Beyond Relevance: Quantifying Distraction of Irrelevant Passages in RAG*

Chen Amiraz[1], Florin Cuconasu[1,2,†], Simone Filice[1] and Zohar Karnin[1]

[1]Technology Innovation Institute, Haifa, Israel
[2]Sapienza University, Rome, Italy

Abstract

Retrieval Augmented Generation (RAG) systems often struggle with irrelevant passages that mislead LLMs during answer generation. This work introduces a comprehensive framework for quantifying and understanding the distracting nature of such passages. We propose a novel metric to measure passage-level distraction effects, demonstrating its robustness across different models. Our methodology combines retrieval-based approaches with controlled synthetic generation techniques that create distracting content spanning multiple categories. Through experimental validation on standard question-answering benchmarks, we show that passages with higher distraction scores consistently degrade model effectiveness, even when relevant content is present. Leveraging this framework, we construct an enhanced training dataset featuring systematically curated distracting passages. When fine-tuned on this dataset, LLMs demonstrate substantial improvements, achieving up to 7.5% accuracy gains over baselines trained on standard RAG data. Our contributions provide both theoretical insights into distraction mechanisms in RAG and practical solutions for developing more robust retrieval-augmented language models.

Keywords

Retrieval Augmented Generation, Large Language Models, Information Retrieval

1. Introduction

The integration of retrieval mechanisms with Large Language Models (LLMs) has become a cornerstone approach for addressing knowledge-intensive tasks [2, 3, 4]. By incorporating external knowledge through retrieved passages, RAG systems effectively mitigate hallucination issues [5] and provide access to current information beyond the model's training data. However, the retrieval process may introduce *distracting passages* that can mislead the generation process [6, 7, 8]. Unlike completely unrelated content, distracting passages exhibit semantic similarity to the input query while failing to contain the correct answer. This subtle relationship creates a particularly problematic scenario: when no relevant content is present, LLMs may generate responses based on misleading information rather than abstaining from answering; whereas, when relevant content is available, distracting passages may prevent the LLM from focusing on the correct information, leading to erroneous responses despite having access to the right answer.

Our work addresses these challenges by developing a systematic approach to identify distracting passages and quantify their *distracting effect*. We demonstrate that despite the apparent model-dependency of distraction susceptibility, the relative distracting effects of passages show remarkable consistency across different LLMs, as evidenced by high correlation scores between models. Furthermore, we validate our measure by showing that passages with higher distracting effect scores cause more significant degradation in answer accuracy, even when relevant information is also available to the model.

Our investigation encompasses two complementary methodologies for fetching distracting content. First, we analyze passages obtained through various retrieval strategies, including a novel answer-

IIR2025: 15th Italian Information Retrieval Workshop, 3th - 5th September 2025, Cagliari, Italy
*This is an extended abstract of [1].
†Corresponding author.
Work conducted while Florin Cuconasu being a research intern at TII.
chen.amiraz@tii.ae (C. Amiraz); cuconasu@diag.uniroma1.it (F. Cuconasu); filice.simone@gmail.com (S. Filice); zohar.karnin@tii.ae (Z. Karnin)

skewed approach designed to surface topically related but answer-irrelevant content. Second, we generate distracting passages across predefined categories, drawing inspiration from established taxonomies of problematic content [9, 10]. This dual approach enables both empirical analysis of naturally occurring distracting content and controlled examination of specific distraction mechanisms.

Finally, we demonstrate that LLMs fine-tuned using training sets enriched with our identified distracting passages achieve substantially improved accuracy compared to models trained on conventional RAG datasets derived through standard retrieval methods.

2. Quantifying the Distracting Effect

We introduce a quantitative framework for measuring the distracting effect of irrelevant passages. Our approach evaluates a passage's ability to mislead an LLM when the passage does not contain the answer to a given query. We test this by tasking the model to respond with "NO-RESPONSE" when the passage content is insufficient to answer the question. For a given query q and an irrelevant passage p, we define the distracting effect $\mathrm{DE}_q(p)$ as:

$$\mathrm{DE}_q(p) = 1 - \mathrm{P}^{\mathrm{LLM}}(\text{NO-RESPONSE}|q, p)$$

This formulation captures the probability that an LLM will attempt to answer a query based on an irrelevant passage rather than appropriately abstaining. The metric ranges from 0 (no distraction) to 1 (maximum distraction), providing an interpretable measure of the passage's distraction.

Our analysis demonstrates remarkable consistency in distracting effect measurements across different language model architectures. Despite varying model sizes and training procedures, we observe strong correlations (Spearman coefficients ranging from 0.47 to 0.76) in distraction assessments across models from different families (Llama, Falcon, and Qwen) ranging from 3B to 70B parameters. This suggests that the distracting effect represents an intrinsic property of passages and that different LLMs share the same weaknesses.

3. Experimental Analysis of Distracting Passages

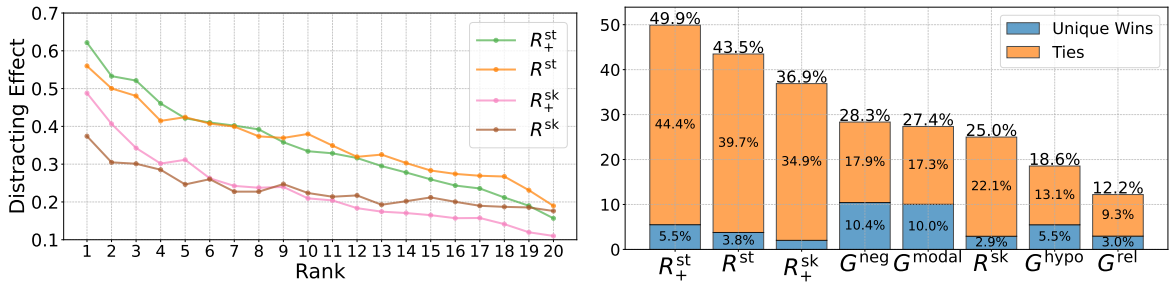
This section presents our empirical investigation into different methods for obtaining distracting passages and analyzes their effectiveness in RAG systems. In this paper, we show results on the Natural Questions (NQ) dataset [11], employing Llama-3.2-3B and Llama-3.1-8B instruct models [12]. Our retrieval pipeline utilizes *E5-base* [13] with optional reranking via the *BGE-M3-v2* cross-encoder [14] applied to the top-20 retrieved passages. Additional results across different models and datasets can be found in our full paper [1]. In our analysis, a passage is considered relevant if it either explicitly contains the ground truth answer or entails the hypothesis "the answer to {question} is {answer}" using the NLI model from Honovich et al. [15]. We systematically exclude such relevant passages when computing distracting effect scores to ensure our measurements focus purely on irrelevant content.

To comprehensively analyze distracting passages, we examine both standard retrieval approaches and novel techniques for obtaining highly distracting content, addressing cases where standard retrieval does not return distracting passages.

Regarding retrieval, we introduce an *answer-skewed retriever* that seeks passages topically related to queries while avoiding answer-relevant content. This approach modifies the standard dense retrieval embedding by subtracting the answer representation from the query embedding: $E^{\mathrm{sub}}(q, a) = E_Q(q) - \lambda E_D(a)$, where λ controls the aggressiveness of answer exclusion[1].

Additionally, we generate irrelevant passages using Claude 3.5 Sonnet V2.0. Following established categorizations [9, 10], we generate four distinct types of distracting content: *Related Topic* passages (G^{rel}) discussing query-adjacent subjects, *Hypothetical* scenarios (G^{hypo}) with alternative answers, *Negation* statements (G^{neg}) providing incorrect information but in negation, and *Modal* statements (G^{modal}) expressing uncertainty about incorrect answers.

[1]We set $\lambda = 1$ in our experiments.



(a) (b)

Figure 1: Experiments on NQ with Llama-3.1-8B. **(a)** Average distracting effect by rank position across retrieval methods. **(b)** Percentage of queries where each method yields the most distracting passage (blue: unique highest scores; orange: highest score is tied with other methods). Additional results in [1].

3.1. Ranking Effects and Method Comparison

Figure 1a demonstrates a critical finding: higher-ranked passages consistently show greater distracting effects compared to lower positions across all retrieval strategies. This pattern holds for both standard dense retrieval (R^{st}) and our answer-skewed approach (R^{sk}). Notably, the integration of reranking modules (R_+^{st} and R_+^{sk}) amplifies this phenomenon. While reranking improves overall retrieval quality, it paradoxically elevates the most distracting passages to top positions. This suggests that contemporary retrieval systems tend to surface passages that are semantically related to queries but factually misleading.

Figure 1b reveals the complementary nature of our different approaches to identifying distracting passages. For the retrieval-based methods, the analysis focuses on the first non-relevant passage returned by each method, which is also the most distracting passage accessible through that approach, as shown in Figure 1a. The vertical bars show the percentage of queries where each method contributes the most distracting passage. The blue parts highlight the times when no other method reaches the same distracting effect as the others. The substantial contribution from both retrieval-based and generation-based methods indicates that a hybrid approach yields superior coverage of distracting passage types. This diversity is particularly valuable for creating robust training datasets, as we will show in Section 4.

3.2. Impact on Answer Generation Quality

To validate our distracting effect measure, we conducted controlled experiments examining how passages with varying distraction levels influence answer accuracy when combined with relevant content. Our experiments account for positional bias [16, 17] by testing both ordering configurations (gold-first and distracting-first) and reporting averaged results. We establish three experimental conditions to demonstrate the progressive impact of distraction. First, when prompts include only the relevant passage, baseline accuracy reaches 82.6 and 80.6 for Llama-3.2-3B and Llama-3.1-8B, respectively. Second, adding a weak distractor (distracting effect smaller than 0.2) alongside the relevant passage causes modest performance degradation, with accuracy declining to 79.4 and 80.1. Third, incorporating a hard distractor (distracting effect greater than 0.8) produces substantially greater impact, with accuracy dropping more dramatically to 71.5 and 73.9 for the same models.

These results validate that our proposed metric effectively identifies truly distracting passages. The substantial accuracy gap between weak and hard distractors, with hard distractors causing accuracy decreases of 6 to 11 percentage points, demonstrates the metric's ability to distinguish between different levels of distracting content.

Test Set	Train Set	Llama-3.2-3B			Llama-3.1-8B		
		acc_u	acc_g	acc	acc_u	acc_g	acc
NQ	*None*	15.2	51.4	37.9	12.8	56.7	40.3
	Retrieve	13.3	**57.1**	40.7	21.0	62.4	46.9
	Rerank	11.5	56.5	39.7	19.9	**63.2**	47.0
	Hard	**21.4**	55.6	**42.8**	**32.0**	59.8	**49.4**
TriviaQA	*None*	38.0	79.2	67.8	39.8	86.4	73.5
	Retrieve	36.9	79.4	67.6	56.8	87.1	78.7
	Rerank	33.3	76.7	64.7	58.4	**87.5**	79.4
	Hard	**54.1**	**82.3**	**74.5**	**68.9**	87.0	**82.0**

Table 1
Answer accuracy averaged over all 4 test sets. *None* is the non-fine-tuned baseline, *Retrieve*, *Rerank* and *Hard* are fine-tuning strategies. Metrics: (1) acc_u, accuracy on ungrounded instances, (2) acc_g, accuracy on grounded instances, and (3) *acc*, overall accuracy. Bold values indicate the highest per model and dataset.

4. RAG Fine-Tuning

Leveraging insights from our distracting effect analysis, we propose a training strategy to fine-tune LLMs for improved robustness in RAG applications. Our approach constructs a training dataset using 800 queries from NQ, where each query is paired with 5 passages selected according to three distinct strategies: (1) **Retrieve** uses the top 5 passages from standard dense retrieval without reranking; (2) **Rerank** same as Retrieve, but in this case we enable reranking; and (3) **Hard** applies a mixed composition where 50% of instances contain the first relevant passage from the reranker plus the 4 most distracting passages, and 50% contain the 5 most distracting passages using methods from Section 3. Lastly, all passages are randomly shuffled to eliminate undesired positional biases during training.

Table 1 demonstrates the effectiveness of training with carefully curated distracting passages. Models trained on our *Hard* dataset configuration achieve 2-3 accuracy point improvements over baseline approaches on NQ. For out-of-distribution evaluation on TriviaQA [18], whose queries were not used during training, the improvements are even more substantial, with the 3B model showing approximately 7 percentage points gains. The benefits prove particularly pronounced for ungrounded examples, where no relevant passage appears in the prompt and correct answers must be retrieved from the model's parametric memory. This pattern suggests that training with distracting passages enhances models' ability to resist misleading contextual information while appropriately relying on their internal knowledge when external context appears unreliable.

5. Conclusions

This work introduces a formal framework for quantifying the distracting effect of irrelevant passages in RAG systems and demonstrates its effectiveness across multiple LLMs. We reveal that stronger retrieval systems paradoxically surface more highly distracting passages, reducing accuracy by up to 11 percentage points when included in LLM contexts. Leveraging this insight, we develop a comprehensive approach combining retrieved and generated passages to create a challenging dataset with distracting content. Fine-tuning on this dataset enhances model robustness, achieving up to 7.5% accuracy improvements on question-answering benchmarks compared to conventional training approaches. We believe that our framework for quantifying distracting effects will enable new approaches to robust information retrieval in LLM-based systems.

Declaration on Generative AI

During the preparation of this work, the authors used Claude Sonnet 4 to check grammar and spelling.

Acknowledgments

This research was conducted while Florin Cuconasu was enrolled in the Italian National Doctorate on Artificial Intelligence at Sapienza University of Rome. The project received support from PNRR MUR project PE0000013-FAIR.

References

[1] C. Amiraz, F. Cuconasu, S. Filice, Z. Karnin, The distracting effect: Understanding irrelevant passages in RAG, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vienna, Austria, 2025, pp. 18228–18258. URL: https://aclanthology.org/2025.acl-long.892/.

[2] D. Chen, A. Fisch, J. Weston, A. Bordes, Reading wikipedia to answer open-domain questions, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 1870–1879.

[3] F. Petroni, A. Piktus, A. Fan, P. Lewis, M. Yazdani, N. De Cao, J. Thorne, Y. Jernite, V. Karpukhin, J. Maillard, V. Plachouras, T. Rocktäschel, S. Riedel, KILT: a benchmark for knowledge intensive language tasks, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 2523–2544.

[4] A. Bacciu, G. Trappolini, A. Santilli, E. Rodolà, F. Silvestri, et al., Fauno: The italian large language model that will leave you senza parole!, in: CEUR WORKSHOP PROCEEDINGS, volume 3448, CEUR-WS, 2023, pp. 9–17.

[5] W. Fan, Y. Ding, L. Ning, S. Wang, H. Li, D. Yin, T.-S. Chua, Q. Li, A survey on RAG meeting LLMs: Towards retrieval-augmented large language models, in: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2024, pp. 6491–6501.

[6] O. Yoran, T. Wolfson, O. Ram, J. Berant, Making retrieval-augmented language models robust to irrelevant context, in: The Twelfth International Conference on Learning Representations, 2024.

[7] 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, 2024, pp. 719–729.

[8] F. Cuconasu, G. Trappolini, F. Siciliano, S. Filice, C. Campagnano, Y. Maarek, N. Tonellotto, F. Silvestri, et al., Rethinking relevance: How noise and distractors impact retrieval-augmented generation, in: CEUR WORKSHOP PROCEEDINGS, volume 3802, CEUR-WS, 2024, pp. 95–98.

[9] V. Basmov, Y. Goldberg, R. Tsarfaty, LLMs' reading comprehension is affected by parametric knowledge and struggles with hypothetical statements, arXiv preprint arXiv:2404.06283 (2024).

[10] R. Abdumalikov, P. Minervini, Y. Kementchedjhieva, Answerability in retrieval-augmented open-domain question answering, arXiv preprint arXiv:2403.01461 (2024).

[11] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, et al., Natural questions: a benchmark for question answering research, Transactions of the Association for Computational Linguistics 7 (2019) 453–466.

[12] A. Grattafiori, et al., The llama 3 herd of models, 2024. URL: https://arxiv.org/abs/2407.21783. arXiv:2407.21783.

[13] L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder, F. Wei, Text embeddings by weakly-supervised contrastive pre-training, arXiv preprint arXiv:2212.03533 (2022).

[14] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, Z. Liu, BGE M3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024. arXiv:2402.03216.

[15] O. Honovich, R. Aharoni, J. Herzig, H. Taitelbaum, D. Kukliansy, V. Cohen, T. Scialom, I. Szpektor, A. Hassidim, Y. Matias, True: Re-evaluating factual consistency evaluation, in: Proceedings of the

2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2022, pp. 3905–3920.

[16] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, P. Liang, Lost in the middle: How language models use long contexts, Transactions of the Association for Computational Linguistics 12 (2024) 157–173. URL: https://aclanthology.org/2024.tacl-1.9/. doi:10.1162/tacl_a_00638.

[17] F. Cuconasu, S. Filice, G. Horowitz, Y. Maarek, F. Silvestri, Do rag systems suffer from positional bias?, 2025. URL: https://arxiv.org/abs/2505.15561. arXiv:2505.15561.

[18] M. Joshi, E. Choi, D. S. Weld, L. Zettlemoyer, TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 1601–1611.