Reasoning-Aware Utility Estimation for Retrieval-Augmented Generation

Abbas Saleminezhad
University of Toronto
Toronto, Ontario, Canada
a.salemi@utoronto.ca

Sajad Ebrahimi
University of Toronto
Toronto, Ontario, Canada
s.ebrahimi@utoronto.ca

Negar Arabzadeh
University of California, Berkeley
Berkeley, California, USA
negara@berkeley.edu

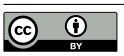
Ebrahim Bagheri
University of Toronto
Toronto, Ontario, Canada
ebrahim.bagheri@utoronto.ca

Abstract

Retrieval-Augmented Generation (RAG) systems increasingly rely on retrieved documents to ground large language model outputs, yet evaluating whether those documents are truly useful for generation remains an open challenge. Recent work has shown that retrieval quality measured by conventional relevance-based metrics does not necessarily align with end-to-end RAG performance: a document judged relevant to a query may not help the model answer correctly, while an irrelevant document may actively distract generation[6, 18, 24]. Existing utility-aware evaluation methods address this limitation by estimating the utility of retrieved documents through isolated document-level probes. However, these approaches estimate document utility independently of the reasoning process. We argue that a document is useful not because it can answer a question in isolation, but because it contributes to the reasoning process that produces the final answer. To capture this behavior, we introduce **Trace Utility Gain (TUG)**, a family of retrieval evaluation metrics that estimate document utility directly from chain-of-thought (CoT) reasoning traces generated over the full retrieval context. Specifically, TUG measures utility through three complementary signals: citation behavior, semantic alignment between retrieved documents and reasoning traces, and model self-assessment of document reliance. By estimating utility from the model's observed reasoning process, TUG captures how retrieved evidence contributes to answer generation rather than how useful it appears in isolation. We evaluate TUG on five question-answering benchmarks using three instruction-tuned large language models. Results show that TUG consistently achieves stronger correlation with end-to-end answer correctness than both traditional retrieval metrics and prior utility-aware approaches.

CCS Concepts

• **Information systems** → **Retrieval tasks and goals**; **Evaluation of retrieval results**.



CIKM '26, Rome, Italy

ACM ISBN 979-8-4007-2539-5/2026/11
https://doi.org/10.1145/3799682.3840041

Keywords

RAG evaluation, retrieval evaluation, document utility

ACM Reference Format:
Abbas Saleminezhad, Sajad Ebrahimi, Negar Arabzadeh, and Ebrahim Bagheri. 2026. Reasoning-Aware Utility Estimation for Retrieval-Augmented Generation. In *Proceedings of the 35th ACM International Conference on Information and Knowledge Management (CIKM '26), November 07–11, 2026, Rome, Italy.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3799682.3840041

1 Introduction

Retrieval-Augmented Generation (RAG) has become a standard paradigm for knowledge-intensive Natural Language Processing (NLP), combining a retriever with a large language model (LLM) to provide external evidence during generation [5, 12, 13]. While retrieval often improves factuality and access to up-to-date information, assessing the quality of the retrieved context remains challenging. Traditional retrieval evaluation relies on relevance judgments and rank-based metrics. However, recent studies have shown that retrieval effectiveness measured by these metrics correlates only weakly with end-to-end RAG performance [21, 23]. A document judged relevant by a human assessor may contribute little to the generated answer, while another document may substantially support the model's reasoning process.

End-to-end evaluation, which directly measures generated answer quality against reference outputs [17], provides a more faithful assessment of the final RAG output but offers limited information about how individual retrieved documents contribute to generation. This motivates retrieval evaluation methods that estimate not only whether a document is relevant to the question, but whether and how it is useful to the generator. The deeper issue is a mismatch between how classical Information Retrieval (IR) metrics characterize retrieval quality and how an LLM utilizes retrieved information. Metrics such as nDCG, MAP, and MRR were designed for human users who examine ranked results with diminishing attention toward lower-ranked documents [4, 8, 16]. In contrast, an LLM processes the retrieved context jointly and exhibits different sensitivity to document ordering and positional effects [14]. Relevance and generation utility are not equivalent. A document judged relevant by a human assessor may contribute little to generation, for example when its information is already available to the model or when the model does not rely on it during reasoning. Conversely, a document with modest relevance may provide evidence that is

particularly useful for producing the answer. Therefore, relevance does not necessarily imply utility for an LLM [1, 7].

Moreover, non-supporting documents are not always harmless. Retrieved documents introduce distracting or conflicting evidence that alters the model's reasoning and degrades answer quality, with different distractors producing substantially different levels of harm [3, 9]. A useful RAG retrieval metric should capture the contribution of retrieved documents to the generation process rather than treating retrieval quality as a function of human relevance alone.

Recent work has begun to move toward utility-oriented retrieval evaluation. eRAG [2, 19] evaluates each query-document pair independently by conditioning the generator on a single retrieved document and measuring the quality of the resulting answer. These per-document answer-quality scores are then treated as generator-oriented utility signals and aggregated across the retrieved list using the ranking-based formulation proposed in eRAG. UDCG [3] also operates at the query-document level, but estimates utility from two components: an LLM-based relevance judgment for the document and the model's probability of producing an answer rather than abstaining when conditioned on that document. These per-document signals are then combined to form the final UDCG score for the retrieved context. Together, these approaches illustrate a shift from purely human-relevance-based evaluation toward signals that more directly reflect the behavior of the generator.

Our key observation is that a reasoning trace generated over the full retrieved context already reveals how the model uses its evidence. It indicates which documents are cited, which reasoning steps they support, and whether some documents are rejected as non-supporting. We therefore introduce *trace-derived utility*, a framework for estimating document utility from a single reasoning rollout over the complete retrieved context. We instantiate this idea through several complementary signals. TUG-Cite derives utility from citation behavior while incorporating abstention and answer confidence. TUG-Role additionally models how documents participate in the reasoning process by distinguishing supporting, background, and non-supporting evidence and using citation polarity to identify distracting documents. We further study TUG-Sem, which measures semantic correspondence between each document and the reasoning trace, and TUG-Self, which directly assesses the model's reliance on each document. Unlike document-wise probing approaches, these signals are derived from the model's behavior when processing the retrieved context jointly. This allows utility estimation to reflect the realized reasoning process while requiring only a single full-context generation for citation-based variants. The resulting document-level signals are then aggregated into a context-level score that can be used to evaluate retrieval quality independently of the final answer correctness label.

The main contributions of this paper are as follows:

- We introduce *trace-derived utility*, a standalone framework for estimating the contribution of retrieved documents from a single reasoning trace generated over the complete RAG context.
- We propose two primary trace-based utility estimators: TUG-Cite, which derives utility from citation, abstention, and confidence signals, and TUG-Role, which models whether retrieved documents support, provide background for, or fail to support the generated reasoning. We additionally study TUG-Sem and TUG-Self as complementary semantic- and self-assessment-based utility signals.
- Across multiple datasets and generator models, we show that trace-derived utility provides a stronger indicator of end-to-end answer quality than conventional relevance-based retrieval metrics and existing utility-oriented baselines.

2 Methodology

Given a question q and a retrieved context $C = \{d_1, \ldots, d_k\}$ consisting of k documents, our goal is to estimate the quality of the retrieval context before evaluating the final generated answer. Recent work has shown that traditional IR metrics are weak predictors of RAG performance because they assume that all relevant documents contribute equally to generation quality and ignore how documents are actually utilized by the language model during reasoning.

Utility-based metrics such as UDCG address this limitation by replacing relevance with document utility. However, existing utility estimation approaches evaluate documents independently, typically by probing the model with a single document at a time. Such estimates ignore the reasoning process occurring when the model receives the full retrieved context. We define document utility in terms of observed contribution to the model's reasoning trace. To capture this behavior, we propose a family of *trace-based utility signals* that estimate document utility directly from CoT reasoning traces generated over the complete retrieval context.

2.1 Trace-Based Utility Estimation

Given the question q and retrieved context C, we prompt an LLM to generate a reasoning trace T and a final answer a. Documents are presented with identifiers $[1], [2], \ldots, [k]$ that allow the model to explicitly reference supporting evidence during reasoning.

For each document d_i, we estimate a utility score $\tau_i \in [0, 1]$ representing its contribution to the reasoning process. We estimate the utility of each document directly from the generated reasoning trace. Let

$$c_i = \begin{cases} 1, & \text{if document } d_i \text{ is cited in the reasoning trace,} \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

We further introduce an abstention-aware gate g that modulates citation-based utility according to the model's response behavior. The gate suppresses citation credit when the model determines that the retrieved context does not provide sufficient evidence to produce an answer, preventing citations made during abstention or evidence rejection from being interpreted as positive utility.

Cite-Abstention (CA). A citation alone does not necessarily indicate that a document contributes positively to the generated answer. For example, a model may cite documents while explaining that the available evidence is insufficient. We therefore define Cite-Abstention as

$$U_i^{\mathrm{CA}} = g\, c_i. \tag{2}$$

This formulation retains the citation signal when the model produces an evidence-supported response, while reducing or suppressing utility when the generation indicates insufficient support.

2.2 Role-Based Utility Estimation

We estimate the utility of each retrieved document from how it is used in the reasoning trace. The role-based prompt distinguishes documents that SUPPORT the answer, provide BACKGROUND information, or are IRRELEVANT. We also use citation polarity to separate citations that support the reasoning from those used to reject or dismiss a document.

We decompose the trace into reasoning steps $\{s_1, \ldots, s_m\}$. A citation is treated as *positive* when the cited document contributes evidence to a reasoning step and as *negative* when the step indicates that the document does not provide the required evidence. Since not all positive citations contribute equally to the final answer, each step is assigned an answer-linkage score $h(s) \in [0, 1]$ based on its lexical overlap with the generated answer. Steps more closely connected to the answer therefore receive larger weights. The positive citation signal for document d_i is then

$$C_i^+ = \sum_{s:\, d_i \in C^+(s)} h(s), \tag{3}$$

where $C^+(s)$ denotes the documents positively cited in step s. Similarly, we define the negative citation signal as

$$C_i^- = \sum_{s=1}^{m} \mathbb{I}\left[d_i \in C^-(s)\right], \tag{4}$$

where $C^-(s)$ denotes citations appearing in steps that reject the document as supporting evidence.

Citation-based evidence is more informative when the generated answer is actually supported by the cited documents. We therefore use a grounding-aware gate η that combines the model's response behavior with the support between the generated answer and its positively cited evidence. The gate suppresses utility when the model abstains and down-weights utility when the answer has weak support in the cited documents. For answered queries, the grounding signal is based on the strongest answer-support score among the positively cited documents:

$$G = \max_{i:\, C_i^+ > 0} E(d_i, a), \tag{5}$$

where $E(d_i, a) \in [0, 1]$ measures how strongly document d_i supports the generated answer a. The answer-support score $E(d_i, a)$ is computed as the lexical overlap between the generated answer and document d_i, measuring the fraction of answer content tokens that also appear in the document.The gate assigns higher weight when the generated answer is better supported by the cited documents, preserving a small minimum weight for non-abstaining responses.

The Role-Based Utility estimates document utility directly from citation behavior in the reasoning trace. Positive, answer-linked citations increase utility, whereas citations used to reject a document decrease it:

$$U_i^{\text{Parse}} = \eta \max\left(0,\, \lambda_{\text{cite}} C_i^+ - \beta C_i^-\right), \tag{6}$$

where λ_{cite} controls the contribution of positive citations and β controls the penalty assigned to negative citations. This formulation avoids treating every cited document as useful, since a document can also be cited when the model explains why it does not support the answer.

2.3 Semantic Trace Utility

Explicit citations may not always be available. As a citation-free alternative, we estimate utility using semantic similarity between the generated reasoning trace and each retrieved document.

Let $e(\cdot)$ denote a sentence embedding model. The semantic utility of document d_i is defined as

$$\tau_i^{\text{Sem}} = \frac{\cos\left(e(T), e(d_i)\right) + 1}{2}. \tag{7}$$

This signal measures the extent to which the content of a document is reflected in the generated reasoning process.

2.4 Self-Assessed Utility

Our final utility estimator directly asks the LLM to assess its reliance on each retrieved document after generating the trace. Given the question, retrieved context, generated reasoning trace, and final answer, the model assigns a utility score

$$\tau_i^{\text{self}} \in [0, 1] \tag{8}$$

to every document.

Unlike citation-based utility, self-assessed utility can capture implicit dependencies that may not appear explicitly in the reasoning trace and does not require citation generation.

2.5 Context-Level Aggregation

To obtain a single context-level score, we aggregate the document-level utilities using a simple signed average. Let τ_i denote the estimated utility of document d_i. We define

$$S(q, C) = \frac{1}{k} \sum_{i=1}^{k} w_i \tau_i, \tag{9}$$

where the weight w_i reflects the role of the document in the generated reasoning.

This aggregation rewards contexts containing evidence that contributes to the generated reasoning while mildly penalizing documents that are explicitly identified as non-supporting. The same aggregation is used across utility estimators, allowing differences in performance to primarily reflect the document-level utility signal rather than the aggregation strategy.

2.6 Computational Complexity

A practical advantage of trace-based utility estimation is efficiency. Original UDCG estimates document utility through independent document-level probes, requiring an additional model invocation for each retrieved document. Consequently, evaluating a context of k documents requires $O(k)$ utility estimation calls.

In contrast, our citation-based and semantic utility estimators reuse a single reasoning trace generated over the full retrieval context. Utility extraction is performed as a lightweight post-processing step without additional generation cost. The self-assessment variant requires only one additional model call per query, resulting in complexity independent of the number of retrieved documents.

Therefore, trace-based utility estimation provides a computationally efficient alternative while leveraging information already produced during answer generation.

Table 1: Spearman correlation between retrieval metrics and end-to-end RAG answer correctness. Results are reported across five datasets and three LLMs. Bold indicates the best metric for each model-dataset combination.

	Llama-3.1-8B					Mistral-7B					Gemma-3-4B				
Metric	PopQA	Trivia	NQ	BioASQ	NoMIRACL	PopQA	Trivia	NQ	BioASQ	NoMIRACL	PopQA	Trivia	NQ	BioASQ	NoMIRACL
NDCG	0.559	0.336	0.404	0.364	0.388	0.502	0.312	0.373	0.367	0.281	0.575	0.336	0.376	0.309	0.217
MAP	0.556	0.337	0.403	0.366	0.384	0.501	0.310	0.373	0.366	0.277	0.572	0.336	0.376	0.309	0.215
MRR	0.617	0.407	0.456	0.377	0.385	0.503	0.402	0.320	0.382	0.283	0.630	0.416	0.437	0.309	0.218
eRAG	0.047	0.373	0.430	0.290	0.116	0.062	0.371	0.366	0.269	0.033	0.017	0.411	0.415	0.235	0.005
UDCG	0.593	0.416	0.432	0.372	0.425	0.453	0.397	0.337	0.335	0.221	0.597	0.424	0.404	0.293	0.235
TUG-Sem	0.622	0.415	0.455	0.372	0.404	0.539	0.413	0.396	0.382	0.322	0.608	0.426	0.448	0.312	0.236
TUG-Self	0.661	0.426	0.465	0.401	0.460	0.516	0.384	0.376	0.388	0.337	0.630	0.435	0.455	0.316	0.276
TUG-Cite	**0.830**	**0.634**	0.562	**0.539**	0.638	**0.634**	**0.538**	**0.538**	**0.578**	**0.704**	**0.734**	**0.659**	0.503	**0.421**	**0.683**
TUG-Role	0.783	0.530	**0.602**	0.534	**0.741**	0.595	0.468	0.443	0.296	0.510	0.710	0.569	**0.504**	0.413	0.640

3 Experiments

3.1 Experimental Setup

Datasets. We evaluate on five question-answering benchmarks spanning diverse domains and retrieval settings: PopQA [15], TriviaQA [10], Natural Questions (NQ) [11, 17], BioASQ[22], and NoMiracl [20]. Following prior work, each query is associated with a set of retrieved documents and binary relevance annotations. We use the top-10 retrieved documents for context construction and metric computation.

Models. To assess the robustness of our approach across model families, we evaluate three open-source instruction-tuned LLMs: Llama-3.1-8B-Instruct[1], Mistral-7B-Instruct-v0.3[2], and Gemma-3-4B-Instruct[3]. The same model is used for answer generation, CoT generation, and self-evaluation. All models are served using vLLM with greedy decoding ($T = 0$).

Baselines. We compare our trace-derived utility metrics against classical retrieval metrics including nDCG, MAP, and MRR, as well as retrieval-aware metrics designed for RAG, namely eRAG and UDCG. Classical metrics rely exclusively on retrieval rankings and relevance labels. eRAG and UDCG estimate document utility through isolated per-passage probing. In contrast, our methods derive utility directly from reasoning traces generated over the complete retrieval context.

Trace generation strategies. We consider two single-rollout trace generation strategies. In the *citation-based* setting, the generator reasons step by step over the retrieved context and cites the supporting document for each reasoning step. If the context does not provide sufficient evidence, it returns `NO-RESPONSE`. This produces a lightweight attribution signal that reveals which documents are used during generation. The *role-based* strategy augments the trace with an explicit assessment of each document. After generating the reasoning and answer, the model assigns every document one of three roles: Support, when the answer depends on its evidence; Background, when it is related but not required; or Irrelevant, when it does not contribute useful evidence. This provides a richer signal than citation alone by distinguishing supporting evidence from related or unused documents. Both strategies use the same retrieved context, allow abstention, and do not expose the gold answer.

Evaluation Protocol. For each query, we compute the score assigned by each retrieval metric and generate an answer using the corresponding retrieval context. Answers are graded against the reference answer set using an LLM judge and converted into binary correctness labels. Following Cuconasu et al. [3], we evaluate retrieval metrics by measuring their Spearman correlation with answer correctness. Spearman is the appropriate choice because a useful retrieval metric must rank contexts consistently with downstream accuracy rather than reproduce its exact value: it captures any monotonic association, while remaining robust to outliers and directly comparable to the rank correlations reported by eRAG and UDCG. Higher correlation indicates stronger alignment between retrieval quality estimates and end-to-end RAG performance.[4]

3.2 Correlation with End-to-End Accuracy

Table 1 reports the Spearman correlation between retrieval metrics and end-to-end answer correctness across three LLMs and five datasets.

Overall, the trace-derived utility metrics achieve the strongest correlation with answer quality across all fifteen model dataset combinations. In particular, either TUG-Cite or TUG-Role obtains the best result in every setting. TUG-Cite is the strongest individual method, achieving the highest correlation in twelve of the fifteen settings, while TUG-Role performs best in the remaining three.

The strongest improvements are observed for Llama-3.1-8B. TUG-Cite achieves an average correlation of 0.641, closely followed by TUG-Role at 0.638, compared with 0.448 for MRR and 0.448 for UDCG. TUG-Cite obtains particularly strong results on PopQA (0.830), TriviaQA (0.634), and BioASQ (0.539), while TUG-Role performs best on NQ (0.602) and NoMIRACL (0.741).

A similar pattern is observed for Mistral-7B, where TUG-Cite achieves the best result on all five datasets and obtains an average correlation of 0.598. This substantially exceeds the strongest classical baseline, MRR (0.378), as well as UDCG (0.349). TUG-Role

[1]https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[2]https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3
[3]https://huggingface.co/google/gemma-3-4b-it

[4]Code, data, and experimental artifacts are available at: https://github.com/Saleminezhad/CoTRAG-eval.

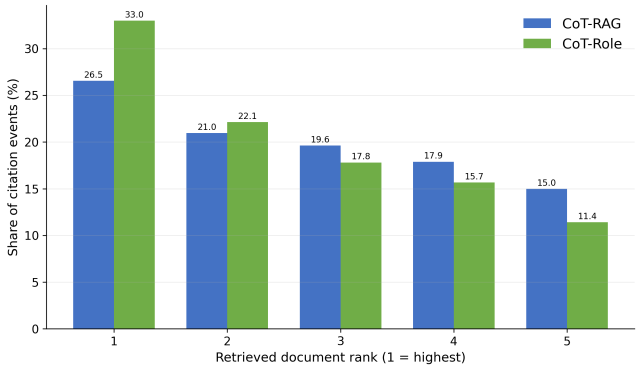


Figure 1: Distribution of cited documents by retrieval rank for CoT-RAG and CoT-Role across five datasets. Rank 1 denotes the highest-ranked retrieved document.

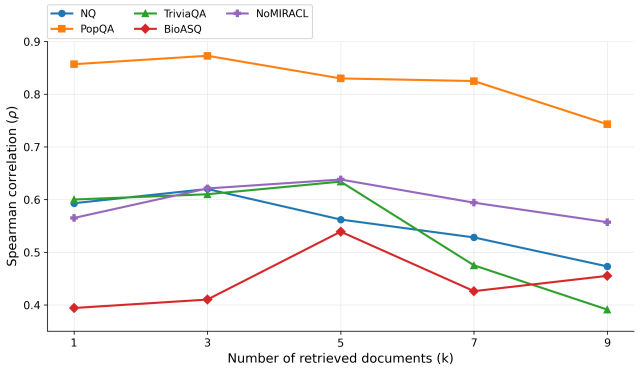


Figure 2: Effect of the number of retrieved documents k on the Spearman correlation between TUG-Cite and end-to-end answer correctness.

remains competitive with an average correlation of 0.462, although its performance varies more across datasets.

For Gemma-3-4B, TUG-Cite achieves the strongest average correlation (0.600), followed by TUG-Role (0.567). TUG-Cite performs best on four of five datasets, while TUG-Role is strongest on NQ (0.504); both substantially outperform UDCG (0.391 average). TUG-Sem and TUG-Self remain competitive with conventional retrieval metrics but consistently trail TUG-Cite and TUG-Role, suggesting that explicitly modeling document use in the reasoning process provides a stronger signal of downstream generation quality. eRAG performs considerably weaker overall, with average correlations of 0.251, 0.220, and 0.217 for Llama-3.1-8B, Mistral-7B, and Gemma-3-4B, respectively. These results further indicate that utility signals derived from full-context reasoning are more closely aligned with final answer correctness.

3.3 Trace Citation Behavior

To better understand how role-based prompting changes evidence use, we compare the citation patterns of CoT-RAG and CoT-Role using 1,000 examples per dataset. Cited documents are mapped to their retrieval rank, with repeated citations within a response counted once and distributions normalized separately for each prompting strategy.

As shown in Figure 1, CoT-Role produces a more top-heavy citation distribution. Rank 1 accounts for 33.0% of CoT-Role citations versus 26.5% for CoT-RAG, while the top two ranks account for 55.1% versus 47.5%. In contrast, rank 5 contributes 15.0% of CoT-RAG citations but only 11.4% of CoT-Role citations. This indicates that role-based prompting concentrates evidence use more strongly on higher-ranked documents.

We also examine whether CoT-Role citations agree with the assigned document roles. Among cited documents, 84.4% are labeled Support, 12.4% are labeled Background or Irrelevant, and 3.2% remain unclassified because of incomplete role parsing. This shows that citation alone does not always imply evidential support. Overall, role-based prompting leads to more selective and support-oriented citation behavior, although this analysis does not by itself establish higher answer accuracy.

3.4 Effect of Context Size

We further analyze the sensitivity of TUG-Cite to the number of retrieved documents k. As shown in Figure 2, performance is generally strongest for moderate context sizes. NQ and PopQA achieve their highest correlations at $k = 3$ (0.620 and 0.873, respectively), while TriviaQA, BioASQ, and NoMIRACL peak at $k = 5$ (0.634, 0.539, and 0.638). Beyond this range, correlation generally decreases as additional documents are introduced, with the largest degradation observed at $k = 9$. These results suggest that TUG-Cite benefits from sufficient supporting evidence, while excessively large contexts may introduce distracting or weakly useful documents that make the trace-derived utility signal less reliable.

4 Conclusion

We revisited document utility estimation for RAG retrieval evaluation and showed that useful utility signals can be extracted directly from the model's reasoning over the full retrieved context. Unlike approaches that evaluate query-document pairs independently, our framework estimates utility from a single full-context reasoning trace, better reflecting how documents are jointly used during generation. Our strongest variants, TUG-Cite and TUG-Role, derive utility from citation behavior and document roles in the reasoning process, while TUG-Sem and TUG-Self provide complementary semantic and self-assessment signals. Across three LLMs and five datasets, either TUG-Cite or TUG-Role achieves the highest correlation with end-to-end answer correctness in all evaluated model-dataset settings.

The proposed framework is also efficient: citation, role, and semantic-based utility require only a single full-context rollout, while TUG-Self requires one additional model call per query. Overall, our results show that reasoning traces provide a strong and inexpensive signal for evaluating how retrieved evidence contributes to RAG generation. Future work can further investigate causal document contribution, contradictory evidence, and robustness across broader retrieval and reasoning settings.

GenAI Usage Disclosure

Generative artificial intelligence (GenAI) tools were used exclusively to improve the readability, clarity, and grammatical correctness of the final manuscript.No GenAI tools were used to generate research ideas, formulate hypotheses, design experiments, analyze results. The intellectual contributions, analyses, and conclusions presented in this paper are entirely those of the authors, who take full responsibility for the content of this work.

References

[1] Chen Amiraz, Florin Cuconasu, Simone Filice, and Zohar Karnin. 2025. The Distracting Effect: Understanding Irrelevant Passages in RAG. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vienna, Austria, 18228–18258.

[2] Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Börschinger, and Tal Schuster. 2022. Tomayto, Tomahto: Beyond Token-Level Answer Equivalence for Question Answering Evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 291–305.

[3] Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. 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 (SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 719–729.

[4] Edward Cutrell and Zhiwei Guan. 2007. What Are You Looking For? An Eye-Tracking Study of Information Usage in Web Search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. Association for Computing Machinery, New York, NY, USA, 407–416.

[5] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv preprint arXiv:2312.10997* (2024).

[6] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A Survey on LLM-as-a-Judge. *arXiv preprint arXiv:2411.15594* (2025).

[7] Jan Hutter, David Rau, Maarten Marx, and Jaap Kamps. 2025. Lost but Not Only in the Middle: Positional Bias in Retrieval Augmented Generation. In *Advances in Information Retrieval: 47th European Conference on Information Retrieval (ECIR 2025)*. Springer, Berlin, Heidelberg, 247–261.

[8] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Transactions on Information Systems* 20, 4 (2002), 422–446.

[9] 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*.

[10] Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. 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)*. 1601–1611.

[11] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and Kristina Toutanova. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics* 7 (2019), 453–466.

[12] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*. Vancouver, BC, Canada, 9459–9474.

[13] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, Vol. 33. 9459–9474.

[14] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics* 12 (2024), 157–173.

[15] Alex Troy Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.

[16] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

[17] Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. 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*. Association for Computational Linguistics, Online, 2523–2544.

[18] Hossein A. Rahmani, Clemencia Siro, Mohammad Aliannejadi, Nick Craswell, Charles L. A. Clarke, Guglielmo Faggioli, Bhaskar Mitra, Paul Thomas, and Emine Yilmaz. 2024. Report on the 1st Workshop on Large Language Models for Evaluation in Information Retrieval (LLM4Eval 2024) at SIGIR 2024. *arXiv preprint arXiv:2408.05388* (2024).

[19] Alireza Salemi and Hamed Zamani. 2024. Evaluating Retrieval Quality in Retrieval-Augmented Generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 2395–2400. doi:10.1145/3626772.3657957

[20] Nandan Thakur, Luiz Bonifacio, Crystina Zhang, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Boxing Chen, Mehdi Rezagholizadeh, and Jimmy Lin. 2024. "Knowing When You Don't Know": A Multilingual Relevance Assessment Dataset for Robust Retrieval-Augmented Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics, Miami, Florida, USA, 12508–12526.

[21] Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2023. Large Language Models Can Accurately Predict Searcher Preferences. *arXiv preprint arXiv:2309.10621* (2023).

[22] George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artieres, Axel Ngonga, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. 2015. An Overview of the BioASQ Large-Scale Biomedical Semantic Indexing and Question Answering Competition. *BMC Bioinformatics* 16 (2015), 138.

[23] Hamed Zamani and Michael Bendersky. 2024. Stochastic RAG: End-to-End Retrieval-Augmented Generation through Expected Utility Maximization. In *Proceedings of the 47th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[24] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.