2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR) | 979-8-3503-5142-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/MIPR62202.2024.00031

Blended RAG: Improving RAG (Retriever-Augmented Generation) Accuracy with Semantic Search and Hybrid Query-Based Retrievers

1st Kunal Sawarkar
IBM
Kunal@ibm.com

2nd Abhilasha Mangal
IBM
Abhilasha.Mangal@ibm.com

3rd Shivam Raj Solanki
IBM
Shivam.Raj.Solanki@ibm.com

***Abstract*—Retrieval-Augmented Generation (RAG) is a prevalent approach to infuse a private knowledge base of documents with Large Language Models (LLM) to build Generative Q&A (Question-Answering) systems. However, RAG accuracy becomes increasingly challenging as the corpus of documents scales up, with Retrievers playing an outsized role in the overall RAG accuracy by extracting the most relevant document from the corpus to provide context to the LLM. In this paper, we propose the 'Blended RAG' method of leveraging semantic search techniques, such as Dense Vector indexes and Sparse Encoder indexes, blended with hybrid query strategies. Our study achieves better retrieval results and sets new benchmarks for IR (Information Retrieval) datasets like NQ and TREC-COVID datasets. We further extend such a 'Blended Retriever' to the RAG system to demonstrate far superior results on Generative Q&A datasets like SQUAD, even surpassing fine-tuning performance.**

***Index Terms*—RAG, Retrievers, Semantic Search, Dense Index, Vector Search**

I. Introduction

RAG represents an approach to text generation that is based not only on patterns learned during training but also on dynamically retrieved external knowledge [1]. This method combines the creative flair of generative models with the encyclopedic recall of a search engine. The efficacy of the RAG system relies fundamentally on two components: the Retriever (R) and the Generator (G), the latter representing the size and type of LLM.

The language model can easily craft sentences, but it might not always have all the facts. This is where the Retriever (R) steps in, quickly sifting through vast amounts of documents to find relevant information that can be used to inform and enrich the language model's output. Think of the retriever as a researcher part of the AI, which feeds the contextually grounded text to generate knowledgeable answers to Generator (G). Without the retriever, RAG would be like a well-spoken individual who delivers irrelevant information.

II. Related Work

Search has been a focal point of research in information retrieval, with numerous studies exploring various methodologies. Historically, the BM25 (Best Match) algorithm, which uses similarity search, has been a cornerstone in this field, as explored by Robertson and Zaragoza (2009). [2]. BM25 prioritizes documents according to their pertinence to a query, capitalizing on Term Frequency (TF), Inverse Document Frequency (IDF), and Document Length to compute a relevance score.

Dense vector models, particularly those employing KNN (k Nearest Neighbours) algorithms, have gained attention for their ability to capture deep semantic relationships in data. Studies by Johnson et al. (2019) demonstrated the efficacy of dense vector representations in large-scale search applications. The kinship between data entities (including the search query) is assessed by computing the vectorial proximity (via cosine similarity etc.). During search execution, the model discerns the 'k' vectors closest in resemblance to the query vector, hence returning the corresponding data entities as results. Their ability to transform text into vector space models, where semantic similarities can be quantitatively assessed, marks a significant advancement over traditional keyword-based approaches. [3]

On the other hand, sparse encoder based vector models have also been explored for their precision in representing document semantics. The work of Zaharia et al. (2010) illustrates the potential of these models in efficiently handling high-dimensional data while maintaining interpretability, a challenge often faced in dense vector representations. In Sparse Encoder indexes the indexed documents, and the user's search query maps into an extensive array of associated terms derived from a vast corpus of training data to encapsulate relationships and contextual use of concepts. The resultant expanded terms for documents and queries are encoded into sparse vectors, an efficient data representation format when handling an extensive vocabulary.

A. Limitations in the current RAG system

Most current retrieval methodologies employed in Retrieval-Augmented Generation (RAG) pipelines rely on keyword and similarity-based searches, which can restrict the RAG system's overall accuracy. Table 1 provides a summary of the current benchmarks for retriever accuracy.

2770-4319/24/$31.00 ©2024 IEEE
DOI 10.1109/MIPR62202.2024.00031

TABLE I: Current Retriever Benchmarks

Dataset	Benchmark Metrics	NDCG@10	p@20	F1
NQDataset	P@20	0.633	86	79.6
Trec Covid	NDCG@10	80.4		
HotpotQA	F1 , EM			0.85

While most of prior efforts in improving RAG accuracy is on G part, by tweaking LLM prompts, tuning etc.,[9] they have limited impact on the overall accuracy of the RAG system, since if R part is feeding irreverent context then answer would be inaccurate. Furthermore, most retrieval methodologies employed in RAG pipelines rely on keyword and similarity-based searches, which can restrict the system's overall accuracy.

Finding the best search method for RAG is still an emerging area of research. The goal of this study is to enhance retriever and RAG accuracy by incorporating Semantic Search-Based Retrievers and Hybrid Search Queries.

III. Blended Retrievers

For RAG systems, we explored three distinct search strategies: keyword-based similarity search, dense vector-based, and semantic-based sparse encoders, integrating these to formulate hybrid queries. Unlike conventional keyword matching, semantic search delves into the nuances of a user's query, deciphering context and intent. This study systematically evaluates an array of search techniques across three primary indices: BM25 [4] for keyword-based, KNN [5] for vector-based, and Elastic Learned Sparse Encoder (ELSER) for sparse encoder-based semantic search.

1) BM25 Index: The BM25 index is adept at employing full-text search capabilities enhanced by fuzzy matching techniques, laying the groundwork for more sophisticated query operations.
2) Dense Vector Index: We construct a dense vector index empowered by sentence transformers. It identifies the proximity of vector representations derived from document and query content.
3) Sparse Encoder Index: The Sparse EncodeR Retriever Model index is an amalgam of semantic understanding and similarity-based retrieval to encapsulate the nuanced relationships between terms, thereby capturing a more authentic representation of user intent and document relevance.

A. Methodology

Our methodology unfolds in a sequence of progressive steps, commencing with the elementary match query within the BM25 index. We then escalate to hybrid queries that amalgamate diverse search techniques across multiple fields, leveraging the multi-match query within the Sparse Encoder-Based Index. This method proves invaluable when the exact location of the query text within the document corpus is indeterminate, hence ensuring a comprehensive match retrieval.

The multi-match queries are categorized as follows:

- Cross Fields: Targets concurrence across multiple fields
- Most Fields: Seeks text representation through different lenses across various fields.
- Best Fields: Pursues the aggregation of words within a singular field.
- Phrase Prefix: Operates similarly to Best Fields but prioritizes phrases over keywords.

After initial match queries, we incorporate dense vector (KNN) and sparse encoder indices, each with their bespoke hybrid queries. This strategic approach synthesizes the strengths of each index, channeling them towards the unified goal of refining retrieval accuracy within our RAG system. We calculate the top-k retrieval accuracy metric to distill the essence of each query type.

In Figure 1, we introduce a scheme designed to create Blended Retrievers by blending semantic search with hybrid queries.

B. Constructing RAG System

From the plethora of possible permutations, a select sextet (top 6) of hybrid queries—those exhibiting paramount retrieval efficacy—were chosen for further scrutiny. These queries were then subjected to rigorous evaluation across the benchmark datasets to ascertain the precision of the retrieval component within RAG. The sextet queries represent the culmination of retriever experimentation, embodying the synthesis of our finest query strategies aligned with various index types. The six blended queries are then fed to generative question-answering systems. This process finds the best retrievers to feed to the Generator of RAG, given the exponential growth in the number of potential query combinations stemming from the integration with distinct index types.

The intricacies of constructing an effective RAG system are multi-fold, particularly when source datasets have diverse and complex landscapes. We undertook a comprehensive evaluation of a myriad of hybrid query formulations, scrutinizing their performance across benchmark datasets, including the Natural Questions (NQ), TREC-COVID, Stanford Question Answering Dataset (SqUAD), and HotPotQA.

IV. Experimentation for Retriever Evaluation

We used top-10 retrieval accuracy to narrow down the six best types of blended retrievers (index + hybrid query) for comparison for each benchmark dataset.

1) Top-10 retrieval accuracy on the NQ dataset : For the NQ dataset [6], our empirical analysis has demonstrated the superior performance of hybrid query strategies, attributable to the ability to utilize multiple data fields effectively. In Figure 2, our findings reveal that the hybrid query approach employing the **Sparse Encoder with Best Fields** attains the highest retrieval accuracy, reaching an impressive 88.77%. This result surpasses the efficacy of all other formulations, establishing a new benchmark for retrieval tasks within this dataset.

2) Top-10 Retrieval Accuracy on TREC-Covid dataset: For the TREC-COVID dataset [7], which encompasses relevancy scores spanning from -1 to 2, with -1 indicative of irrelevance

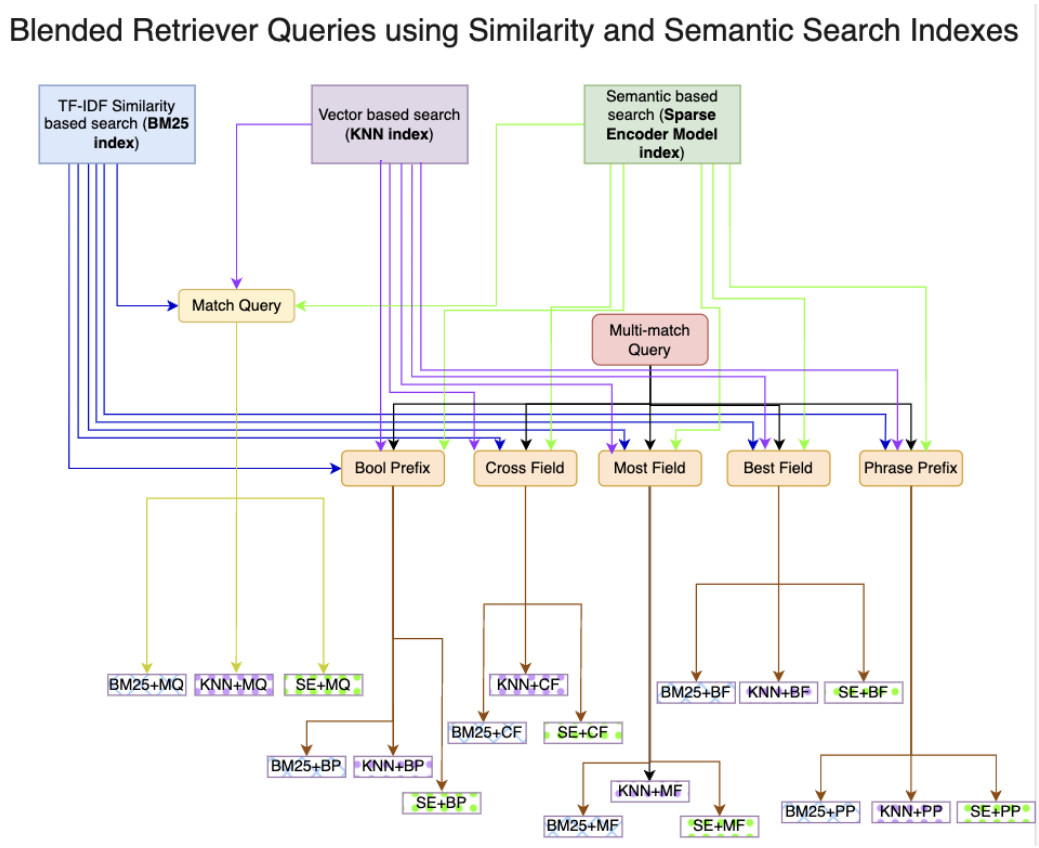

Fig. 1: Scheme of Creating Blended Retrievers using Semantic Search with Hybrid Queries.

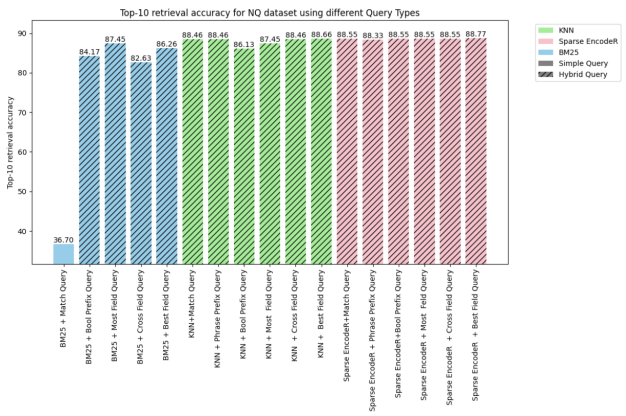

Fig. 2: Top-10 Retriever Accuracy for NQ Dataset

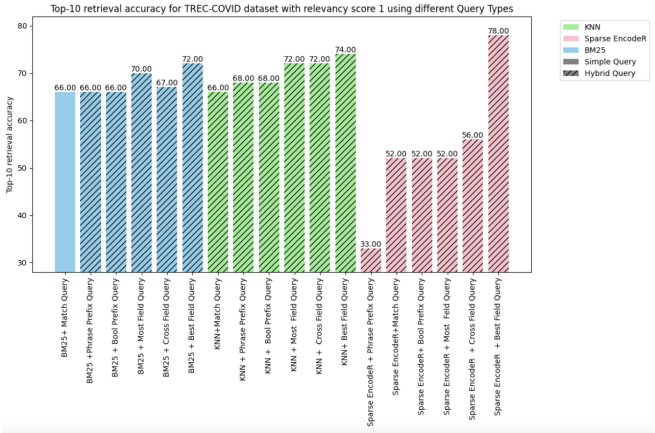

Fig. 3: Top 10 retriever accuracy for Trec-Covid Score-1

and 2 denoting high relevance, our initial assessments targeted documents with a relevancy of 1, deemed partially relevant.

Figure 3 analysis reveals a superior performance of vector search hybrid queries over those based on keywords. In particular, hybrid queries that leverage the **Sparse EncodeR utilizing Best Fields** demonstrate the highest efficacy across all index types at 78% accuracy.

Subsequent to the initial evaluation, the same spectrum of queries was subjected to assessment against the TREC-COVID dataset with a relevancy score of 2, denoting that the documents were entirely pertinent to the associated queries. Figure 4 illustrated with a relevance score of two, where documents fully meet the relevance criteria for associated queries, reinforce the efficacy of vector search hybrid queries

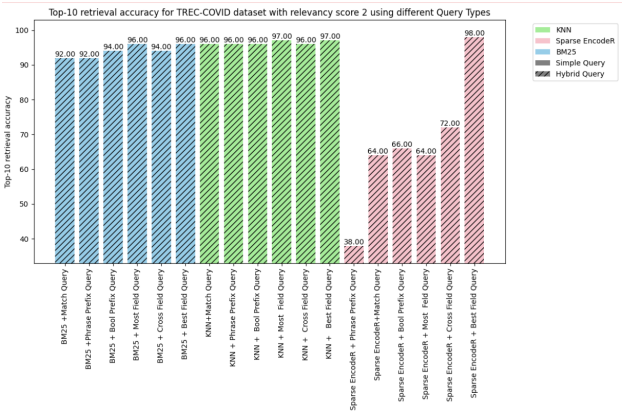

Fig. 4: Top 10 retriever accuracy for Trec-Covid Score-2

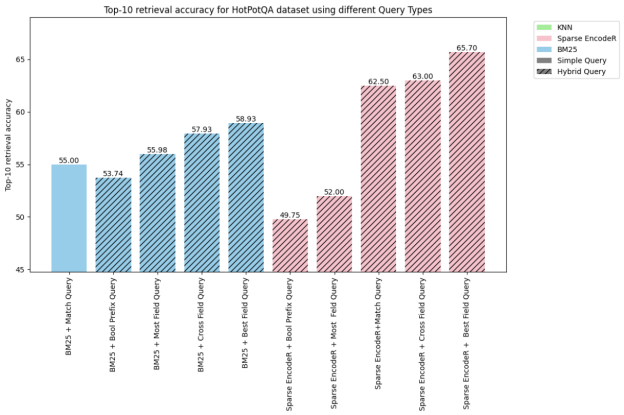

Fig. 5: Top 10 retriever accuracy for HotPotQA dataset

over conventional keyword-based methods. Notably, the hybrid query incorporating **Sparse Encoder with Best Fields** demonstrates a 98% top-10 retrieval accuracy, eclipsing all other formulations. This suggests that a methodological pivot towards more nuanced blended search, particularly those that effectively utilize the Best Fields, can significantly enhance retrieval outcomes in information retrieval (IR) systems.

3) Top-10 Retrieval Accuracy on the HotPotQA dataset : The HotPotQA [8] dataset, with its extensive corpus of over 5M documents and a query set comprising 7,500 items, presents a formidable challenge for comprehensive evaluation due to compute requirements. Consequently, the assessment was confined to a select subset of hybrid queries. Despite these constraints, the analysis provided insightful data, as reflected in the accompanying visualization in Figure 5.

Figure 5 shows that hybrid queries, specifically those utilizing Cross Fields and Best Fields search strategies, demonstrate superior performance. Notably, the hybrid query that blends Sparse EncodeR with Best Fields queries achieved the highest efficiency, of 65.70% on the HotPotQA dataset.

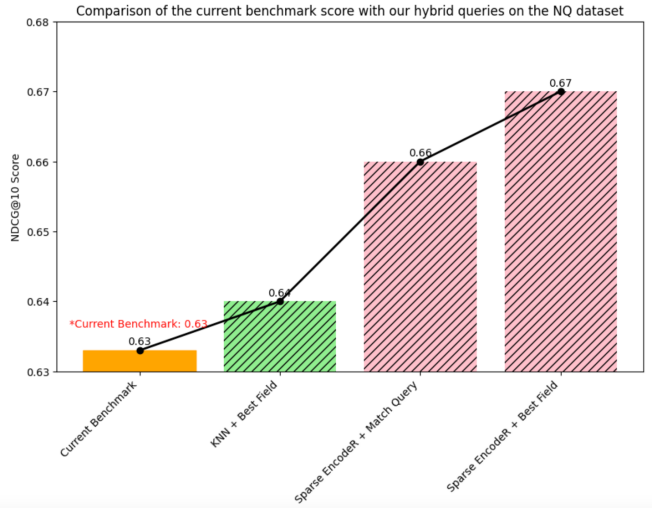

Fig. 6: NQ dataset Benchmarking using NDCG@10 Metric

TABLE II: Retriever Benchmarking using NDCG@10 Metric

Dataset	Model/Pipeline	NDCG@10
Trec-covid	COCO-DR Large	0.804
Trec-covid	Blended RAG	0.87
NQ dataset	monoT5-3B	0.633
NQ dataset	Blended RAG	0.67

A. Retriever Benchmarking

Now that we have identified the best set of combinations of Index + Query types, we will use these sextet queries on IR datasets for benchmarking using NDCG@10 [9] scores (Normalised Discounted Cumulative Gain metric).

1) NQ dataset benchmarking: The results for NDCG@10 using sextet queries and the current benchmark on the NQ dataset are shown in the chart Figure 7. Our pipeline provides the best NDCG@10 score of 0.67, which is 5.8% higher than the current benchmark score of 0.633 achieved by the monoT5-3B model. Table II shows that all semantic search-based hybrid queries outperform the current benchmark score, which indicates that our hybrid queries are a better candidate for developing the RAG pipeline.

2) TREC-Covid Dataset Benchmarking : In our research, the suite of hybrid queries devised has demonstrably exceeded the current benchmark of 0.80 NDCG@10 score, signaling their superior candidature for the RAG pipeline. Figure 7 shows the results for NDCG@10 using sextet queries. Blended Retrievers achieved an NDCG@10 score of 0.87, which marks an 8.2% increment over the benchmark score of 0.804 established by the COCO-DR Large model (Table II).

3) SqUAD Dataset Benchmarking: The SqUAD (Stanford Question Answering Dataset) [10] is not an IR dataset, but we evaluated the retrieval accuracy of the SquAD dataset for consistency. Firstly, we created a corpus from the SqUAD dataset using the title and context fields in the dataset. Then, we indexed the corpus using BM25, dense vector, and Sparse Encoder. The top-k (k=5,10, and 20) retrieval accuracy results

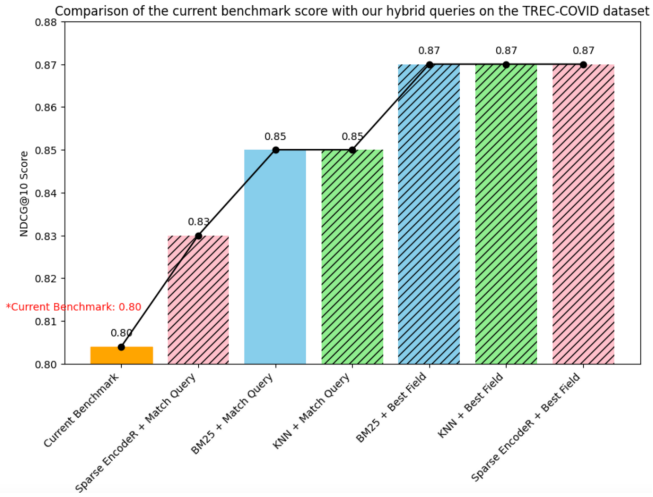

Fig. 7: TREC-Covid Dataset Benchmarking using NDCG@10 Metric

for the SqUAD dataset are calculated. Table III illustrates that for SQuAD, dense vector (KNN)-based semantic searches achieve higher accuracy than sparse vector-based semantic searches and traditional similarity-based searches, particularly for top-k retrieval performance with k values of 5, 10, and 20. (See Appendix for more details)

B. Summary of Retriever Evaluation

We evaluated the retrieval accuracy using our approach, quantified by Top-k metrics where $k \in \{5, 10, 20\}$, across NQ, TREC-COVID, SQUAD, and CoQA datasets. This synopsis demonstrates the capability of our **Blended Retrieval** methodology within diverse informational contexts. Key observations are

- Enhanced retrieval accuracy is exhibited in all datasets except for CoQA [11]. This enhancement is attributable to the capability of our hybrid queries to effectively utilize available metadata to source the most pertinent results.
- Implementing dense vector-based (KNN) semantic search results in a marked improvement over keyword-based search approaches.
- Employing semantic search-based hybrid queries realizes better retrieval precision compared to all conventional keyword-based or vector-based searches.
- Furthermore, it is discernible that the Sparse Encoder-based semantic search, when amalgamated with the 'Best Fields' hybrid query, often provides superior results than any other method.

V. RAG Experimentation

From the retriever evaluation experiments, we know the best retriever, i.e., the best combination of indices + query. In this section, we extend this knowledge to evaluate the RAG pipeline. To avoid the effect of LLM size or type, we perform all experiments using FLAN-T5-XXL.

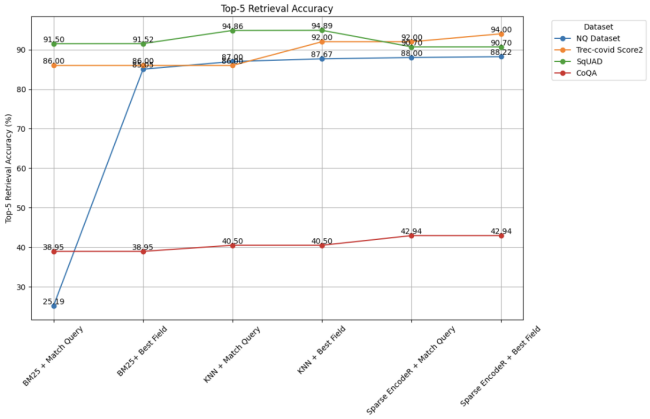

Fig. 8: Top-5 Retrieval Accuracy across Datasets

A. RAG Evaluation on the SqUAD Dataset

SqUAD is a commonly bench-marked dataset for RAG systems or Generative Q&A using LLMs. Our study juxtaposes three variations of the RAG pipeline from prior work using the evaluation metrics of Exact Match (EM) and F1 scores to gauge the accuracy of answer generation, as well as Top-5 and Top-10 for retrieval accuracy.

- RAG-original [12]: This variant, a model fine-tuned on the Natural Questions dataset, has been appraised without domain-specific adaptation.
- RAG-end2end [12]: As an extension of RAG-original, this model undergoes additional fine-tuning, tailored for domain adaptation to the SQuAD.
- Blended RAG: Distinctively, our Blended RAG variant has not undergone training on the SQuAD dataset or any related corpora. It harnesses an optimized amalgamation of field selections and hybrid query formulations with semantic indices to feed LLMs to render the most precise responses possible.

Consequently, as shown in Table IV, our Blended RAG showcases enhanced performance for Generative Q&A with F1 scores higher by 50%, even without dataset-specific fine-tuning. This characteristic is particularly advantageous for large enterprise datasets, where fine-tuning may be impractical or unfeasible, underscoring this research's principal application.

B. RAG Evaluation on the NQ Dataset

Natual Questions (NQ) is another commonly studied dataset for RAG. The Blended RAG pipeline, utilizing zero-shot learning, was evaluated to ascertain its efficacy against other non-fine-tuned models. The assessment focused on the following metrics: Exact Match (EM), F1 Score, and retrieval accuracy (Top-5 and Top-20) in Table V.

Blended RAG (Zero-shot): Demonstrated superior performance with an EM of 42.63, improving the prior benchmark by 35%.

TABLE III: Blended Retriever Performance SqUAD Dataset

SqUAD	BM25+MQ	BM25+BF	KNN+MQ	KNN+BF	SPARSE ENCODER+MQ	SPARSE ENCODER+BF
Top-5	91.5	91.52	94.86	94.89	90.7	90.7
Top-10	94.43	94.49	97.43	97.43	94.13	94.16
Top-20	96.3	96.36	98.57	98.58	96.49	96.52

TABLE IV: Evaluation of the RAG Pipeline on the SquAD Dataset

Model/Pipeline	EM	F1	Top-5	Top-20
RAG-original	28.12	39.42	59.64	72.38
RAG-end2end	40.02	52.63	75.79	85.57
Blended RAG	57.63	68.4	94.89	98.58

TABLE V: Evaluation of the RAG pipeline on the NQ dataset

Model/Pipeline	EM	F1	Top-5	Top-20
GLaM (Oneshot) [13]	26.3			
GLaM (Zeroshot) [13]	24.7			
PaLM540B (Oneshot) [14]	29.3			
Blended RAG (Zero-shot)	42.63	53.96	88.22	88.88

VI. Discussion

While RAG is a commonly used approach in the industry, we realized during the course of this study that various challenges still exist, like there are no standard datasets on which both R (Retriever) and RAG benchmarks are available. Retriever is often studied as a separate problem in the IR domain, while RAG is studied in the LLM domain. We thus attempted to bring synergy between the two domains with this work. In this section, we share some learning on limitations and appropriate use of this method.

A. Trade-off between Sparse and Dense Vector Indices

The HotPotQA corpus presents substantial computational challenges with 5M documents, generating a dense vector index to an approximate size of 50GB, a factor that significantly hampers processing efficiency. Dense vector indexing, characterized by its rapid indexing capability, is offset by a relatively sluggish querying performance. Conversely, sparse vector indexing, despite its slower indexing process, offers expeditious querying advantages. Furthermore, a stark contrast in storage requirements is observed; for instance, the sparse vector index of the HotPotQA corpus occupied a mere 10.5GB as opposed to the 50GB required for the dense vector equivalent.

In such cases, we recommend sparse encoder indexes. Furthermore, for enterprises with this volume, we found it better to use multi-tenancy with federated search queries.

B. Blended Retrievers without Metadata

When datasets are enriched with metadata or other relevant informational facets, they improve the efficacy of blended retrievers. Conversely, for datasets devoid of metadata, such as CoQA, it is not as impressive. You can see the results in Table VI.

TABLE VI: Top-5 retrieval accuracy CoQA Dataset

COQA	BM25+MQ	BM25+BF	KNN+MQ	KNN+BF	SE+MQ	SE+BF
Top-5	45.3	45.3	47.56	47.56	49.94	49.94

The absence of metadata in the CoQA dataset resulted in hybrid queries offering no improvement over basic queries. This limitation underscores the critical role of metadata in enhancing the efficacy of complex query structures. However, Sparse Encoder-based semantic searches still yield the most favorable outcomes than traditional methods.

Additionally, we would like to note that while NDCG@10 scores for Retriever and F1,EM scores for RAG are commonly used metrics, we found them to be poor proxies of Generative Q&A systems for human alignment. Better metrics to evaluate the RAG system is a key area of future work.

VII. Conclusion

Blended RAG pipeline is highly effective across multiple datasets despite not being specifically trained on them. Notably, this approach does not necessitate exemplars for prompt engineering which are often required in few-shot learning, indicating a robust generalization capability within the zero-shot paradigm. This study demonstrated:

- Optimization of R with Blended Search: Incorporating Semantic Search, specifically Sparse Encoder indices coupled with 'Best Fields' queries, has emerged as the superior construct across all, setting a new benchmark of 87% for Retriever Accuracy on TREC-COVID.
- Enhancement of RAG via Blended Retrievers: The significant amplification in retrieval accuracy is particularly pronounced for the overall evaluation of the RAG pipeline, surpassing prior benchmarks on fine-tuned sets by a wide margin. Blended RAG sets a new benchmark at 68% F1 Score on SQUAD and 42% EM Score on NQ dataset; for non-tuned Q&A systems.

The empirical findings endorse the potency of Blended Retrievers in refining RAG systems beyond focusing on LLM size & type, getting better results with relatively smaller LLM and thus setting a foundation for more intelligent and contextually aware Generative Q&A systems.

ACKNOWLEDGMENT

Authors would like to acknowledge the below members for making this study possible.

- **IBM Ecosystem** The authors conducted this study while employed at IBM Ecosystem. They would like to express their gratitude to the Ecosystem team and leadership for their support in carrying out this work.
- **IBM Research** The authors have received generous feedback on their work from colleagues at IBM Research, particularly Radu Florian, whom the authors would like to acknowledge.
- **Elastic** - The authors have been granted access to the Elastic Search platform and ELSER index as an embodiment of sparse index. They would like to thank Elastic for their support.

REFERENCES

[1] T. Merth, Q. Fu, M. Rastegari, and M. Najibi, "Superposition prompting: Improving and accelerating retrieval-augmented generation," *arXiv preprint arXiv:2404.06910*, 2024.

[2] S. Robertson and H. Zaragoza, "The bm25 algorithm," *Foundations and Trends in Information Retrieval*, 2009.

[3] M. Johnson *et al.*, "Knn algorithms for semantic search," in *Proceedings of the International Conference on Machine Learning*, 2019.

[4] G. Amati, *BM25*, pp. 257–260. Boston, MA: Springer US, 2009.

[5] K. Taunk, S. De, S. Verma, and A. Swetapadma, "A brief review of nearest neighbor algorithm for learning and classification," in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, pp. 1255–1260, 2019.

[6] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, M. Kelcey, J. Devlin, K. Lee, K. N. Toutanova, L. Jones, M.-W. Chang, A. Dai, J. Uszkoreit, Q. Le, and S. Petrov, "Natural questions: a benchmark for question answering research," *Transactions of the Association of Computational Linguistics*, 2019.

[7] L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Burdick, D. Eide, K. Funk, Y. Katsis, R. Kinney, *et al.*, "Cord-19: The covid-19 open research dataset," *ArXiv*, 2020.

[8] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. W. Cohen, R. Salakhutdinov, and C. D. Manning, "Hotpotqa: A dataset for diverse, explainable multi-hop question answering," *arXiv preprint arXiv:1809.09600*, 2018.

[9] Y. Wang, L. Wang, Y. Li, D. He, and T.-Y. Liu, "A theoretical analysis of ndcg type ranking measures," in *Conference on learning theory*, pp. 25–54, PMLR, 2013.

[10] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100,000+ questions for machine comprehension of text," *arXiv preprint arXiv:1606.05250*, 2016.

[11] S. Reddy, D. Chen, and C. D. Manning, "Coqa: A conversational question answering challenge," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 249–266, 2019.

[12] S. Siriwardhana, R. Weerasekera, E. Wen, T. Kaluarachchi, R. Rana, and S. Nanayakkara, "Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1–17, 2023.

[13] N. Du, Y. Huang, A. M. Dai, S. Tong, D. Lepikhin, Y. Xu, M. Krikun, Y. Zhou, A. W. Yu, O. Firat, *et al.*, "Glam: Efficient scaling of language models with mixture-of-experts," in *International Conference on Machine Learning*, pp. 5547–5569, PMLR, 2022.

[14] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel, "Palm: Scaling language modeling with pathways," *Journal of Machine Learning Research*, vol. 24, no. 240, pp. 1–113, 2023.