DOI: https://doi.org/10.57233/ijsgs.v11i2.872 ISSNp: 2488-9229; ISSNe: 3027-1118

IJSGS FUGUSAU VOLXXX WEBSITE: https://fugus-ijsgs.com.ng

INTERNATIONAL JOURNAL OF SCIENCE FOR GLOBAL SUSTAINABILITY
(A PUBLICATION OF FACULTY OF SCIENCE, FEDERAL UNIVERSITY GUSAU, NIGERIA)

Analyzing the Impact of Various Indexing Techniques on Retrieval-Augmented Generation (RAG) Performance in Closed-Domain Question Answering

[1]Buhari Bashir, [2]Ismail Ahmad Ibrahim, [3]Ameenu Abdulhameed Rabiu and [4] Salihu Abdullahi Audu

[1]No. 8 Funtua Crescent K/Kaura New Layout, Katsina.
[2]No. 13 Dutsinma Street Tudun Wada, Kaduna.
[3]No. 70 Kundilar Gandu, Zoo Road Housing Estate, Kano
[4]Department of Computer Science, Nasarawa State University, Nigeria

Corresponding Author's Email: abdulgiz@nsuk.edu.ng

Received on: April, 2025 **Revised and Accepted on:** May, 2025 **Published on:** July, 2025

ABSTRACT

Rapid and recent advancements in large language models (LLMs) have become the driving force of many Natural Language Processing (NLP) applications, revolutionizing tasks such as Question Answering (QA) with the aim of improving human-computer interaction. LLMs have been explored to be capable of predicting reasonably good answers for provided questions using its ability to memorize information seen during training of the model. But to what extent can these models remember training data? Retrieval-Augmented Generation (RAG) augments LLMs with knowledge to avoid relying on their memorization capabilities by combining generative capabilities of LLMs with retrieval-based methods to enhance answer accuracy and relevance of answers. With or without RAG, the aim is to mitigate the generation of plausible but non factual responses by LLMs which is hallucination. RAG has demonstrated significant performance in reducing hallucination and improving accuracy, speed and relevance of LLM generation in QA tasks. This research evaluates how different indexing methods such as BM25, DPR, and hybrid techniques affect RAG system performance in closed-domain QA, examining accuracy of retrieval and generation using performance metrics such as hit rate, precision, faithfulness and f1-score. The study also considers passage length variability and domainspecific adaptation across multiple datasets and domains with the hypothesis that focusing on a single domain can enhance system performance. Experiments are carried out on multiple datasets within both closed and open domains, using pre-trained models. This study explores optimal indexing strategies for RAG systems, balancing accuracy and efficiency. Results show that increasing passage length and retrieved samples improves hit rate and recall but reduces precision. The best performance achieved was 91% Hit-rate@9 and 93% relevancy using the BioASQ dataset.

Keywords: Language Processing, Retrieval-Augmented Generation, large language models, Question Answering

1.0 INTRODUCTION

1.1 Background

The advents of LLM-powered applications, such as ChatGPT, Bing Chat, and CoPilot, have undeniably reshaped human-computer interaction, showcasing unprecedented levels of performance Chen et al. (2021). These Large Language Models (LLMs) have demonstrated excellent performance in tasks ranging from language understanding to text generation, paving the way for more sophisticated human-computer interactions (Andriopoulos and Pouwelse, 2023). LLMs are used for Question Answering (QA) tasks, generating responses to user questions. These responses can be based on the model's fine-tuned knowledge (closed-book approach) or external data sources such as search engines to enrich the input context with more relevant information (open-book approach) (Ram et al., 2023). While closed-book approaches showcase LLMs' capabilities, they face scalability challenges due to size and computational requirements. Larger models are needed to improve accuracy and reduce hallucinations—plausible but non-factual predictions common in LLMs.

The fundamental framework of RAG comprises of three key stages; chunking and indexing documents, retrieving relevant information based on vector similarity to user queries, and generating responses using an LLM augmented with the retrieved context (Lewis et al., 2021).

Despite growing interest in RAG systems, there's a gap in literature comparing indexing techniques and their impact on RAG performance. This study will analyze different RAG retrievers, including BM25, dense passage retrieval, and hybrid retrieval, to explore various indexing methods (Ma et al., 2021). The impact of these techniques on closed-domain QA, along with the influence of passage length on retrieval performance Hsia et al. (2024), remains largely unexplored. Closeddomain QA has advantages over open-domain when using RAG due to higher overlap of

DOI: https://doi.org/10.57233/ijsgs.v11i2.872 ISSNp: 2488-9229; ISSNe: 3027-1118

question-answer pairs within the data source. This comprehensive evaluation aims to contribute to developing more effective and scalable closed-domain QA solutions,

1.2 Aim and Objectives

The aim of this paper is to analyze and evaluate the impact of various indexing techniques on the performance of RAG systems in the context of closed-domain QA. This investigation will provide insights into how different techniques of indexing documents influence the overall performance of RAG systems in terms of effectiveness and accuracy. Thus providing best approach that aims to enhance the retrieval and generation of relevant answers in closed-domain QA tasks.

Objective 1: Evaluate the effect of increasing passage length on RAG performance in closed-domain QA by conducting experiments with varying passage length and measuring the resulting impact on retrieval and generation accuracy of RAG systems.

Objective 2: Investigate the influence of different indexing techniques on RAG performance by comparing the effectiveness of various ways of indexing and retrieving documents relevant to closed-domain QA tasks.

Objective 3: Assess the impact of domain adaptation of indexing techniques on RAG performance across different domains, including closed and open domains.

Objective 4: Comparing the retrieval performance of a generic pretrained model versus domain-specific fine-tuned model to assess the impact of domain adaptation on passage retrieval.

Objective 5: Analyze the scalability of RAG systems under different indexing techniques by evaluating their performance across a diverse range of closed-domain QA tasks and datasets.

1.3 Hypotheses

Hypothesis 1: Increasing passage length impacts the performance of RAG in closeddomain QA by providing more contextual information.

Hypothesis 2: The choice of RAG indexing technique will impact performance of an entire RAG system, showcasing various levels of effectiveness in closed-domain QA. This hypothesis suggests that certain indexing techniques may be better suited for indexing 3 and retrieving documents relevant to specific domains, thus influencing the overall performance of RAG systems.

Hypothesis 3: Domain adaptation of RAG systems differs significantly across closed and open domain QA. Closed domains that have well-defined and limited subject matter, are expected to achieve higher accuracy and effectiveness compared to open domains where the subject matter is more diverse and extensive when using RAG.

2.0 MATERIALS AND METHOD(S)

2.1 Requirement Analysis

Hardware Requirements The following hardware specifications are recommended to conduct the experiments for analyzing RAG systems:

i. **Processing Power** to ensure that Multi-core CPU for parallel processing tasks and High-performance GPU Liu et al. (2024) are capable of handling large-scale data processing and complex model computations. Specifically, GPUs with high CUDA cores and VRAM (>32GB) for efficient parallel processing and running computationally expensive libraries like PyTorch were used. If the required compute is not available locally, it can be rented from the cloud.

ii. **High Memory Capacity** (>64GB) to support in-memory operations to handle large datasets and intensive computing tasks and avoid bottlenecks during data processing and retrieval of relevant passages in QA task.

iii. **Large Storage Space** such as Solid Sate Drive (SSD) is recommended for efficient storage Mielke et al. (2017) of datasets, indexed databases, model checkpoints, and logs due to their fast data access and retrieval.

Software Requirements The following software specifications are recommended to conduct the experiments for analyzing RAG systems:

i. **Operating System** such as Linux-based OS will be used which is great for handling deep learning libraries compared to windows-based and others, making linuxbased OS easier integration with server environments.

ii. **Integrated Development Environment (IDE)** like PyCharm or Visual Studio Code. Jupyter Notebook will be used for experimentation.

iii. **Version control systems** like Git to manage codebase changes.

iv. **Programming Languages**: Python (version 3.8 or higher) due to its extensive support for data science and machine learning libraries was used.

2.2 Data Sources and Selection Criteria

The literature review is centred on published work between 2010 and 2023 from peer-reviewed journal

DOI: https://doi.org/10.57233/ijsgs.v11i2.872 ISSNp: 2488-9229; ISSNe: 3027-1118

articles, conference proceedings, and scholarly book chapters. The following inclusion criteria were applied:

i. **Relevance:** Studies must examine ML, LLMs, data harmonisation, or decision support systems within the context of higher education.
ii. **Quality:** Only publications from reputable, peer-reviewed journals or established academic conferences were selected to ensure methodological rigor and reliability.
iii. **Diversity:** The review incorporates studies from different geographical regions and educational systems to provide a comprehensive and globally informed perspective.

Searches were conducted across major academic databases, including Google Scholar, JSTOR, IEEE Xplore, and Scopus, using combinations of the following keywords: "machine learning," "large language models," "data harmonisation," and "higher education." Boolean operators and filters were applied to refine results and exclude duplicates or non-relevant works.

2.3 Research Design

The research design for this study employs a quantitative research design, using experimental methods to test the hypotheses, meet the objectives outlined in Chapter 1 and to systematically investigate the impact of various indexing techniques, passage length and domain adaptation on RAG performance in closed-domain question answering.

2.4 Dataset Selection and Preparation

This study will make use of two (2) different datasets centered around the medical information. The CovidQA dataset, having a single domain of COVID-19 information, is categorized as a closed-domain dataset, while the BioASQ dataset, which comprises a wider range of domains within the biomedical field, can be classified as open-domain to some extent.

2.5 Libraries and Frameworks

The following libraries and frameworks are essential for implementing a robust RAG system for QA:

- **Core Technologies**:

i. Python: The primary programming language
ii. PyTorch Imambi et al. (2021): For building and training neural networks
iii. Hugging Face Transformers : For pre-trained NLP models such as embeded model or LLM
iv. LlamaIndex : For RAG-specific indexing, retrieval and generation

- **Other Technologies**:

i. Pandas and NumPy: For data manipulation and numerical computing
ii. Datasets: For efficient dataset handling
iii. Json: For parsing and manipulating JSON data.
iv. Best asyncio: For enabling asynchronous operations in Jupyter notebooks.
v. Matplotlib and Seaborn: For data visualization.

2.6 RAG System Architecture

RAG Retriever is responsible for indexing and retrieving relevant passages while RAG Generator is any language model that can generate answers based on retrieved passages Yu et al. (2024).

RAG pipeline will be integrated as follows:

i. Indexing of documents using selected indexing techniques
ii. User query input
iii. Retriever selects relevant passages based on created index
iv. Retrieved passages and original query are fed to the generator using prompt
v. Generator produces the final answer
vi. Both Retrieval and generation will be evaluated

2.7 Experimental Setup

Passage Length Variation: To address Hypothesis 1 and Objective 1, it is important to consider splitting the documents into chunks in order to experiment with multiple passage lengths Finardi et al. (2024)

Motivation: This range of passage lengths allows us to investigate the trade-off between conciseness and contextual information. Shorter passages (128, 512 tokens) may provide more focused information, while longer passages (1024, 2048 tokens) offer more context. This approach is supported by research showing that passage length can significantly impact retrieval performance Dai and Callan (2019).

Table 1: Hyperparameter Tuning Range for QA RAG System

Parameter	Tuning Range
Chunk size	(128, 512, 1024, 2048)
Similarity (TopK)	(3, 5, 9)

2.3 Evaluation Metrics

The effectiveness of each indexing technique of RAG in closed-domain QA will be evaluated based on performance metrics such as Gao et al. (2024) Accuracy, Precision, Recall, AP, NDCG, Hit-rate and MRR for retrieval, while F1-score, faithfulness, semantic similarity and relevancy will be used for generation. The ability to reduce hallucination in generated answers alongside newer metrics tailored to

DOI: https://doi.org/10.57233/ijsgs.v11i2.872 ISSNp: 2488-9229; ISSNe: 3027-1118

assess the contextual relevance of the answers.

Retrieval Evaluation:

All retrieval evaluation metrics are measured @topK value. Therefore, choice of similarity k values are important. This study will be using (3, 5, 9) as k values.

Retrieval evaluation is greatly simplified when working with LlamIndex. The retriever evaluator class is capable of generating evaluation results from provided metrics. However, QA-Dataset must be provided which is a ”EmbeddingQAFinetuneDataset” type and contains list of queries, corpus and relevant docs.

Accuracy is the proportion of correctly answered questions (Yacouby and Axman, 2020).

Hit-Rate refers to the proportion of queries for which the retrieval system successfully returns at least one relevant document within a specified number of top results Tamm et al. (2021).

Mean Reciprocal Rank (MRR) at K evaluates how quickly a ranking system can show the first relevant item in the top-K results. In RAG systems, a higher MRR indicates that the system is generally retrieving relevant documents closer to the top of the list, which is desirable for providing accurate context to RAG generator Tamm et al. (2021) .

Precision is the effectiveness of retrieving relevant documents only. It indicates how well the model is at predicting a positive class by calculating the percentage of true positive predictions among all positive predictions.

True Positives are instances correctly predicted as positive while False Positives are instances incorrectly predicted as positive (Yacouby and Axman, 2020).

Recall is the effectiveness of retrieving relevant documents only. The ability of the model to catch every positive case in the dataset. It indicates how well the model can detect positive cases by calculating the percentage of genuine positive predictions among all actual positive instances.

True Positives are instances correctly predicted as positive while False Negatives are instances incorrectly predicted as negative when they are actually positive.

Average Precision (AP) measures the average precision at different recall levels, focusing on the ranking of relevant documents. It’s useful for assessing how well a RAG system retrieves relevant information (Tamm et al. 2021).

Normalized Discounted Cumulative Gain (NDCG) valuates the quality of ranked retrieval results, considering both the relevance and the position of retrieved documents. It’s particularly useful for RAG as it accounts for the order of retrieval, which is crucial when passing context to language models.

Semantic Similarity evaluates how closely the meaning of the generated response matches the meaning of the relevant parts of the retrieved documents Risch et al. (2021). The semantic similarity evaluator on LlamaIndex takes all the ground-truth answers in the case of CovidQA dataset and compares them with the generated response. A threshold of 0.8 is set as a criteria for passing. Semantic similarity evaluation can be conducted using cosine similary, BERT-score or Sentence-BERT.

Response Time Time taken for the model to retrieve documents and generate an answer, measured in seconds.

Performance achieved in this study will be compared against the results from the study of Glass et al. (2021). The study also applied different indexing technique/RAG retrievers such as DPR and BM25 were used and evaluated based similar performance metrics of this study.

3.0 RESULTS AND DISCUSSION

3.1 Overview of Experiments

Experiments were conducted to analyze the impact of various indexing techniques, passage lengths, and embedding models on the performance of RAG systems in closeddomain QA using two datasets: CovidQA: small dataset with 25 questions and 102 documents focused on COVID-19. BioASQ: larger dataset with 707 questions (test split) and 40.2k documents covering broader biomedical topics. Parameters such as: indexing techniques, embedding models, passage length (using chunking) and top-k retrieved passages were varied while performance evaluation was conducted using various metrics, including hit rate, mean reciprocal rank (MRR), faithfulness, relevancy, and F1-score. The findings will be presented in relation to the hypotheses and objectives outlined in earlier chapters

3.2 Effect of Top-k Retrieved Passages

Table 2: Average Performance Metrics Across Top-k Values for conducted on DPR experiment of CovidQA dataset.

DOI: https://doi.org/10.57233/ijsgs.v11i2.872 ISSNp: 2488-9229; ISSNe: 3027-1118

IJSGS FUGUSAU VOLXXX WEBSITE: https://fugus-ijsgs.com.ng

top-k	hit rate	mrr	precision	recall	Ndcg	semantic similarity	Faithfulness	f1
3	0.71	0.56	0.40	0.31	0.41	0.55	0.71	0.19
5	0.83	0.59	0.36	0.43	0.38	0.56	0.76	0.20
9	0.92	0.60	0.29	0.56	0.33	0.56	0.78	0.18

Table 3: Average Performance Metrics Across Top-k Values for conducted on DPR experiment of BioASQ dataset

top-k	hit rate	mrr	precision	recall	ndcg	semantic similarity	Faithfulness	f1
3	0.88	0.81	0.64	0.35	0.67	0.70	0.62	0.20
5	0.91	0.82	0.57	0.43	0.61	0.72	0.66	0.22
9	0.93	0.82	0.47	0.53	0.53	0.74	0.70	0.23

Key Observations:

The search for optimal top-k value indicates trade-offs between different metrics as the top-k value changes, with larger k values generally improving recall and hit rate at the cost of precision and faithfulness. This has been explored in the study of Khanna and Subedi (2024) whereby large value of top-k tends to produce low precision since more false positives retrieval is made with higher k value. Table 2 and Table 3 shows how top-k affects experiments of this study and all illustration in this paper are made using topk = 5.

3.3 Impact of Passage Length on RAG Performance

Table 4: Average Performance Metrics Across Passage Length for conducted on DPR experiment of CovidQA dataset @ 5

Chunk	hit rate	mrr	semantic similarity	faithfulness	Relevancy	f1
128	0.80	0.60	0.54	0.75	0.93	0.18
512	0.87	0.59	0.57	0.83	0.84	0.18
1024	0.83	0.59	0.57	0.69	0.87	0.22
2048	0.83	0.59	0.57	0.76	0.89	0.20

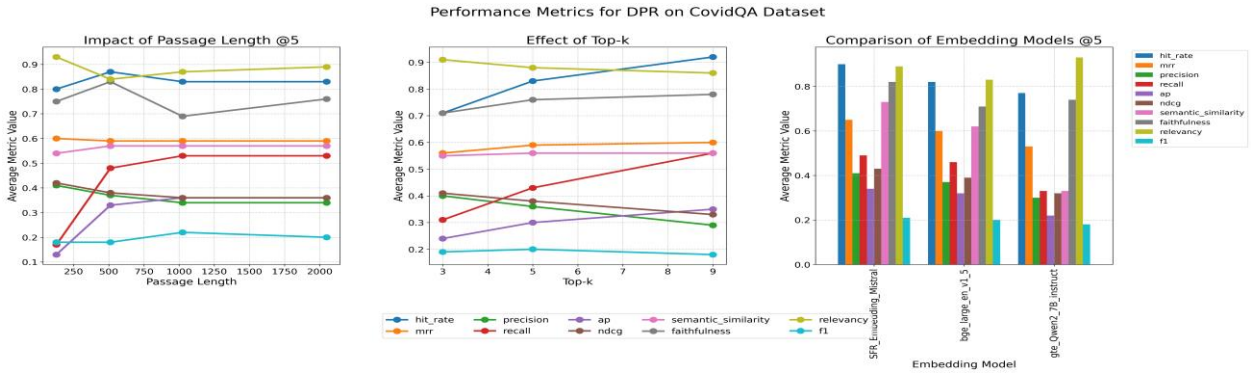

Figure 1: Average Performance Metrics Across Passage Length for conducted on DPR experiment of CovidQA dataset @ 5

Key Observations:

Figure 1 shows how passage length can dictate the performance of RAG QA system, particularly when observing the trend from 128 to 512 tokens per document which reveals a significant correlation between passage length and RAG performance.

Table 7 results indicate that increasing passage length generally improved RAG performance up to a certain point. The 1024-token chunks consistently outperformed shorter passages across all metrics. However, further increasing chunk size to 2048 tokens led to a slight decrease in performance, suggesting a potential information

DOI: https://doi.org/10.57233/ijsgs.v11i2.872 ISSNp: 2488-9229; ISSNe: 3027-1118

IJSGS FUGUSAU VOLXXX WEBSITE: https://fugus-ijsgs.com.ng

overload for the model.

This study found out that: Short Passages: Provided quick retrieval but often lacked sufficient context for generation of an accurate answer. Medium Passages: Provides balanced performance, supplying adequate context without significantly increasing retrieval time. Long Passages: Typically results in improved hit-rate however, at the cost of longer retrieval and generation time.

Table 5: Average Performance Metrics Across Passage Length for conducted on DPR experiment of BioASQ dataset @ 5

Chunk	hit rate	mrr	semantic similarity	faithfulness	relevancy	f1
128	0.91	0.83	0.69	0.66	0.97	0.19
512	0.90	0.82	0.73	0.66	0.92	0.23
1024	0.90	0.82	0.73	0.66	0.91	0.23
2048	0.90	0.82	0.73	0.66	0.91	0.22

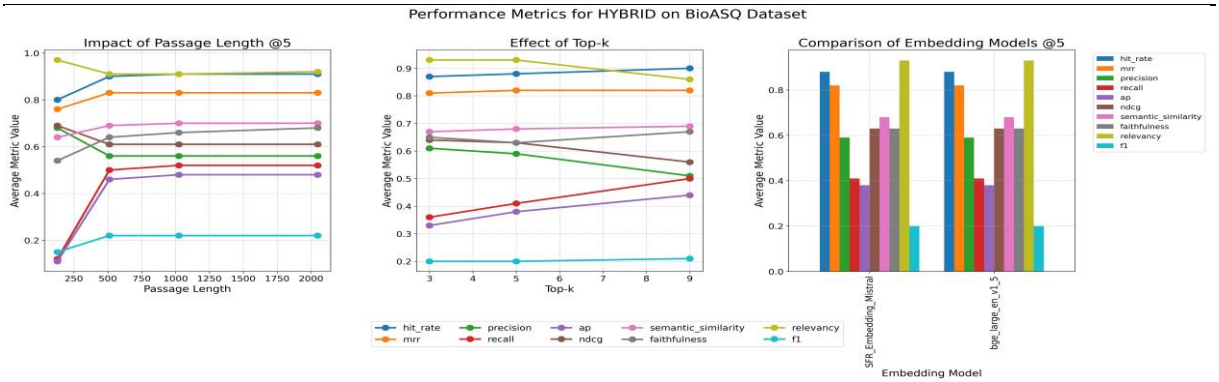

Figure 2: Passage Length for conducted on HYBRID experiment of BioASQ dataset @ 5

3.4 Comparison of Indexing Techniques

Table 6: Average Performance Metrics Across Indexing Techniques for conducted on DPR experiment of CovidQA dataset @ 5. This shows how retrieval/generation metrics very across various indexing techniques

Retriever	hit rate	mrr	Precision	recall	ndcg	faithfulness	Relevancy	f1
bm25	0.72	0.60	0.47	0.30	0.47	0.75	0.95	0.26
Dpr	0.83	0.59	0.36	0.43	0.38	0.76	0.88	0.20
Hybrid	0.72	0.60	0.47	0.30	0.47	0.75	0.92	0.24

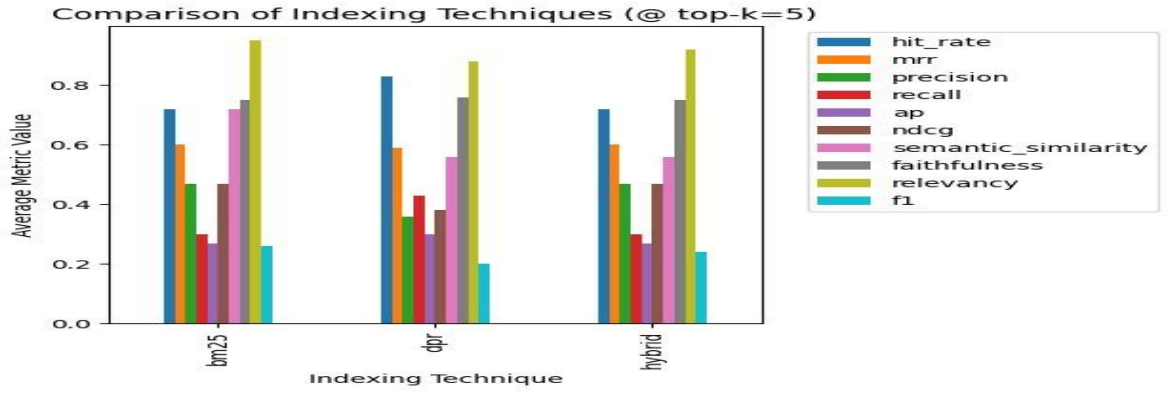

Figure 3: Average Performance Metrics Across Indexing Techniques for conducted on DPR experiment of CovidQA dataset @ 5

DOI: https://doi.org/10.57233/ijsgs.v11i2.872 ISSNp: 2488-9229; ISSNe: 3027-1118
IJSGS FUGUSAU VOLXXX WEBSITE: https://fugus-ijsgs.com.ng

Key Observations:
Three indexing techniques were evaluated namely: DPR, BM25, and a hybrid as shown in Table 7 and Table 8. It is observed that BM25 was best at generation with highest f1-score of 26% and 95% relevancy on CovidQA dataset indicating that retrieved documents highly relevant to the given queries. This shows that the keyword matching was not sufficient to retrieve relevant passages in this study. DPR was however, best at retrieval of documents showcasing it's ability to capture semantic relationships between queries and passages. All three indexing techniques have a reasonably good faithfulness score, suggesting that the generated responses aligns with retrieved documents/passages.

The hybrid in approach, which combined the strengths of BM25 and DPR was expected to have a novel performance, however it's best performance was on BioASQ dataset with 59% precision over 57% that of both DPR and BM25.

Table 7: Average Performance Metrics Across Indexing Techniques for conducted on DPR experiment of BioASQ dataset @ 5

Retriever	Hit Rate	MRR	Precision	Recall	NDCG	Faithfulness	Relevancy	F1
BM25	0.90	0.82	0.57	0.43	0.61	0.64	0.92	0.21
DPR	0.91	0.82	0.57	0.43	0.61	0.66	0.93	0.22
Hybrid	0.88	0.82	0.59	0.41	0.63	0.63	0.93	0.20

3.5 Embedding Model Comparison

Table 8: Average Performance Metrics Across Embedding Models for conducted on DPR experiment of CovidQA dataset @ 5

Embed	Hit_rate	Mrr	Semantic_similarity	faithfulness	relevancy	F1
SFR_Embedding_Mistral	0.90	0,65	0.73	0.82	0.89	0.21
Bge_large_v1.5	0.82	0.60	0.62	0.71	0.83	0.20
Gte_Qwen2_7B_instruc	0.77	0.53	0.33	0.74	0.93	0.18

Table 9: Average Performance Metrics Across Embedding Models for conducted on DPR experiment of BioASQ dataset @ 5

Embed	Hit_rate	Mrr	Semantic_similarity	faithfulness	relevancy	F1
SFR_Embedding_Mistral	0.91	0,82	0.75	0.66	0.93	0.22
Bge_large_v1.5	0.91	0.82	0.69	0.65	0.93	0.22

Key Observations:
Table 9 present embedding models against performance metrics where faithfulness decreases from SFR model down to GTE model. The SFR and BGE models showed scalability to the larger dataset in the transition from CovidQA to BioASQ, however, the SFR model maintained its performance advantage over BGE. The GTE model was resilient in terms of relevancy with a top score of 93% on the CovidQa dataset. BGE model's performance is outstanding and it's low disk usage of 1.25GB sets it at an advantage over other larger embedding models

3.6 Domain Adaptation Analysis

RAG system demonstrated strong performance on the closed-domain CovidQA dataset, achieving 92% hit-rate@9. When applied to the broader BioASQ dataset, performance increased slightly and became more robust, with 91% hit-rate@5 and 93% relevancy. This suggests effective domain adaptation, in the transition from a relatively small and narrow (closed-domain) to a wider domain such as biomedical.

DOI: https://doi.org/10.57233/ijsgs.v11i2.872 ISSNp: 2488-9229; ISSNe: 3027-1118

4.0 DISCUSSION

4.1 Implications for Hypothesis 1: Impact of Passage Length

Results support Hypothesis 1, with longer passages (512-1024 tokens) generally improving performance. However, very long passages (2048 tokens) showed diminishing returns, suggesting an optimal range exists. Observations shows that the peak performance for RAG system occurs when using between 512 and 1024 tokens per chunk size which helps in achieving an optimal balance between context and conciseness. Moderate passage lengths (512-1024 tokens) achieved the best balance of faithfulness and relevancy.

This finding aligns with previous research by Lewis et al. (2020) where 512 token was adopted as the optimal. The slight performance degradation at 2048 tokens indicates a potential information overload where excessive context begins to hinder rather than help the model's understanding. This finding has important consequences for designing RAG systems. It indicates that developers should target passage sizes in the range of 512 to 1024 tokens when setting up their systems. Doing so is likely to yield the best results for question-answering tasks within specific, limited domains.

These findings partially support Hypothesis 1, showing that increased passage length generally improves performance, but with diminishing returns beyond a certain threshold.

4.2 Implications for Hypothesis 2: Effectiveness of Indexing Techniques

This study confirms the second hypothesis, demonstrating that the choice of indexing technique significantly impacts RAG performance. The superior performance of DPR retriever in retrieval of relevant documents highlights the ability of the retriever to capture semantic relationships between queries and passage. The hybrid retriever demonstrates consistent performance at generation of response which shows that its performance it tilted more toward BM25 since it is also much better at generation than retrieval when compared to DPR. An 88% hit-rate@5 for hybrid is sill an achievement despite DPR having 91% hit-rate@5.

Queries and responses for this experiment were more scientific as opposes to plain English, perhaps BM25 is not good at retrieving scientific terms as compared to DPR. The findings of this study are consistent with previous research that highlights the limitations of traditional retrieval methods like BM25 in handling complex queries (Sawarkar et al., 2024).

4.3 Implications for Hypothesis 3: Domain Adaptation and Scalability

The successful application of top-performing models from CovidQA to BioASQ supports Hypothesis 3. Performance of SFR model remained strong on the larger dataset, indicating good scalability of the chosen techniques. The performance comparison between CovidQA and BioASQ datasets provides strong evidence for our third hypothesis regarding domain adaptation. The RAG system's ability to maintain 90% hit-rate@5 when transitioning from a narrow COVID-19 focus to broader biomedical topics demonstrates robust generalization capabilities. However, the observed performance drop (from 82% to 66% faithfulness) highlights the challenges of domain adaptation in RAG systems in the aspect of generation of responses that aligns with the retrieved passages especially in a closed domain QA task.

4.4 Impact of Top-k Retrieved Passages

The number of retrieved passages (top-k) showed clear effects on RAG performance. Higher k values improved hit rates and recall but decreased precision. This suggests a trade-off between comprehensive information retrieval and focused, relevant answers. These results highlight the importance of carefully tuning the k value based on specific application requirements and computational constraints and a value of 5 seems to be a good choice of k.

4.5 Embedding Model Performance Analysis

The comparison of embedding models reveals an interesting trade-off between performance and computational requirements. The larger models (GTE and SFR) required significantly more memory, which may limit their applicability in resource-constrained environments. These findings from this study suggests that SFR should be used over BGE and GTE however, to save cost, BGE model is also equal to the task which emphasize the importance of model selection based on both performance and practical constraints.

4.6 Limitations of the Study

The study was limited to two datasets in the biomedical domain, besides CovidQA dataset had only 25 question which is not enough to justify findings. Future work should explore other closed domains to test generalizability. Computational resources limited the scale of experiments on larger models, despite of accessing the school's cluster, experiments were still very slow. Further study with more powerful hardware could explore even larger models and datasets. The impact of different LLMs in the generation and evaluation phase was not explored and could be a fruitful area for future research or at all to use GPT-4. More sophisticated hybrid retrieval methods, including

learned rankers, could potentially improve performance further.

The focus on biomedical domains may not fully represent the challenges in other specialized fields. The study did not explore the impact of different language models in the generation phase, which could influence overall RAG performance.

4.7 Practical Implications and Recommendations

The choice of embedding model significantly impacts RAG performance, with larger, more recent models like SFR _Embedding Mistral showing clear advantages. Passage length is a critical factor, with a sweet spot around 512-1024 tokens balancing context and specificity. Hybrid retrieval approaches warrant further investigation, as they show potential to combine strengths of different methods in both retrieval and generation.

Therefore, practitioners in the field of RAG and closed domain QA should implore the findings of this study.

5.0 CONCLUSION

5.1 Conclusion

This study provides valuable insights and demonstrates the critical role of indexing techniques, passage length, and embedding model selection in optimizing RAG performance for closed-domain question answering. By carefully making the right selection, practitioners can significantly enhance the accuracy, relevance, and efficiency of their RAG systems, paving the way for more effective AI-powered information retrieval and question answering applications.

5.2 Recommendations for Future Work

Future work should explore other closed domains to test generalizability and better study of domain adaptation. Future research could explore the integration of other advanced retrieval techniques, such as neural retrievers or transformers-based models, to further enhance the performance of RAG systems since the current hybrid system used in this study did not emerge as the optimal performance. Additionally, the technique of calculating f1-score in this study made the metric have very low score despite the systems ability to make good generation. An example of an output scoring low is when the expected response is ”two”, but the model generates the number 2.

REFERENCES

Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. Evaluation of retrieval-augmented generation: A survey, 2024. URL https://arxiv.org/abs/ 2405.07437.

Jennifer Hsia, Afreen Shaikh, Zhiruo Wang, and Graham Neubig. Ragged: Towards informed design of retrieval augmented generation systems, 2024.

Julian Risch, Timo M¨oller, Julian Gutsch, and Malte Pietsch. Semantic answer similarity for evaluating question answering models, 2021. URL https://arxiv.org/abs/2108. 06130.

Konstantinos Andriopoulos and Johan Pouwelse. Augmenting llms with knowledge: A survey on hallucination prevention, 2023.

Kunal Sawarkar, Abhilasha Mangal, and Shivam Raj Solanki. Blended rag: Improving rag (retriever-augmented generation) accuracy with semantic search and hybrid querybased retrievers, 2024. URL https://arxiv.org/abs/2404.07220.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code, 2021.

Michael Glass, Gaetano Rossiello, and Alfio Gliozzo. Zero-shot slot filling with dpr and rag, 2021. Ruiqi Guo, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, and Sanjiv Kumar. Accelerating large-scale inference with anisotropic vector quantization, 2020.

Mike Lewis, Marjan Ghazvininejad, Gargi Ghosh, Armen Aghajanyan, Sida Wang, and Luke Zettlemoyer. Pre-training via paraphrasing, 2020.

Neal R. Mielke, Robert E. Frickey, Ivan Kalastirsky, Minyan Quan, Dmitry Ustinov, and Venkatesh J. Vasudevan. Reliability of solid-state drives based on nand flash memory. Proceedings of the IEEE, 105(9):1725–1750, 2017. doi: 10.1109/JPROC.2017.2725738.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models, 2023. URL https://arxiv.org/abs/2302.00083.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Ku¨ttler, Mike Lewis, Wen tau Yih, Tim Rockt¨aschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021.

Paulo Finardi, Leonardo Avila, Rodrigo Castaldoni, Pedro Gengo, Celio Larcher, Marcos Piau, Pablo Costa, and Vinicius Carid´a. The chronicles of rag: The retriever, the chunk and the generator, 2024. URL https://arxiv.org/abs/2401.07883.

Reda Yacouby and Dustin Axman. Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models.

In Steffen Eger, Yang Gao, Maxime Peyrard, Wei Zhao, and Eduard Hovy, editors, Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems, pages 79–91, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.eval4nlp-1.9. URL https://aclanthology.org/2020.eval4nlp-1.9.

Roi Blanco and Paolo Boldi. Extending bm25 with multiple query operators. In Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '12, page 921–930, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450314725. doi: 10.1145/2348283.2348406. URL https://doi.org/10.1145/2348283.2348406.

Sagar Imambi, Kolla Bhanu Prakash, and G. R. Kanagachidambaresan. PyTorch, pages 87–104. Springer International Publishing, Cham, 2021. ISBN 978-3-03057077-4. doi: 10.1007/978-3-030-57077-4 10. URL https://doi.org/10.1007/978-3-030-57077-4_10.

Sara Rosenthal, Avirup Sil, Radu Florian, and Salim Roukos. Clapnq: Cohesive longform answers from passages in natural questions for rag systems, 2024.

Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, and Suranga Nanayakkara. Fine-tune the entire rag architecture (including dpr retriever) for question-answering, 2021.

Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, and Suranga Nanayakkara. Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering, 2022.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-pack: Packaged resources to advance general chinese embedding, 2023.

Siva Reddy, Danqi Chen, and Christopher D. Manning. Coqa: A conversational question answering challenge, 2019. URL https://arxiv.org/abs/1808.07042.

Sujit Khanna and Shishir Subedi. Tabular embedding model (tem): Finetuning embedding models for tabular rag applications, 2024. URL https://arxiv.org/abs/2405. 01585.

Vladimir Karpukhin, Barlas O˘guz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. Dense passage retrieval for open-domain question answering, 2020.

Weijia Zhang, Mohammad Aliannejadi, Yifei Yuan, Jiahuan Pei, Jia-Hong Huang, and Evangelos Kanoulas. Towards fine-grained citation evaluation in generated text: A comparative analysis of faithfulness metrics, 2024. URL https://arxiv.org/abs/ 2406.15264.

Xueguang Ma, Kai Sun, Ronak Pradeep, and Jimmy Lin. A replication study of dense passage retriever, 2021.

Yan-Martin Tamm, Rinchin Damdinov, and Alexey Vasilev. Quality metrics in recommender systems: Do we calculate metrics consistently? In Fifteenth ACM Conference on Recommender Systems, RecSys '21. ACM, September 2021. doi: 10.1145/3460231.3478848. URL http://dx.doi.org/10.1145/3460231.3478848.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. arXiv preprint arXiv:2308.03281, 2023.

Zhengliang Liu, Xin Zhang, Shu Zhang, Xintao Hu, Tuo Zhang, Ning Qiang, Tianming Liu, and Bao Ge. Understanding llms: A comprehensive overview from training to inference, 2024.

Zhuyun Dai and Jamie Callan. Context-aware sentence/passage term importance estimation for first stage retrieval, 2019. URL https://arxiv.org/abs/1910.10687.