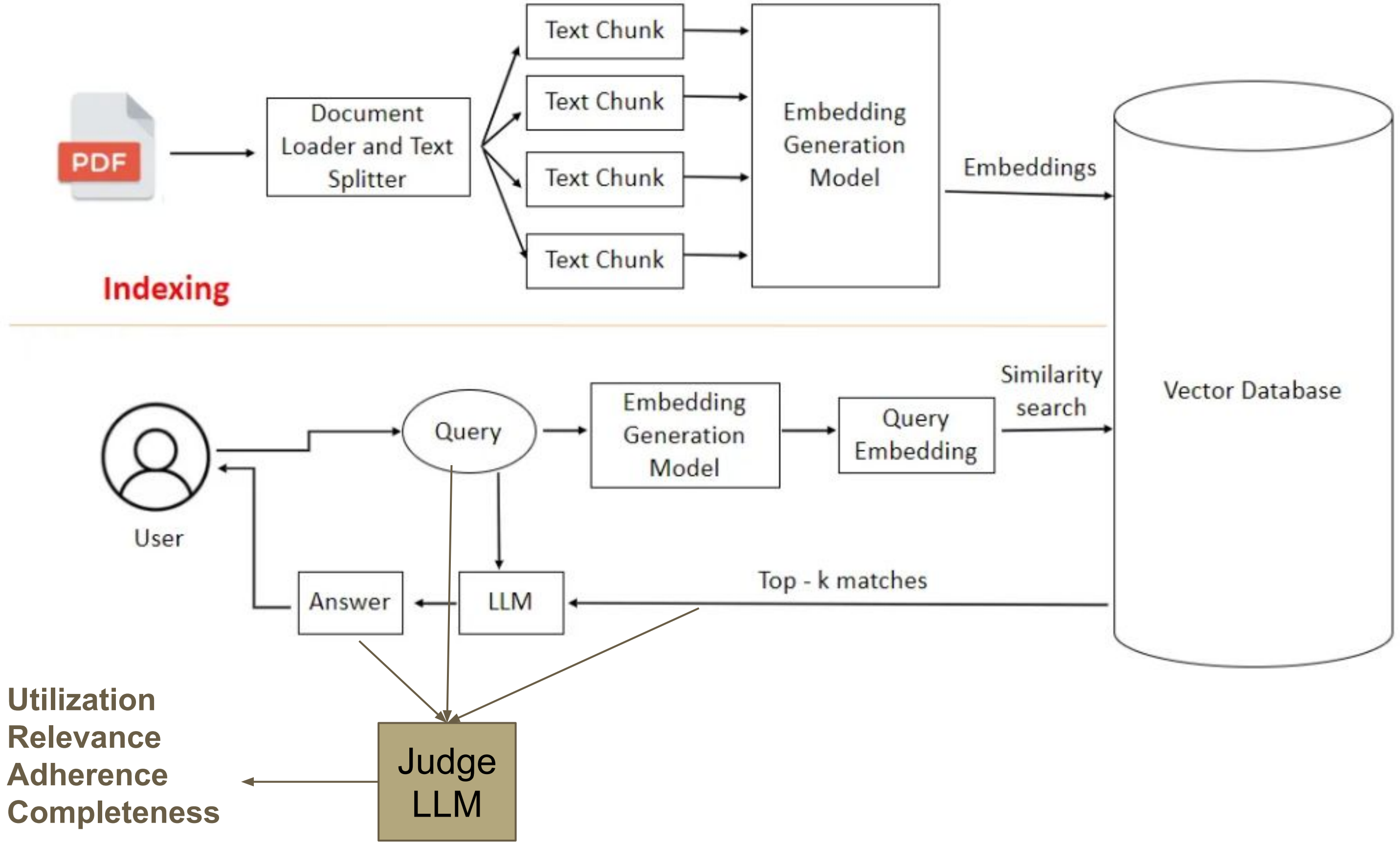
RAG-RGB Project

Group 25

Srinivas Gundu
Kaveri Pulibandla
Rahul Pelluri

Overview


PDF
Document Loader and Text Splitter
Text Chunk
Text Chunk
Text Chunk
Text Chunk
Embedding Generation Model
Embeddings
Indexing
Vector Database
Similarity search
User
Query
Embedding Generation Model
Query Embedding
Top - k matches
Answer
LLM
Judge LLM
Utilization
Relevance
Adherence
Completeness

Scope of Experiments

- Data chunking
- Embedding Model changes
- Using different Vector DBs
- Different Retrieval techniques
- Modifying Judge LLM Prompt
- Gradio Implementation

Experiments Conducted

- Chunking strategies
    - Experiments
        - Sentence division
        - Recursive Character Text Splitter
        - Semantic Chunking
        - We stuck to recursive character text splitting. Semantic chunking took too much time and quickly exhausted our GPU credits.
- Embedding Model changes
    - Evaluated Models on MTEB with existing ratings for different DB types
    - Experimented with small and large param models
    - Experimented with 11 models overall
- Vector DBs
    - Used FAISS and Chroma
    - Chroma offers more flexibility
    - Metadata generated from local LLM is integrated with Chroma to refine query. The process was GPU intensive and time taking but improved retrieval
- Retrieval Techniques
    - Hybrid, Vector retrieval mechanisms best combination was 0.1 ratio

Finance Results

								Hybrid			Hybrid 25pc			Vector			MetadataSearch		
id	gpt3_ad	gpt3_co	gpt35_u	relevan	utilizati	complet	Adhere	utilizatio	complete	relevanc	utilizatio	complete	relevanc	utilizatio	complete	relevanc	utilizatio	complete	relevance
finqa_66	null	null	null	0.0588	0.0588	1	1	0.03455	1	0.02827	0.0911	1	0.15602	0.15393	1	0.14136	0.26154	1	0.54136
finqa_63	1	0.3333	0.1111	0.1111	0.1111	1	1	0.97059	0.33333	0.84454	NA	NA	NA	NA	NA	NA	NA	NA	NA
finqa_70	1	0.04	0.04	0.04	0.04	1	1	0.28558	1	0.06437	0.28105	0.75	0.11786	0.11242	1	0.11786	0.271	1	0.67664
finqa_70	1	0.2	0.05	0.05	0.05	1	1	0.09524	1	0.04762	0.08163	1	0.08163	0.33673	1	0.33673	NA	NA	NA

Embedding Model: “multilingual-e5-large-instruct”
Vector DB: “**Chroma**”
Chunking Size: 800
Overlap: 100
Retriever LLM: “llama3-8b-8192”
Judge LLM: “llama3-70b-8192”

General Knowledge Results

Id	Question	Relevance	Utilzation	Adherence	Completeness
hagrid_4293_0	What is the dominate religion in Pristina?	1	1	1	1
hagrid_3369_0	What's the melting point of Silicon?	1.25	1	1	1
hagrid_534_0	When did Iain Norman Macleod die?	1.65	0.67	1	1
hagrid_363_0	How old is Drake Hogestyn?	1	1	1	1
expertqa_1471	What type of genu valgum is normal?	1	1	1	1
expertqa_493	What are examples of special education needs?	1	2	1	1
expertqa_130	How can I build a portfolio?	0.87	1.18	1	1
8187	what causes ghost ants	1	1	1	1
7188	uses of copper sulphate crystals	1	1	1	1
1929	what pollutants come from factories	0.6	2.5	1	1
5a83093c55429966c78a6b01	Were both Léopold Eyharts and Ulrich Walter a General in the French Air Force?	1	1	1	1
5a875154554299211dda2be7	Who founded the honky tonk that is at the center of John Travolta's third major acting role?	1	1	1	1
5abccb235542996583600497	Where does the team coached by someone with the nickname "Coach K" play?	2	0.5	1	1

Embedding Model: "sentence-transformers/all-MiniLM-L6-v2"
Vector DB: "**FAISS**"
Chunking Size: 700
Overlap: 100
Retriever LLM: "llama3-8b-8192"
Judge LLM: "llama3-70b-8192"

Medical Results

id	question	relevance_score	utilization_score	completeness_score	relevance_score	utilization_scor	completeness_score
677	When was the first case of COVID-19 identified?	0.269231	0.076923	0.285714	0.125	1	1
1756	Is there an Influenza vaccine?	0.333333	0.333333	1	1	1	1
766	What is a future potential of filamentous phage?	0.1	0.1	1	1	0.5	1
1421	Which viruses may not cause prolonged inflammation due to strong induction of antiviral clearance?	0.411765	0.176471	0.428571	0.833	1	0.5

Embedding Model: “pritamdeka/BioBERT-mnli-snli-scinli-scitail-mednli-stsb”
Vector DB: “**Chroma**”
Chunking Size: 800
Overlap: 100
Retriever LLM: “llama3-8b-8192”
Judge LLM: “llama3-70b-8192”

Customer Support Results

Customer Support (TechQA)

id	question	relevance_score	utilization_score	completeness_score	relevance_score	utilization_score	completeness_score
techqa_DEV_Q243	Using cobol copybooks Sometimes, there will be errors/fields missing in typetree, while importing cobol copybooks. Is there any format for copybooks(specifically to be used in wtx), that we need to follow?	0.011719	0.011719	1	1	1	1
techqa_DEV_Q253	Want to find out if Microsoft Edge is supported with ICC? Want to find out Microsoft Edge is supported with ICC?	0.00266	0.00133	0.5	1	1	1
techqa_DEV_Q008	How can I export a private key from DataPower Gateway Appliance? How can I export a private key	0.062718	0.062718	1	1	1	1
techqa_DEV_Q266	How to install Packaging utility? How to install Packaging utility?	0.070093	0.046729	0.533333	0.8	1.25	1

Embedding Model: “sentence-transformers/all-MiniLM-L6-v2”
Vector DB: “**FAISS**”
Chunking Size: 500
Overlap: 100
Retriever LLM: “llama3-8b-8192”
Judge LLM: “llama3-70b-8192”

Legal Results

id	relevance_score	utilization_score	completeness_score	relevance_score	utilization_score	completeness_score
PerformanceSportsBrandsInc_20110909_S-1_ EX-10.10_7220214_EX-10.10_Endorsement Agreement__Ip Ownership Assignment	0.014085	0.014085	1	0.4	1	0.5
SPARKLINGSPRINGWATERHOLDINGSLTD_07_03_2002- EX-10.13-SOFTWARE LICENSE AND MAINTENANCE AGREEMENT__Volume Restriction	0.010811	0.005405	0.5	0.4	1	1

Embedding Model: “sentence-transformers/all-MiniLM-L6-v2”
Vector DB: “**FAISS**”
Chunking Size: 500
Overlap: 100
Retriever LLM: “llama3-8b-8192”
Judge LLM: “llama3-70b-8192”

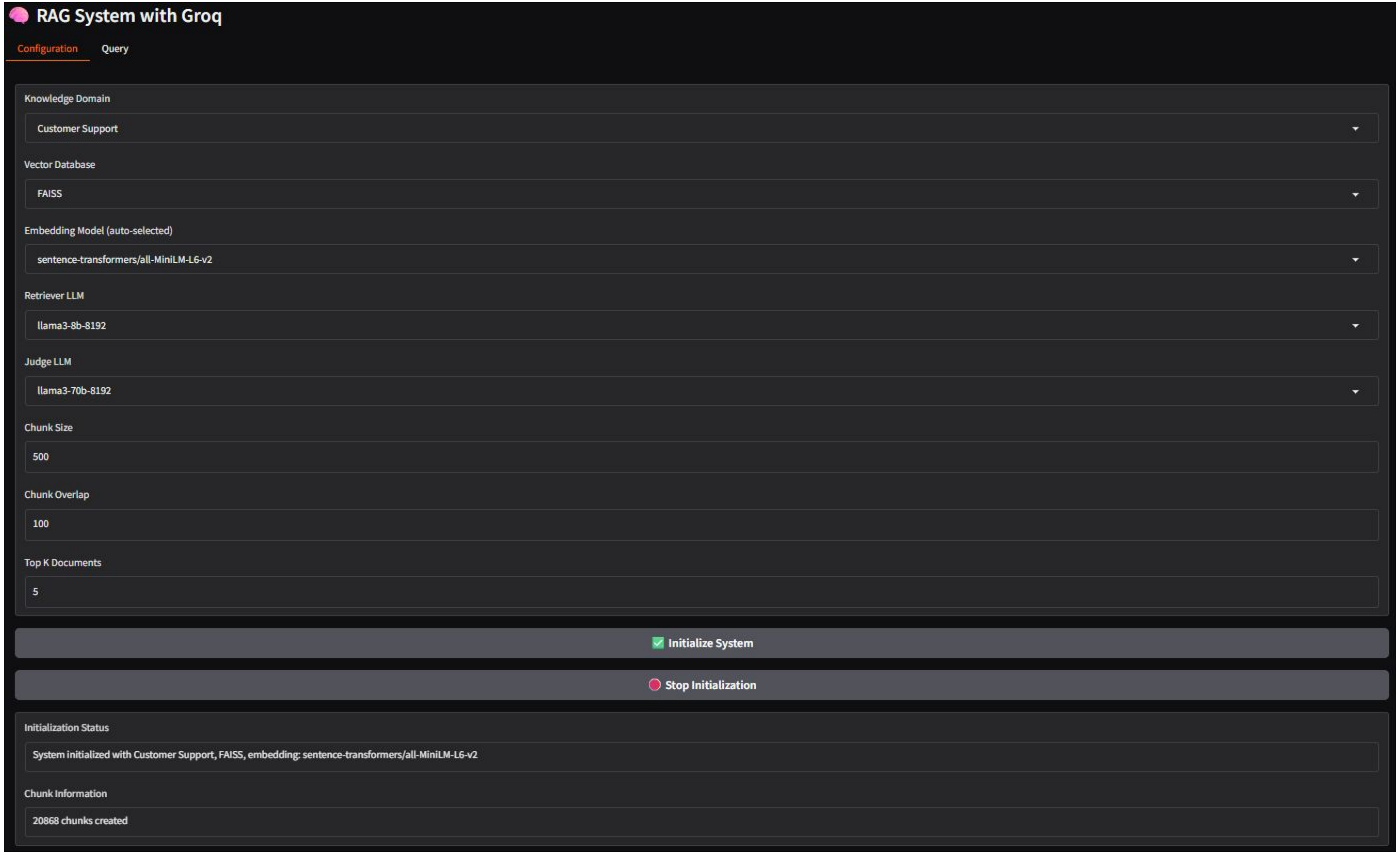
Observations

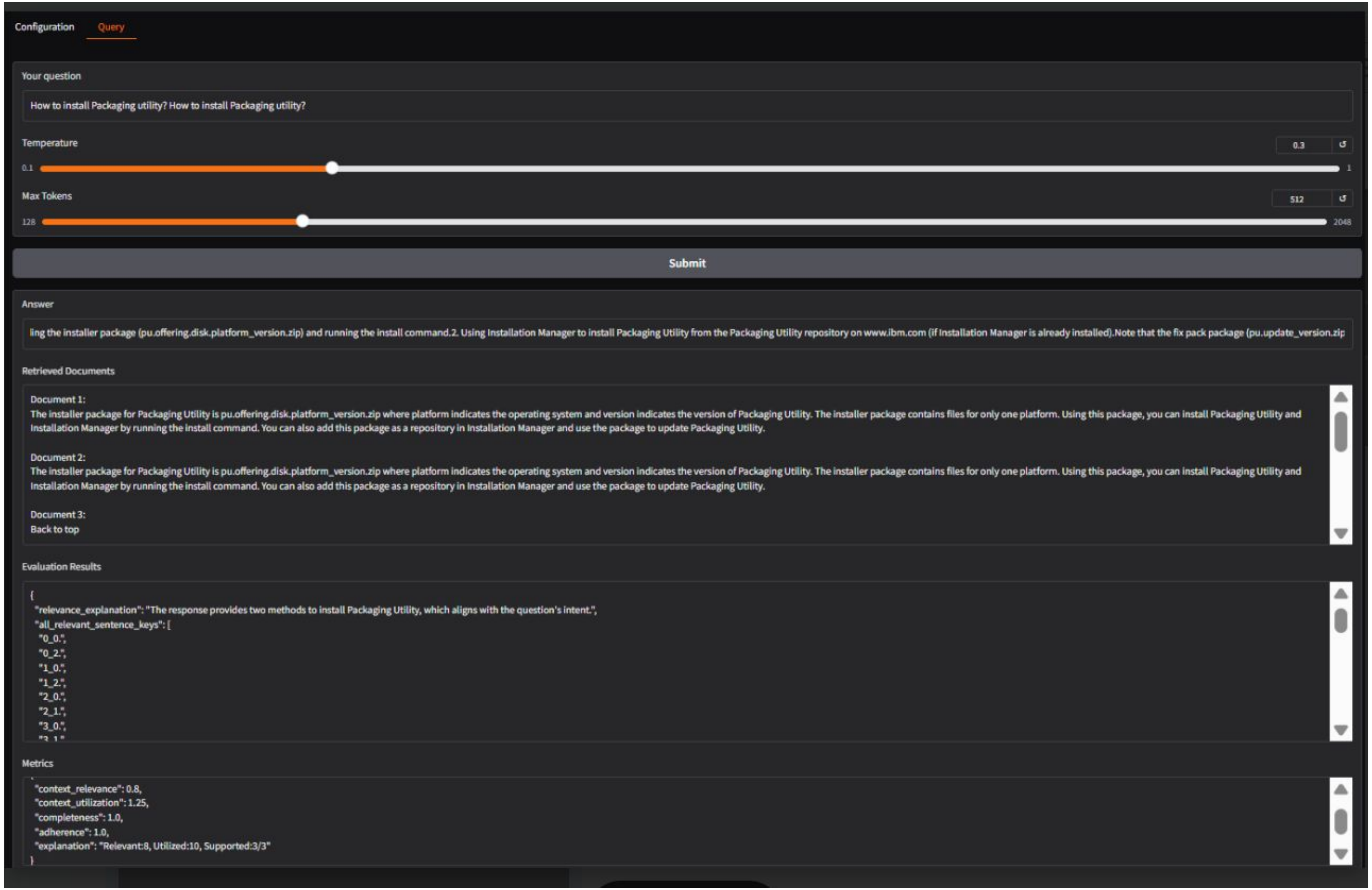
- Domain-specific embeddings (BioBERT) significantly improved retrieval relevance.
- Sentence-level metadata helped in accurate sentence mapping for judge LLM.
- Metadata search on Chroma gave better results but the GPU usage became very high and time taking.
- Query decomposition took time. But once implemented, it increased the number of retrieved documents and affected the final KPIs.
- Query decomposition is a good idea but takes more overall inference time.

Best Practices for Performance Improvement

- Use multiproc for chunking the data
- Chunk size less than 1000 for best results (As per research papers)
- ChromaDB Langchain wrapper does not support some functionality (progress bar, metadata incorporation).
  - Use Native Chroma library for creating the DB and Langchain for wrapper to query later.
  - Get summarization words for each chunk from a 7B LLM and feed as metadata. Query on the metadata as well as vector retrieval for best results.
  - Run a local LLM on Collab GPU for metadata creation. (Used Mistral 7B)
- Tweaked the judge prompt for more straightforward results.
- Cache the results for repeated queries. More work needed.

Gradio Implementation




Configuration
Query
Your question
How to install Packaging utility? How to install Packaging utility?
Temperature
0.3
0.1
1
Max Tokens
512
128
2048
Submit
Answer
ling the installer package (pu.offering.disk.platform_version.zip) and running the install command.2. Using Installation Manager to install Packaging Utility from the Packaging Utility repository on www.ibm.com (if Installation Manager is already installed).Note that the fix pack package (pu.update_version.zip
Retrieved Documents
Document 1:
The installer package for Packaging Utility is pu.offering.disk.platform_version.zip where platform indicates the operating system and version indicates the version of Packaging Utility. The installer package contains files for only one platform. Using this package, you can install Packaging Utility and Installation Manager by running the install command. You can also add this package as a repository in Installation Manager and use the package to update Packaging Utility.
Document 2:
The installer package for Packaging Utility is pu.offering.disk.platform_version.zip where platform indicates the operating system and version indicates the version of Packaging Utility. The installer package contains files for only one platform. Using this package, you can install Packaging Utility and Installation Manager by running the install command. You can also add this package as a repository in Installation Manager and use the package to update Packaging Utility.
Document 3:
Back to top
Evaluation Results
{
"relevance_explanation": "The response provides two methods to install Packaging Utility, which aligns with the question's intent.",
"all_relevant_sentence_keys": [
"0_0.",
"0_2.",
"1_0.",
"1_2.",
"2_0.",
"2_1.",
"3_0.",
Metrics
"context_relevance": 0.8,
"context_utilization": 1.25,
"completeness": 1.0,
"adherence": 1.0,
"explanation": "Relevant:8, Utilized:10, Supported:3/3"
}

Overall Results

After using metadata based search, overall search quality shoots up.

All Datasets	Relevance	Utilization	Completion	Adherence
cuad	0.4	1	0.5	1
finqa	0.84	0.62	0.9	1
tatqa	1	0.94	1	1
covidqa	0.125	1	1	1
pubmedqa	0.833	1	1	1
hotpotqa	0.65	1	1	1
nsmarco	1	0.72	1	1
expertqa	0.87	1	1	1
hagrid	1	0.67	1	1
techqa	0.8	0.666	1	1
enaul	0.86	1	1	1
delucionqa	1	1	1	1

Domain-wise Averages:

1. Biomedical (pubmedqa, covidqa)

Relevance: (0.833 + 0.125) / 2 = 0.479

Utilization: (1 + 1) / 2 = 1

Completion: (1 + 1) / 2 = 1

Adherence: (1 + 1) / 2 = 1

2. Finance (finqa, tatqa)

Relevance: (0.84 + 1) / 2 = 0.92

Utilization: (0.62 + 0.94) / 2 = 0.78

Completion: (0.9 + 1) / 2 = 0.95

Adherence: (1 + 1) / 2 = 1

3. Lega(cuad)

Relevance: 0.4

Utilization: 1

Completion: 0.5

Adherence: 1

4. Customer Support (enaul, techqa, delucion

Relevance: (0.86 + 0.8 + 1) / 3 = 0.886

Utilization: (1 + 0.666 + 1) / 3 ≈ 0.889

Completion: (1 + 1 + 1) / 3 = 1

Adherence: (1 + 1 + 1) / 3 = 1

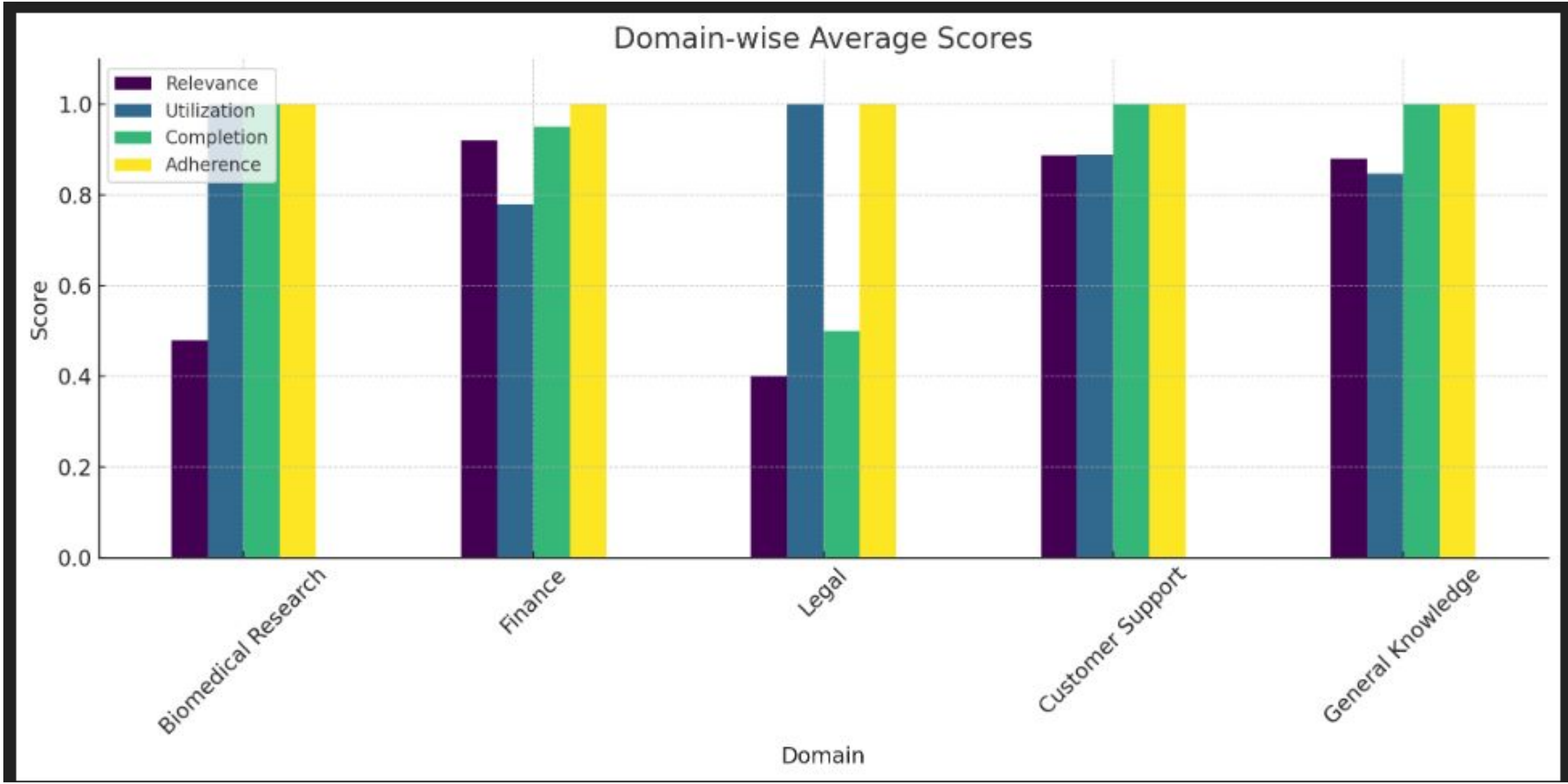
5. General Knowledge (hotpotqa, hagrid, nsm

Relevance: (0.65 + 1 + 1 + 0.87) / 4 = 0.88

Utilization: (1 + 0.67 + 0.72 + 1) / 4 ≈ 0.8475

Completion: 1 (all 1s)

Adherence: 1 (all 1s)


Domain-wise Average Scores
Relevance
Utilization
Completion
Adherence
Score
1.0
0.8
0.6
0.4
0.2
0.0
Biomedical Research
Finance
Legal
Customer Support
General Knowledge
Domain

RMSE & AUC-ROC (Overall Results)

We compared predicted scores with ground-truth annotations from RAGBench to compute regression & classification metrics.

RMSE - Measures average error between predicted and actual scores.

- RMSE for context relevance

Dataset	RMSE	AUCROC
FINANCE	0.11	0.72
LEGAL	0.2	0.76
GENERAL KNOWLEDGE	0.15	0.89
CUSTOMER SUPPORT	0.12	0.91
BIOMEDICAL RESEARCH	0.14	0.87

RGB Dataset

Observations

- Code uploaded in github is buggy.
- Tried to run a LLAMA based quantised LLM on Collab for judge model but requires a Pro connection.

Scope of Experimentation

- Large param models anyways have a higher accuracy score.
- Idea has been to check for smaller LLMs which will be deployed onto edge devices
- We have run LLAMA 2 based models to check for performance between fine tuned models vs community models

Results

- Tiny LLM is a fine tuned model
- Mini LLAMA is a smaller community model
- Everything degrades with noise.

Feature	Mini LLAMA (1.1B)				TinyLLAMA (1.1B)			
	Noise 0%	Noise 25%	Noise 50%	Noise 100%	Noise 0%	Noise 25%	Noise 50%	Noise 100%
Negative Rejection				40%				45%
Counterfactual robustness	20%	12%	8%		26%	22%	16%	22%
Information Integration	32%	27%	22%		38%	35%	32%	35%

Thank You