Rethinking Relevance: How Noise and Distractors Impact Retrieval-Augmented Generation*

Florin Cuconasu[1,†], Giovanni Trappolini[1,†], Federico Siciliano[1], Simone Filice[2], Cesare Campagnano[1], Yoelle Mareek[2], Nicola Tonellotto[3] and Fabrizio Silvestri[1]

[1] *Sapienza University, Rome, Italy*
[2] *Technology Innovation Institute, Haifa, Israel*
[3] *University of Pisa, Pisa, Italy*

Abstract

Retrieval-Augmented Generation (RAG) systems enhance the performance of Large Language Models (LLMs) by incorporating external information fetched from a retriever component. While traditional approaches prioritize retrieving "relevant" documents, our research reveals that these documents can be a double-edged sword. We explore the counterintuitive benefits of integrating noisy, non-relevant documents into the retrieval process. In particular, we conduct an analysis of how different types of retrieved documents—relevant, distracting, and random—affect the overall effectiveness of RAG systems. Our findings reveal that the inclusion of random documents, often perceived as noise, can significantly improve LLM accuracy, with gains up to 35%. Conversely, highly scored but non-relevant documents from the retriever negatively impact performance. These insights challenge conventional retrieval strategies and suggest a paradigm shift towards rethinking information retrieval for neural models.

Keywords

Information Retrieval, Retrieval-Augmented Generation, Large Language Models

1. Introduction

Large Language Models (LLMs) [2, 3, 4, 5, 6] have shown unprecedented capabilities in generating human-like text and answering complex questions (and beyond [7, 8, 9, 10]). Despite their ability, these models are limited by their incapacity to update or expand their knowledge beyond their pre-training data. This limitation becomes particularly evident when handling queries that require up-to-date information or specialized knowledge. To address this, among other issues [11, 12], Retrieval-Augmented Generation (RAG) [13] systems have emerged, which extend the functionality of LLMs by retrieving relevant information from external sources to augment the original prompts.

Traditionally, the retrieval component in RAG systems has focused on fetching documents that are "relevant" to the query [14, 15]. The underlying assumption is that the more relevant

IIR'24: 14th Italian Information Retrieval Workshop, September 05–06, 2024, Udine, Italy

[*] This is an extended abstract of [1].

[†] These authors contributed equally.

cuconasu@diag.uniroma1.it (F. Cuconasu); trappolini@diag.uniroma1.it (G. Trappolini); siciliano@diag.uniroma1.it (F. Siciliano); filice.simone@gmail.com (S. Filice); campagnano@diag.uniroma1.it (C. Campagnano); yoelle@yahoo.com (Y. Mareek); nicola.tonellotto@unipi.it (N. Tonellotto); fsilvestri@diag.uniroma1.it (F. Silvestri)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

the information, the more accurate the LLM's responses will be. However, this approach was created for another scenario, where retrieved documents would be passed down to a human to read and review. In this study, we challenge this assumption by investigating the impact of different types of retrieved documents—including highly relevant, semantically related but non-relevant (distractors), and completely random documents (noise)—on the performance of RAG systems.

Our analysis reveals a surprising phenomenon: the inclusion of random documents, often dismissed as noise, can enhance the accuracy of LLM responses. We observe that strategic placement of these random documents within the context can lead to accuracy improvements of up to 35%. In contrast, top-scoring distractor documents, which do not contain the direct answer but are contextually related, can degrade performance by misguiding the model.

These findings suggest a paradigm shift in the design of retrieval strategies for RAG systems. Instead of solely focusing on maximizing relevance, incorporating a balanced mix of document types, including noise, can lead to better overall performance. This counterintuitive approach calls for a re-evaluation of current retrieval methodologies and paves the way for more effective integration of LLMs and retrieval systems.

This study's contributions can be summarized as follows:

- We provide a detailed examination of how different types of retrieved documents—relevant, distracting, and random—impact the effectiveness of RAG systems.
- We uncover the counterintuitive finding that incorporating random documents, perceived as noise, into the retrieval process can significantly enhance RAG accuracy, with improvements of up to 35%.
- Our results suggest a need to shift retrieval strategies laying the groundwork for future research to optimize RAG system performance by leveraging both relevance and informational noise.

2. RAG

RAG, along with its variations [16, 17, 18, 19], is a technique that enhances the capabilities of LLMs by combining two key components: information retrieval and text generation. In this study, we will concentrate on the task of Open-domain Question Answering (OQA), where the goal is to answer a question q with the support of a corpus of documents $\mathcal{D}$.

Retriever The retriever's role is to find a sufficiently small subset of documents $\mathcal{D}_r$ to allow the reasoner to answer the query correctly. Among the various retrieval methodologies, the use of a dense retriever has gained prominence due to its effectiveness in handling semantic matches. The dense retriever processes both the query q and potential source documents to generate corresponding embeddings $\vec{q}$ for the query and $\vec{d_i}$ for each document $d_i \in \mathcal{D}$. The embedding process can be represented as $\vec{q} = Encoder_q(q)$; $\vec{d_i} = Encoder_d(d_i)$ where $Encoder_q$ and $Encoder_d$ are neural network-based encoders. Once the embeddings are generated, the retrieval process involves computing the similarity, for instance, cosine similarities, between the query embedding and each document embedding. According to these scores, the top-ranked documents are selected for further processing in the generator component.

Reasoner The second step involves a generator component in charge of synthesizing an answer, typically implemented via an LLM. Generative language models operate by predicting the probability distribution of the next token, given the previous tokens. For a given sequence of words $w_1, w_2, \dots, w_n$, a generative language model aims to maximize the likelihood of this sequence, expressed using the chain rule of probability: $P(w_1, w_2, \dots, w_n) = \prod_{i=1}^{N} P(w_i | w_1, w_2, \dots, w_{i-1})$ where $P(w_i | w_1, w_2, \dots, w_{i-1})$ is the conditional probability of the word w_i given the preceding sequence of words $w_1, w_2, \dots, w_{i-1}$. In RAG, the generative language model takes a query q and the retrieved documents $\mathcal{D}_r$ as input and generates a response by sequentially predicting the next token in the sequence. More formally, $P_{rag}(y|q) \approx \prod_{i}^{N} \sum_{d \in \mathcal{D}_r} p_\eta(d|q) p_\theta(y_i | q, d, y_{1:i-1})$ where $p_\eta(d|q)$ is the retrieval component that provides a (truncated) probability distribution for the top-scoring documents, and $p_\theta(y_i | q, d, y_{1:i-1})$ is a probability distribution parameterized by θ that generates a current token based on the previously generated tokens, the query, and the retrieved document; this role is filled by the LLM. In the case of dense retrieval, the probability distribution for the top-scoring documents may assume a functional form of the kind $p_\eta(d|q) \propto \exp(\vec{q} \cdot \vec{d})$.

3. Experimental Evaluation

We perform our analysis on the open-domain version of Natural Questions (NQ-open) [20, 21] dataset, and we show results for Llama 2 7B-Chat [3]. As a retriever, we utilize the Contriever model [15], which selects documents from the English Wikipedia corpus. Results are reported in terms of accuracy; specifically, the answer is considered correct if contains the ground truth. We perform two sets of experiments.

Table 1
Accuracy results of Llama2 7B-Chat when evaluated with prompts composed of the gold document and a varying number of distracting 🔗 documents. Full results are available in [1].

# 🔗	0	1	2	4	6	8	10	12	14	16	18
Llama2	56.42	42.83	39.74	37.95	38.80	37.48	37.16	39.91	41.18	38.89	37.81

Results I In the first sets of experiments, we study the impact of distracting documents. Those are the documents that are scored highly by the retriever but are not useful in answering the question; we indicate these documents with 🔗. To study their impact, we evaluate the LLMs' effectiveness in an oracle setup, where an increasing number of distracting documents are added to the gold document. As an example, we might ask "What color is Napoleon's horse?" the gold document would say that it is grey, while a distracting one could say that Napoleon's wife's horse is brown. The results, displayed in Table 1, show that LLM accuracy declines as the number of distracting documents in the context increases, with a decrease of 18.61 points (-33%) when the context includes 18 distracting documents. This indicates that adding semantically aligned yet non-relevant documents introduces a layer of complexity that can misguide LLMs from identifying the correct response.

Table 2
Accuracy of Llama2 7B-Chat in configurations involving random ⚅ Wikipedia documents and retrieved documents 🗎.

# 🗎 # ⚅	1	2	3	4	5	8	10
0	16.20	18.66	18.76	18.66	19.21	21.98	21.08
1	13.08	16.16	11.77	18.93	19.87	21.35	21.46
2	13.15	16.44	18.95	20.06	21.74	21.56	23.68
3	13.01	17.27	20.08	23.16	22.01	21.98	24.09
5	14.64	20.56	22.33	22.40	21.50	24.51	24.82
8	17.34	20.66	23.36	23.75	24.54	24.16	23.64
10	17.96	21.74	24.00	25.02	24.99	24.20	-
15	20.16	23.54	25.51	25.30	-	-	-
16	20.32	24.71	25.58	-	-	-	-
17	20.39	24.26	-	-	-	-	-
18	20.73	-	-	-	-	-	-

Results II Table 2 presents results from a more realistic scenario where the gold document is not predetermined. We model both the addition of retrieved documents, 🗎 (rows), and that of randomly picked documents, ⚅ (columns). Interestingly, the inclusion of random documents seems to help the model focus on the correct information within the provided documents, as indicated by the increase in model accuracy when these documents are included. For instance, by adding 15 random documents to 4 retrieved ones, there is an increase of 6.64 points (+35%).

4. Conclusion and Future Work

In this study, we explored how different types of retrieved documents affect RAG systems, focusing on the qualities that a retriever should have to enhance prompt effectiveness for RAG configurations. Our findings challenge the prevailing assumptions about document retrieval. Specifically, we discovered that highly ranked retrieved documents that lack the answer can actually be detrimental to the effectiveness of LLMs. Intriguingly, we found that introducing completely random documents can boost the accuracy of these systems. These results warrant a rethinking of traditional IR systems to better suit emerging NLP systems. In our future work, we plan to investigate whether this behavior is consistent across different types of datasets and tasks, involving various models.

Acknowledgments

This work is supported by the Spoke “FutureHPC & BigData” of the ICSC – Centro Nazionale di Ricerca in High-Performance Computing, Big Data and Quantum Computing, the Spoke “Human-centered AI” of the M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - “FAIR - Future Artificial Intelligence Research”, SERICS (PE00000014), IR0000013 - SoBigData.it, funded by European Union – NextGenerationEU, the FoReLab project (Departments of Excellence), and the NEREO PRIN project funded by the Italian Ministry of Education and Research Grant no. 2022AEFHAZ. This work was carried out while Florin Cuconasu was enrolled in the Italian National Doctorate on Artificial Intelligence run by the Sapienza University of Rome.

References

[1] F. Cuconasu, G. Trappolini, F. Siciliano, S. Filice, C. Campagnano, Y. Maarek, N. Tonellotto, F. Silvestri, The power of noise: Redefining retrieval for rag systems, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, volume 17 of *SIGIR 2024*, ACM, 2024, p. 719–729. URL: http://dx.doi.org/10.1145/3626772.3657834. doi:10.1145/3626772.3657834.

[2] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, 2020. URL: https://arxiv.org/abs/2005.14165. arXiv:2005.14165.

[3] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, T. Scialom, Llama 2: Open foundation and fine-tuned chat models, 2023. URL: https://arxiv.org/abs/2307.09288. arXiv:2307.09288.

[4] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7b, 2023. URL: https://arxiv.org/abs/2310.06825. arXiv:2310.06825.

[5] A. Bacciu, G. Trappolini, A. Santilli, E. Rodolà, F. Silvestri, Fauno: The italian large language model that will leave you senza parole!, in: F. M. Nardini, N. Tonellotto, G. Faggioli, A. Ferrara (Eds.), Proceedings of the 13th Italian Information Retrieval Workshop (IIR 2023), Pisa, Italy, June 8-9, 2023, volume 3448 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 9–17. URL: https://ceur-ws.org/Vol-3448/paper-24.pdf.

[6] A. Bacciu, C. Campagnano, G. Trappolini, F. Silvestri, DanteLLM: Let's push Italian LLM research forward!, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 4343–4355. URL: https://aclanthology.org/2024.lrec-main.388.

[7] G. Trappolini, A. Santilli, E. Rodolà, A. Halevy, F. Silvestri, Multimodal neural databases, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 2619–2628. URL: https://doi.org/10.1145/3539618.3591930. doi:10.1145/3539618.3591930.

[8] G. Tolomei, C. Campagnano, F. Silvestri, G. Trappolini, Prompt-to-os (p2os): Revolutionizing operating systems and human-computer interaction with integrated ai gen-

erative models, in: 2023 IEEE 5th International Conference on Cognitive Machine Intelligence (CogMI), IEEE Computer Society, Los Alamitos, CA, USA, 2023, pp. 128–134. URL: https://doi.ieeecomputersociety.org/10.1109/CogMI58952.2023.00027. doi:10.1109/CogMI58952.2023.00027.

[9] G. Barnabò, G. Trappolini, L. Lastilla, C. Campagnano, A. Fan, F. Petroni, F. Silvestri, Cycledrums: automatic drum arrangement for bass lines using cyclegan, Discov. Artif. Intell. 3 (2023). URL: https://doi.org/10.1007/s44163-023-00047-7. doi:10.1007/S44163-023-00047-7.

[10] W. Mucha, F. Cuconasu, N. A. Etori, V. Kalokyri, G. Trappolini, Text2taste: A versatile egocentric vision system for intelligent reading assistance using large language model, in: Computers Helping People with Special Needs: 19th International Conference, ICCHP 2024, Linz, Austria, July 8–12, 2024, Proceedings, Part II, Springer-Verlag, Berlin, Heidelberg, 2024, p. 285–291. URL: https://doi.org/10.1007/978-3-031-62849-8_35. doi:10.1007/978-3-031-62849-8_35.

[11] H. Inan, K. Upasani, J. Chi, R. Rungta, K. Iyer, Y. Mao, M. Tontchev, Q. Hu, B. Fuller, D. Testuggine, M. Khabsa, Llama guard: Llm-based input-output safeguard for human-ai conversations, 2023. URL: https://arxiv.org/abs/2312.06674. arXiv:2312.06674.

[12] F. Petroni, F. Siciliano, F. Silvestri, G. Trappolini, Ir-rag @ sigir24: Information retrieval's role in rag systems, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 3036–3039. URL: https://doi.org/10.1145/3626772.3657984. doi:10.1145/3626772.3657984.

[13] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. URL: https://arxiv.org/abs/2005.11401. arXiv:2005.11401.

[14] S. Robertson, H. Zaragoza, The probabilistic relevance framework: Bm25 and beyond, Found. Trends Inf. Retr. 3 (2009) 333–389. URL: https://doi.org/10.1561/1500000019. doi:10.1561/1500000019.

[15] G. Izacard, M. Caron, L. Hosseini, S. Riedel, P. Bojanowski, A. Joulin, E. Grave, Unsupervised dense information retrieval with contrastive learning, 2022. URL: https://arxiv.org/abs/2112.09118. arXiv:2112.09118.

[16] A. Salemi, H. Zamani, Evaluating retrieval quality in retrieval-augmented generation, 2024. URL: https://arxiv.org/abs/2404.13781. arXiv:2404.13781.

[17] H. Zamani, F. Diaz, M. Dehghani, D. Metzler, M. Bendersky, Retrieval-enhanced machine learning, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '22, ACM, 2022. URL: http://dx.doi.org/10.1145/3477495.3531722. doi:10.1145/3477495.3531722.

[18] S. Pathiyan Cherumanal, L. Tian, F. M. Abushaqra, A. F. Magnossão de Paula, K. Ji, H. Ali, D. Hettiachchi, J. R. Trippas, F. Scholer, D. Spina, Walert: Putting conversational information seeking knowledge into action by building and evaluating a large language model-powered chatbot, in: Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval, CHIIR '24, ACM, 2024. URL: http://dx.doi.org/10.1145/3627508.3638309. doi:10.1145/3627508.3638309.

[19] A. Bacciu, F. Cuconasu, F. Siciliano, F. Silvestri, N. Tonellotto, G. Trappolini, RRAML:

reinforced retrieval augmented machine learning, in: R. Basili, D. Lembo, C. Limongelli, A. Orlandini (Eds.), Proceedings of the Discussion Papers - 22nd International Conference of the Italian Association for Artificial Intelligence (AIxIA 2023 DP) co-located with 22nd International Conference of the Italian Association for Artificial Intelligence (AIxIA 2023), Rome, Italy, November 6-9, 2023, volume 3537 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 29–37. URL: https://ceur-ws.org/Vol-3537/paper4.pdf.

[20] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelcey, M.-W. Chang, A. M. Dai, J. Uszkoreit, Q. Le, S. Petrov, Natural questions: A benchmark for question answering research, Transactions of the Association for Computational Linguistics 7 (2019) 453–466.

[21] K. Lee, M.-W. Chang, K. Toutanova, Latent retrieval for weakly supervised open domain question answering, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 6086–6096. URL: https://www.aclweb.org/anthology/P19-1612. doi:10.18653/v1/P19-1612.