NoobRAG: An Adaptive and Modular RAG System

A F M Mohimenul Joaa*
ScaDS.AI, TU Dresden
Dresden, Germany
mohimenul.joaa@gmail.com

Ramy Boulos*
ScaDS.AI, TU Dresden
Dresden, Germany
ramyboulos@gmail.com

Himanshu Manoj Kaloni*
ScaDS.AI, TU Dresden
Dresden, Germany
himanshu_manoj.kaloni@tu-dresden.de

Michael Färber*
ScaDS.AI, TU Dresden
Dresden, Germany
michael.faerber@tu-dresden.de

Abstract

We propose NoobRAG, a real-time RAG solution designed for accurate and robust responses, leveraging advanced query refinement, adaptive iteration control, dynamic hyperparameter selection, and reranking. To meet strict response-time constraints, we plan to incorporate ideas from PipeRAG [5] and AdaptiveRAG [4], ensuring an optimized pipeline with no idle components. Our hybrid retrieval system combines sparse retrieval (OpenSearch) and dense retrieval (Pinecone) to select the top-k documents, which are then refined through Focus Mode Transformation and reranked using diversity enforcement and Gumbel. Additionally, we will use TII's DataMorgana [1] tool to generate synthetic benchmarks for training and testing.

Keywords

Hybrid Retrieval, Query Refinement, Focus Mode Transformation, Diversity Enforcement, Gumbel Reranking, Adaptive RAG Iterations

ACM Reference Format:
A F M Mohimenul Joaa, Ramy Boulos, Himanshu Manoj Kaloni, and Michael Färber. 2025. NoobRAG: An Adaptive and Modular RAG System. In *Proceedings of SIGIR'25 (SIGIR '25)*. ACM, New York, NY, USA, 2 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Our Solution

Our approach is built on a multi-faceted strategy designed to refine query understanding and improve the relevance of retrieved documents. We begin by addressing query refinement through advanced techniques that enhance the original input. First, an auxiliary query expansion mechanism is employed, where a lightweight language model (with fewer than 10B parameters) enriches the initial query to capture subtle nuances of user intent [7]. In addition to this, an intermediate step involves generating a plausible answer hypothesis using another language model. This hypothesis is then encoded into vector embeddings, which serve as refined queries to boost semantic document retrieval—especially in cases where the original query is ambiguous [2].

In the next stage, our Focus Mode transformation is applied to the retrieved documents. Here, each document is segmented into individual sentences, and these sentences are ranked according to their relevance to the query. Only the highest-ranked sentences are selected for further processing, thereby ensuring that the most pertinent information is retained [7].

To further optimize the system, we introduce adaptive retrieval-generation iterations alongside dynamic hyperparameter selection. A classifier inspired by the AdaptiveRAG framework is utilized to predict the optimal number of retrieval-generation iterations—ranging from 0 to n—for each query. This adaptive strategy ensures computational efficiency without compromising the quality of the results [4]. Furthermore, recognizing the interplay between input token size and output quality, a dedicated neural network predicts the ideal hyperparameters for each query. This network dynamically adjusts parameters such as input token size, iteration count, and other fine-tuning settings, thus achieving balanced performance across various query and model configurations [6].

Finally, our method incorporates both diversity and reranking mechanisms to further refine the retrieved set of documents. By applying differentiable Gumbel Softmax reranking, the system optimizes document selection to prioritize the most relevant content [3]. In parallel, diversity-aware rankers are implemented to enforce a broad and non-redundant selection of information, ensuring that the final output captures a wide array of perspectives [8].

This comprehensive solution thus integrates query refinement, focus mode transformation, adaptive iteration strategies, dynamic hyperparameter selection, and advanced reranking techniques to robustly address the challenges of semantic document retrieval and improve overall system performance.

*All authors contributed equally to this research.

SIGIR '25, 2025, Padua, Italy

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/XXXXXXX.XXXXXXX

References

[1] Simone Filice, Guy Horowitz, David Carmel, Zohar Karnin, Liane Lewin-Eytan, and Yoelle Maarek. 2025. Generating Diverse Q&A Benchmarks for RAG Evaluation with DataMorgana. arXiv:2501.12789 [cs.CL] https://arxiv.org/abs/2501.12789
[2] Yunfan Gao, Yun Xiong, Meng Wang, and Haofen Wang. 2024. Modular RAG: Transforming RAG Systems into LEGO-like Reconfigurable Frameworks. arXiv:2407.21059 [cs.CL] https://arxiv.org/abs/2407.21059
[3] Siyuan Huang, Zhiyuan Ma, Jintao Du, Changhua Meng, Weiqiang Wang, Jingwen Leng, Minyi Guo, and Zhouhan Lin. 2025. Gumbel Reranking: Differentiable End-to-End Reranker Optimization. arXiv:2502.11116 [cs.CL] https://arxiv.org/

abs/2502.11116
[4] Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. 2024. Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity. arXiv:2403.14403 [cs.CL] https://arxiv.org/abs/2403.14403
[5] Wenqi Jiang, Shuai Zhang, Boran Han, Jie Wang, Bernie Wang, and Tim Kraska. 2024. PipeRAG: Fast Retrieval-Augmented Generation via Algorithm-System Co-design. arXiv:2403.05676 [cs.CL] https://arxiv.org/abs/2403.05676
[6] Bowen Jin, Jinsung Yoon, Jiawei Han, and Sercan O Arik. 2025. Long-Context LLMs Meet RAG: Overcoming Challenges for Long Inputs in RAG. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=oU3tpaR8fm
[7] Siran Li, Linus Stenzel, Carsten Eickhoff, and Seyed Ali Bahrainian. 2025. Enhancing Retrieval-Augmented Generation: A Study of Best Practices. arXiv:2501.07391 [cs.CL] https://arxiv.org/abs/2501.07391
[8] Zhchao Wang, Bin Bi, Yanqi Luo, Sitaram Asur, and Claire Na Cheng. 2025. Diversity Enhances an LLM's Performance in RAG and Long-context Task. arXiv:2502.09017 [cs.CL] https://arxiv.org/abs/2502.09017