Retrieval Performance in RAG Systems: A Component-Level Evaluation Framework

Alexander Kreß*, Alexander Lawall*, Thomas Zöller*
*IU International University of Applied Science
Erfurt, Germany
alexander.kress@iu-study.org, alexander.lawall@iu.org, thomas.zoeller@iu.org

***Abstract*—Retrieval-Augmented Generation (RAG) systems are relevant for improving factuality in Large Language Model (LLM) outputs, yet their evaluation remains challenging due to their multi-component architecture. This paper introduces plot-RAG (pRAG), a novel evaluation framework that visualizes component-level performance in RAG systems, providing granular insights into retrieval and re-ranking processes, without requiring resource-intensive LLM-based evaluation. The effectiveness of pRAG is demonstrated by analyzing a real-world technical documentation question-answering system. Additionally, the methodology for generating and validating synthetic evaluation datasets is presented, showing they can match or exceed manually prepared datasets for RAG assessment. The experiments confirm that the retrieval component represents the most critical performance bottleneck in RAG systems, and a formula is provided to determine the optimal retrieval size based on response time requirements. These contributions enable a more efficient and targeted evaluation of RAG systems, particularly in specialized domains where the creation of ground truth data typically requires substantial expert involvement.**

***Index Terms*—retrieval-augmented generation; evaluation framework; synthetic datasets; component-level analysis.**

I. Introduction

Retrieval-Augmented Generation (RAG) systems are important for improving the factuality and reliability of Large Language Model (LLM) outputs, especially in domain-specific applications. Despite their adoption, evaluating these systems remains challenging, particularly when considering their multi-component nature and varying performance across different use cases [1].

A. Motivation and Problem Statement

The evaluation of RAG systems faces several challenges that current LLM approaches fail to adequately address. While LLMs alone can be evaluated using established benchmarks, RAG systems introduce additional complexities due to their multi-stage architecture spanning document processing, retrieval, re-ranking, and generation components [2][3]. As noted by [2], dynamic data environments further complicate evaluation, as the underlying knowledge sources often change over time.

Current evaluation frameworks typically produce aggregate metrics that mask the performance of individual components, making it difficult to identify specific bottlenecks or optimization opportunities [2][4]. Manual evaluation methods are becoming increasingly inefficient, necessitating automated approaches that can scale with system complexity. Additionally, temporal aspects of RAG performance — such as latency variations across different technical configurations — are rarely incorporated into evaluation methodologies despite their critical importance in real-world applications [2][5].

Established benchmark datasets like HotpotQA [6] and MS MARCO [7] have proven inadequate for evaluating modern RAG systems [8], as they fail to capture the nuanced retrieval and generation scenarios encountered in specialized domains. The availability of ground truth data will become rare in the future [9]. While synthetic dataset generation offers promising alternatives [10], systematic approaches for validating these datasets and incorporating them into holistic evaluation frameworks remain underdeveloped.

B. Research Gap

Despite the proliferation of evaluation methods for RAG systems, major gaps persist in current approaches. Existing frameworks like RAGAS [11] rarely provide granular insights into component-level performance, instead focusing on end-to-end evaluation that obscures the contribution of individual technical elements [2][3]. As [12] observes, the various technical alternatives available at each stage of the RAG pipeline create a complex evaluation space that remains largely unexplored. [13] mentioned that this gap cannot be closed by asking LLMs for reasoning.

The role of re-ranking models [14][15] and hybrid retrieval techniques like the combination of embeddings and BM25 [16] in RAG performance is inadequately addressed by current evaluation approaches. Furthermore, while the importance of synthetic datasets for evaluation is increasingly recognized [17], methodologies for generating and validating these datasets remain ad-hoc and not standardized. These gaps demand a comprehensive evaluation framework that addresses both the technical complexity of RAG systems and the practical challenges of meaningful assessment [10].

C. Research Questions

This paper addresses the primary research question: “Which technical concepts are necessary to successfully evaluate RAG systems?”. Secondary research questions are investigated to explore this question more comprehensively:

1) “How can we effectively evaluate the retrieval component in RAG systems?”
2) “How can synthetic datasets be efficiently generated and validated for RAG evaluation?”
3) “What approaches show promise for evaluating the entire RAG pipeline?”

D. Contributions

This paper makes several contributions to the field of RAG system evaluation:

- The introduction of a methodology for generating and validating synthetic evaluation datasets that can scale efficiently across domains and use cases.
- The development of a visualization approach (pRAG) for assessing retrieval component performance that incorporates both quality metrics and temporal analysis.
- The provision of empirical findings from applying the framework to a real-world RAG system designed for technical documentation question answering.

The framework addresses critical gaps in existing evaluation approaches by offering a more granular, component-specific assessment methodology that can adapt to the evolving landscape of RAG system design.

E. Paper Structure

The remainder of this paper is organized as follows: Section II describes the methodology, including the architecture of the evaluation framework, synthetic dataset generation approach, and component-specific assessment techniques. Section III presents the results of applying the framework to a case study RAG system and discusses key findings and implications. Finally, Section IV concludes the paper and outlines directions for future work.

II. Methodology

A. RAG System Architecture

The RAG system employs a microservice-based architecture designed for scalability and modular development. The system processes user queries through these key components: When a user submits a query via the frontend, the middleware API coordinates the workflow. First, the pre-processing API generates keywords and embeddings from the query for semantic comparison. These are passed to a vector handling API that performs hybrid retrieval, combining BM25 [18], keyword matching, and embedding-based semantic search through paradeDB, a PostgreSQL extension supporting vector operations.

Retrieved contexts and metadata flow back to the middleware API, which forwards them to the pre-processing API where a cross-encoder re-ranker prioritizes the most semantically relevant documents. Finally, these re-ranked contexts together with an initial prompt are provided to a LLM that generates a comprehensive response based on the available information and returns it to the user via the frontend.

B. System Implementation

The system is deployed on a Kubernetes cluster with the frontend developed in React and backend services in Python. For the knowledge base, 50 technical documents from HORSCH machinery manuals using Azure Document Intelligence are processed to convert PDF content into processable text.

For embedding generation, the Hugging Face multi-qa-MiniLM-L6-cos-v1 Sentence Transformer model [19] was implemented, selected for its balance of English language capabilities and computational efficiency. Documents were chunked to match the model's maximum token length and stored with machine-specific metadata.

We evaluated two cross-encoder models for re-ranking: msmarco-MiniLM-L6-en-dev1 [20] and ms-marco-MiniLM-L-6-v2 [21], which reorder retrieved contexts based on query relevance. For response generation, we utilized ChatGPT-3.5-Turbo-0125 with crafted prompts to ensure responses were relevant, accurate, and focused on HORSCH machinery documentation.

Performance timing was implemented using Python's time module, capturing execution duration for each component to enable system optimization.

C. plot-RAG (pRAG): A Novel Evaluation Framework

1) Motivation and Design: A key contribution of this work is pRAG, a novel visualization and evaluation framework specifically designed to address the lack of quantitative, interpretable evaluation methods for RAG systems. pRAG provides granular insights into the performance of individual RAG components, particularly the critical retrieval and re-ranking stages. This contrasts with current evaluation approaches, which often focus on end-to-end performance or rely on limited metrics like recall and precision, which are susceptible to outliers [22].

2) Visualization Components: The pRAG visualization (see Figure 1) displays multiple dimensions of system performance simultaneously:

- **Context position tracking:** Visualizes where relevant contexts from ground truth appear in both retrieval and re-ranked results (blue numbers).
- **Retrieval method comparison:** Distinguishes between embedding-based and BM25 keyword-based retrievals (y-axis).
- **Ground truth distribution:** Shows distances between relevant contexts in the document corpus (green numbers).
- **Quantitative metrics overlay:** Presents calculated performance metrics alongside visual representations (top right corner).
- **Right contexts quantity:** Number of relevant contexts from ground truth at this position based on the entire evaluated data set (numbers in parentheses).

3) Metrics Integration: pRAG calculates and visualizes several critical metrics:

- **Specialized recall metrics:**
  - Recall Emb: Effectiveness of embedding-based retrieval
  - Recall BM25: Performance of keyword-based retrieval
  - Recall Full Retrieval: Combined unique contexts retrieval rate
  - Recall Reranking from Retrieval: Preservation of relevant contexts after re-ranking
- **Ranking quality:** Normalized Discounted Cumulative Gain (NDCG) calculation highlighting the importance of positioning relevant information earlier in results
- **Retrieval optimization:** Recommended retrieval sizes for both embedding and BM25 components

4) Actionable Insights: The pRAG framework provides actionable insights by visually exposing:

- Which retrieval method (BM25 or embeddings) more effectively captures relevant contexts
- How effectively the re-ranker prioritizes relevant contexts
- Optimal retrieval configuration parameters
- Performance bottlenecks in specific components

This visualization approach enables the identification of system weaknesses without requiring extensive manual analysis, making it particularly valuable for ongoing RAG system development and optimization.

Figure 1 shows the unitization of the pRAG approach in the analysis of retriever performance. The particular results are further discussed in Section III-A

D. Synthetic Dataset Generation

For evaluation, both manually curated and synthetically generated question-answer pairs based on three technical manuals for products from the HORSCH portfolio: Avatar 12/40 SD, Joker RX, and Tiger MT were created. These documents were selected based on machine sales volume analysis, indicating likely user query subjects.

For each document, we prepared 50 question-answer pairs with relevant contexts as ground truth. From each set, five pairs were randomly selected as examples for synthetic generation. Using these examples iteratively with different ground truth contexts, we generated 45 synthetic question-answer pairs per document using three different language models: GPT-4o-Mini, Gemini-1.5-Flash, and Nemotron-4-340b-Instruct. Also, two comparison methodologies were implemented:

1) **Absolute comparison:** evaluating curated vs. synthetic datasets based on different contexts
2) **Relative comparison:** generating synthetic data using ground truth from the curated dataset

Quality assessment employed a GAN-like approach, using language models (GPT-4o-Mini, Llama-3-Patronus-Lynx-8B-Instruct [23], and Prometheus-7b-v2.0 [24]) as discriminators to evaluate response quality with Pass/Fail determinations and comparative quality judgments.

E. Experimental Setup

Multiple experimental configurations are conceptualized to evaluate different aspects of the RAG system and demonstrate the utility of the pRAG framework. For enhanced retrieval configurations, we used a basic setup and changed specific technical components for enhanced setups:

a) Basic Synthetic Data Evaluation (Setup A):

- Generator: ChatGPT-3.5-Turbo-0125
- Retrieval: paradeDB (BM25+Embeddings)
- Re-ranker: Cross-encoder/msmarco-MiniLM-L6-en-de-v1
- Retrieval size: 8 contexts each for BM25 and embeddings
- Re-ranking size: 4 contexts

b) Enhanced Retrieval Configuration (Setup B):

- Generator: ChatGPT-3.5-Turbo-0125
- Metadata: Machine Name
- Re-ranker: Cross-encoder/msmarco-MiniLM-L6-en-de-v1
- Retrieval size: 60 contexts (BM25+Embeddings)
- Re-ranking size: 20 contexts

c) Enhanced Retrieval Configuration (Setup B-1):

- New Re-ranker: Cross-encoder/ms-marco-MiniLM-L-6-v2

d) Enhanced Retrieval Configuration (Setup B-2):

- New Method: HyDE Integration

For additional quantitative evaluation, we implemented the RAGAS framework to assess context precision, answer credibility, relevance, and accuracy. We compared RAGAS with the pRAG framework to substantiate the validity of the pRAG approach. We supplemented the pRAG approach with timing analysis of each system component, capturing minimum, maximum, and median execution times.

The expert evaluation was conducted with domain specialists who assessed question-answer pair quality in a blinded format, comparing synthesized and manually curated responses without knowledge of their origin to eliminate bias. For this experiment, we used the Basic Setup B with three different datasets.

In this comprehensive methodology using the novel pRAG evaluation framework, we aimed to evaluate not only the overall RAG system performance but also the viability of synthetic data for ongoing system improvement. We aimed to address the issue of available ground truth datasets by generating synthetic datasets automatically based on contexts from the database.

III. Results and Discussion

A. Performance Analysis of RAG Components Using pRAG

1) Retrieval Component Performance: The analysis demonstrates that each component contributes differently to the overall performance and can be individually assessed through visualization with pRAG. Figure 1 illustrates the pRAG visualization, where the positions of contexts in the ground truth collection are mapped against their retrieval positions. It shows that several relevant contexts (positions 6, 7, and 8) were missed by BM25 but captured by embedding-based retrieval. “Large” gaps in the diagram can support decision-making on whether increasing the retrieval size at the cost of performance should be implemented to identify only a few additional relevant contexts. The automated evaluation of RAG systems with pRAG does not require an LLM as a judge. This makes the evaluation more resource-efficient, considering the substantial computational power required by LLMs.

Since pRAG visualizes the full set of retrieved contexts, the precision metric can be omitted. However, integrating the recall value into the diagram is beneficial to complement the visualization with a quantitative metric. The average values from setups in Section II-E b), c), and d) are presented in Table I.

The pRAG visualization enabled a dedicated evaluation of retrieval techniques, revealing that relevant contexts were retrieved either through BM25 or embedding-based methods

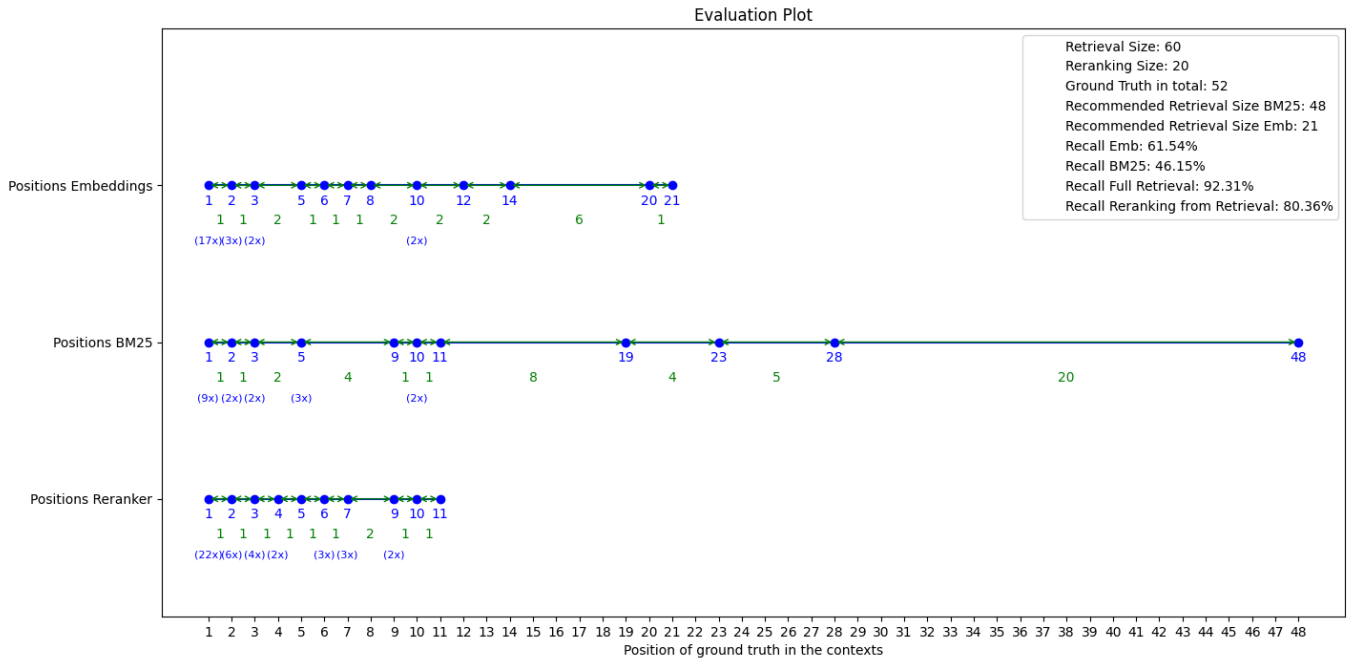

Figure 1. pRAG Visualization Showing Position of Retrieved Contexts Relative to Ground Truth based on Setup B-1.

TABLE I
RECALL METRICS ACROSS ENHANCED RETRIEVAL CONFIGURATION

Metric	Setup B	Setup B-1	Setup B-2
Recall BM25	0.4314	0.4615	0.3654
Recall Embeddings	0.6863	0.6154	0.3846
Recall Full Retrieval	0.9412	0.9231	0.6923

and also in both. This points out the importance of evaluating different retrieval strategies individually, as relevant contexts may be identified in one approach but not in another. Further analysis demonstrated that embedding-based retrieval significantly outperformed lexical methods for datasets containing technical terminology. This underlines the necessity of hybrid retrieval approaches, where the combination of strategies ensures a more comprehensive retrieval process and improves overall performance.

2) Optimal Retrieval Size Determination: The experiments demonstrate a relationship between retrieval size and answer quality. With higher retrieval size RAG systems have to handle more irrelevant contexts. Therefore, the optimal retrieval size can be determined using:

$$\text{Retrieval Size} = \frac{\text{Current Retrieval Size} \times \text{Average Response Time}}{\text{Acceptable Response Time}}$$

This formula provides a practical guideline for balancing response time against completeness. As shown in Figure 1, there is no need to put the retrieval size to 60 because the latest relevant contexts were found in positions 21 by embeddings and position 48 by BM25.

The pRAG analysis revealed diminishing returns beyond certain retrieval sizes. For example, in setup B-2 (cf. Figure 1), increasing BM25 retrieval size from 28 to 48 yielded only one additional relevant context, suggesting a practical cut-off point based on efficiency considerations.

B. Synthetic Dataset Evaluation Results

1) Comparative Quality Assessment: To assess the effectiveness of synthetic versus manually prepared datasets, we evaluated both using specialized discriminator models. The results of this evaluation indicate that synthetically generated data achieves comparable or superior performance. Specifically, manually prepared datasets did not offer a notable advantage, and Lynx even performed better on the synthetic data. This confirms that synthetic datasets can provide a similar level of performance to manually prepared ones. Detailed performance results are presented in Figure 2.

TABLE II
GENERATOR-DISCRIMINATOR COMBINATIONS

No.	Generator - Discriminator Combination
1	GPT-4o Mini - GPT-4o Mini
2	Gemini-1.5-Flash - GPT-4o Mini
3	Neomotron-4-340b-Inst. - GPT-4o Mini
4	GPT-4o Mini - Llama-3-Patronus-Lynx-8B-Inst.
5	Gemini-1.5-Flash - Llama-3-Patronus-Lynx-8B-Inst.
6	Neomotron-4-340b-Inst. - Llama-3-Patronus-Lynx-8B-Inst.
7	GPT-4o Mini - Prometheus-7b-v2.0
8	Gemini-1.5-Flash - Prometheus-7b-v2.0
9	Neomotron-4-340b-Inst. - Prometheus-7b-v2.0

2) Human Expert Validation: Human evaluators assessed pairs of question-answer examples from both dataset types. In

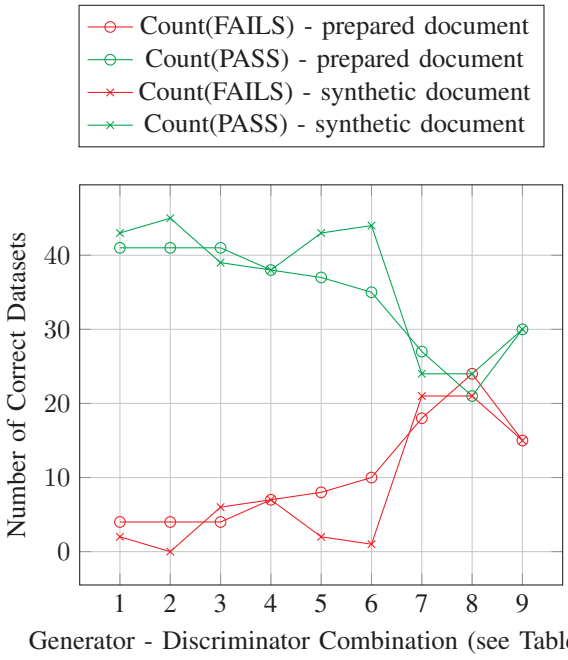

Figure 2. Comparison of Prepared and Synthetic Document for Different Generator-Discriminator Combinations.

68% of comparable cases, experts preferred synthetically generated data, with the remaining 32% showing no clear preference. Table III summarizes these findings. Expert evaluators noted that synthetic datasets showed stronger logical coherence and clearer question formulation. However, they identified a consequential limitation: synthetic datasets generated from tabular data frequently contained factual errors or misinterpretations of numerical relationships, suggesting a specific weakness in current LLM approaches to tabular content. For the evaluation, we used setup B with the three different datasets. The model Nemotron was chosen for its ability to generate better synthetic data [25].

TABLE III
HUMAN EXPERT PREFERENCES IN DATASET EVALUATION

Dataset Pair	Prefer Prepared	Prefer Synthetic	No Preference
Avatar/Nemotron	14	20	16
JokerRX/Nemotron	8	9	31
TigerMT/Nemotron	4	24	22

C. Technical Component Performance Insights

1) Comparative Analysis of Retrieval Enhancements: We evaluated technical enhancements to the base RAG architecture, including re-ranking models and the integration of HyDE (Hypothetical Document Embeddings) [26]. This method decomposes dense retrieval into two distinct tasks: First, it uses an instruction-following language model (like InstructGPT) to generate a hypothetical document in response to a user query. In the second step, an unsupervised contrastively-trained encoder (like Contriever) encodes this hypothetical document into an embedding vector. This vector identifies a neighborhood in the corpus embedding space, from which similar real documents are retrieved based on vector similarity. Table IV summarizes these findings.

TABLE IV
PERFORMANCE COMPARISON OF RETRIEVAL ENHANCEMENT TECHNIQUES

Technique	Recall Re-rank from Retrieval	NDCG	Mean Resp. Time (s)
msmarco-MiniLM-L6-en-de-v1	0.7544	0.54	3.41
ms-marco-MiniLM-L-6-v2	0.8036	0.63	4.08
HyDE Integration	0.7179	0.35	5.43

Contrary to [26], the HyDE approach showed reduced performance despite increased processing time. The pRAG analysis revealed that HyDE's theoretical advantage in generating better query representations did not improve the retrieval of relevant contexts in our test datasets.

Among re-ranking models, ms-marco-MiniLM-L-6-v2 demonstrated the best performance with 80% recall from retrieval but required 20% more processing time than msmarco-MiniLM-L6-en-de-v1. The time-performance analysis shows this tradeoff across various system components.

IV. CONCLUSION AND FUTURE WORK

This paper contributes to the evaluation methodology of RAG systems. Our primary findings are the critical role of ground truth data in conducting valid evaluations of domain-specific RAG applications.

A. Key Contributions

Our research has validated three key advances in RAG evaluation:

1) **The pRAG Visualization:** pRAG provides granular insights into component-level performance that conventional aggregated metrics cannot reveal. This approach allows precise identification of retrieval bottlenecks and optimization opportunities within complex RAG architectures. The visualization-based approach of pRAG offers insights into system performance beyond what metrics-only frameworks provide. pRAG is a resource-saving evaluation technology for RAG systems without any usage of LLM-powered evaluation.
2) **Viability of Synthetic Datasets:** The results confirm comparable or superior evaluation quality of synthetically generated question-answer pairs compared to manually prepared datasets. This significantly reduces the resource burden for domain-specific RAG applications while maintaining evaluation rigor.
3) **Retrieval Optimization Guidelines:** The retrieval component represents the most critical performance bottleneck in RAG systems and we provide a practical formula for determining optimal retrieval size based on response time requirements.

The integration of these approaches enables more efficient and targeted evaluation of RAG systems, particularly in specialized domains where ground truth data creation conventionally requires substantial expert involvement.

B. Future Research Directions

Several promising research directions emerge from this work:

1) **Dynamic Evaluation of Evolving RAG Systems:** Future research should explore automated evaluation approaches for continuously changing RAG systems, potentially integrating pRAG with streaming metrics.
2) **Multi-modal Data Analysis:** Our work focused exclusively on textual data. Extending these evaluation methods to incorporate images, tables, and other data modalities represents an important next step.
3) **Enhanced Synthetic Data Generation:** While our synthetic datasets performed well, specific weaknesses were identified with tabular data. Future work should address these limitations and explore character-based generation approaches to increase dataset heterogeneity.
4) **Generator Component Analysis:** The relationship between retrieval metrics and generation quality is of interest. Future work should explore how retrieved contexts influence the generation process and final answer quality.

In conclusion, the combination of pRAG visualization and synthetic dataset generation represents an advancement in RAG system evaluation methodology. These approaches provide practical tools for researchers and practitioners seeking to optimize RAG implementations for specialized knowledge domains in a more efficient and targeted assessment of individual components.

REFERENCES

[1] X. Wang *et al.*, "Searching for best practices in retrieval-augmented generation," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 17 716–17 736.

[2] H. Yu *et al.*, "Evaluation of retrieval-augmented generation: A survey," in *CCF Conference on Big Data*. Springer, 2024, pp. 102–120.

[3] S. Barnett, S. Kurniawan, S. Thudumu, Z. Brannelly, and M. Abdelrazek, "Seven failure points when engineering a retrieval augmented generation system," in *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI*, 2024, pp. 194–199.

[4] A. Salemi and H. Zamani, "Evaluating retrieval quality in retrieval-augmented generation," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024, pp. 2395–2400.

[5] T. Kenneweg, P. Kenneweg, and B. Hammer, "Retrieval augmented generation systems: Automatic dataset creation, evaluation and boolean agent setup," *arXiv preprint arXiv:2403.00820*, 2024, [retrieved: May, 2025].

[6] Z. Yang *et al.*, "Hotpotqa: A dataset for diverse, explainable multi-hop question answering," *arXiv preprint arXiv:1809.09600*, 2018, [retrieved: May, 2025].

[7] D. F. Campos *et al.*, "Ms marco: A human generated machine reading comprehension dataset," *ArXiv*, vol. abs/1611.09268, 2016, [retrieved: May, 2025]. [Online]. Available: https://api.semanticscholar.org/CorpusID:1289517

[8] K. Zhu *et al.*, "Rageval: Scenario specific rag evaluation dataset generation framework," *arXiv preprint arXiv:2408.01262*, 2024, [retrieved: May, 2025].

[9] R. Liu *et al.*, "Best practices and lessons learned on synthetic data," *arXiv preprint arXiv:2404.07503*, 2024, [retrieved: May, 2025].

[10] S. Kim *et al.*, "Evaluating language models as synthetic data generators," *arXiv preprint arXiv:2412.03679*, 2024, [retrieved: May, 2025].

[11] S. Es, J. James, L. E. Anke, and S. Schockaert, "Ragas: Automated evaluation of retrieval augmented generation," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 2024, pp. 150–158.

[12] Y. Gao *et al.*, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, vol. 2, 2023, [retrieved: May, 2025].

[13] I. Mirzadeh *et al.*, "Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models," *arXiv preprint arXiv:2410.05229*, 2024, [retrieved: May, 2025].

[14] Y. Yu *et al.*, "Rankrag: Unifying context ranking with retrieval-augmented generation in llms," *Advances in Neural Information Processing Systems*, vol. 37, pp. 121 156–121 184, 2024.

[15] G. d. S. P. Moreira *et al.*, "Enhancing q&a text retrieval with ranking models: Benchmarking, fine-tuning and deploying rerankers for rag," *arXiv preprint arXiv:2409.07691*, 2024, [retrieved: May, 2025].

[16] P. Mandikal and R. Mooney, "Sparse meets dense: A hybrid approach to enhance scientific document retrieval," *arXiv preprint arXiv:2401.04055*, 2024, [retrieved: May, 2025].

[17] Y. Lu, M. Shen, H. Wang, X. Wang, C. van Rechem, T. Fu, and W. Wei, "Machine learning for synthetic data generation: a review," *arXiv preprint arXiv:2302.04062*, 2023, [retrieved: May, 2025].

[18] S. Robertson, H. Zaragoza *et al.*, "The probabilistic relevance framework: Bm25 and beyond," *Foundations and Trends® in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.

[19] Hugging Face, "Model card for multi-qa-minilm-l6-cos-v1," [Online]. Available: https://huggingface.co/sentence-transformers/multi-qa-MiniLM-L6-cos-v1, 2021, [retrieved: May, 2025].

[20] ——, "Model card for msmacro-minilm-l6-en-de-v1," [Online]. Available: https://huggingface.co/cross-encoder/msmarco-MiniLM-L6-en-de-v1, 2021, [retrieved: May, 2025].

[21] ——, "Model card for ms-macro-minilm-l-6-v2," [Online]. Available: https://huggingface.co/cross-encoder/ms-marco-MiniLM-L6-v2, 2021, [retrieved: May, 2025].

[22] D. Park and S. Kim, "Probabilistic precision and recall towards reliable evaluation of generative models," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 20 099–20 109.

[23] S. S. Ravi, B. Mielczarek, A. Kannappan, D. Kiela, and R. Qian, "Lynx: An open source hallucination evaluation model," *arXiv preprint arXiv:2407.08488*, 2024, [retrieved: May, 2025].

[24] S. Kim *et al.*, "Prometheus 2: An open source language model specialized in evaluating other language models," *arXiv preprint arXiv:2405.01535*, 2024, [retrieved: May, 2025].

[25] B. Adler *et al.*, "Nemotron-4 340b technical report," *arXiv preprint arXiv:2406.11704*, 2024, [retrieved: May, 2025].

[26] L. Gao, X. Ma, J. Lin, and J. Callan, "Precise zero-shot dense retrieval without relevance labels, 2022," *URL https://arxiv. org/abs/2212.10496*, 2022.