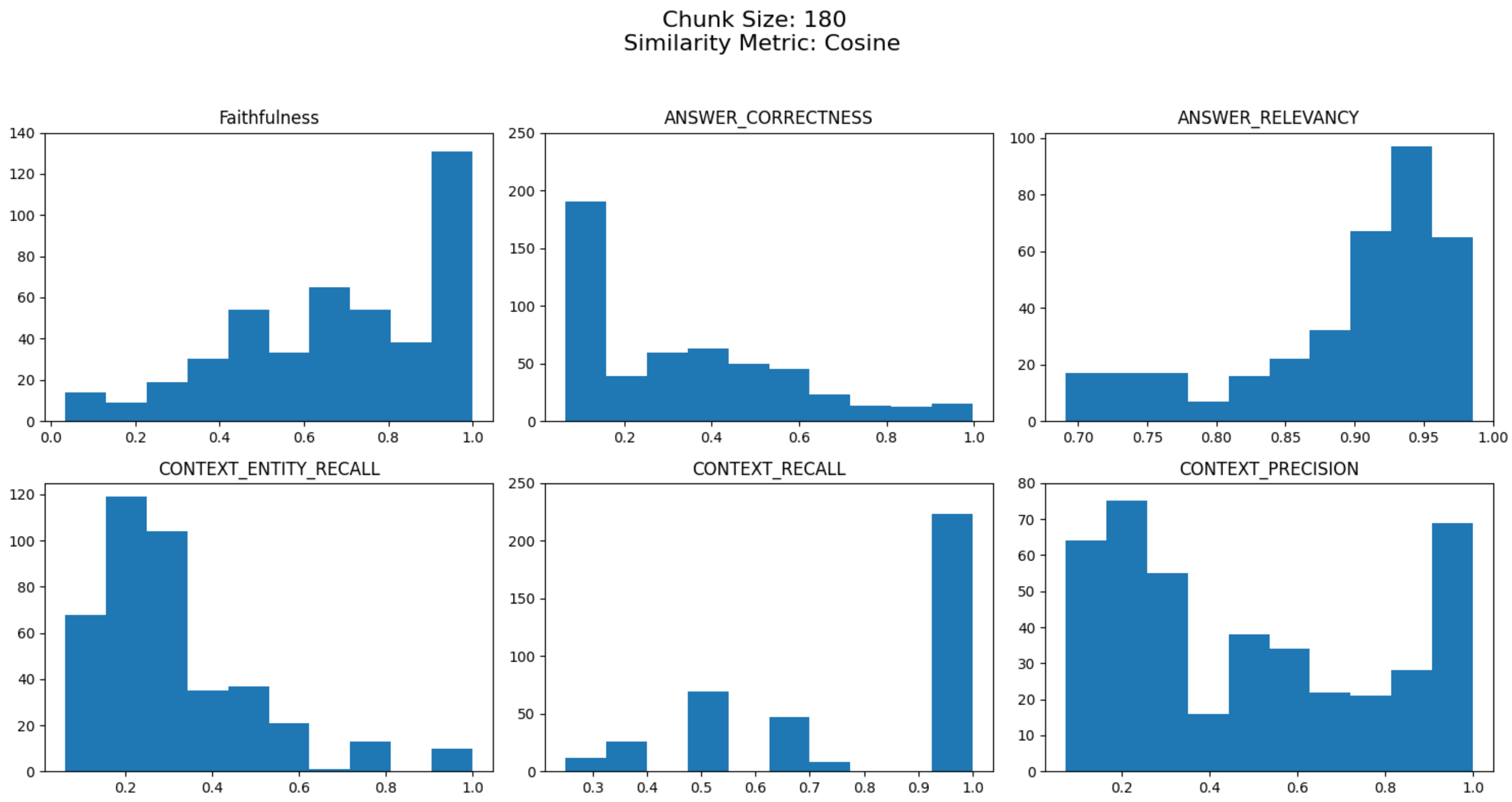
Chunk Size: 120
Similarity Metric: Cosine
Faithfulness
ANSWER_CORRECTNESS
ANSWER_RELEVANCY
CONTEXT_ENTITY_RECALL
CONTEXT_RECALL
CONTEXT_PRECISION

Results:



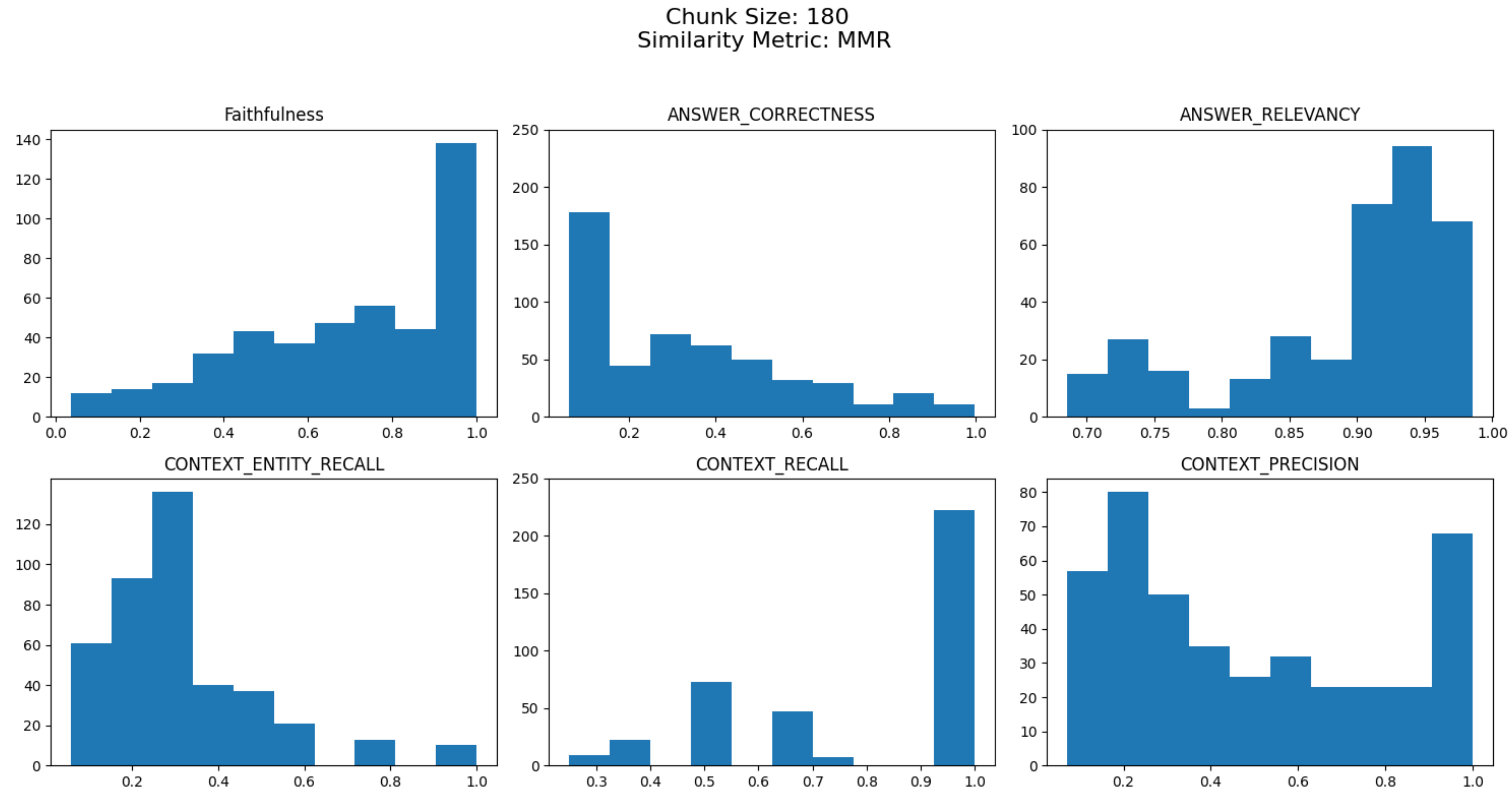
Chunk Size: 120
Similarity Metric: MMR
Faithfulness
ANSWER_CORRECTNESS
ANSWER_RELEVANCY
CONTEXT_ENTITY_RECALL
CONTEXT_RECALL
CONTEXT_PRECISION

Evaluation:



Chunk Size: 180
Similarity Metric: Cosine
Faithfulness
ANSWER_CORRECTNESS
ANSWER_RELEVANCY
CONTEXT_ENTITY_RECALL
CONTEXT_RECALL
CONTEXT_PRECISION

Evaluation:


Chunk Size: 180
Similarity Metric: MMR
Faithfulness
ANSWER_CORRECTNESS
ANSWER_RELEVANCY
CONTEXT_ENTITY_RECALL
CONTEXT_RECALL
CONTEXT_PRECISION

Results: Latency in sec (26 Gb GPU)

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