Redefining Retrieval Evaluation in the Era of LLMs

Giovanni Trappolini[1*†], **Florin Cuconasu**[1,2*‡§]
Simone Filice[2], **Yoelle Maarek**[2], **Fabrizio Silvestri**[1]

[1]Sapienza University of Rome, [2]Technology Innovation Institute

Abstract

Traditional Information Retrieval (IR) metrics, such as nDCG, MAP, and MRR, assume that human users sequentially examine documents with diminishing attention to lower ranks. This assumption breaks down in Retrieval Augmented Generation (RAG) systems, where search results are consumed by Large Language Models (LLMs), which, unlike humans, process all retrieved documents as a whole rather than sequentially. Additionally, traditional IR metrics do not account for related but irrelevant documents that actively degrade generation quality, rather than merely being ignored. Due to these two major misalignments, namely human vs. machine position discount and human relevance vs. machine utility, classical IR metrics do not accurately predict RAG performance. We introduce a utility-based annotation schema that quantifies both the positive contribution of relevant passages and the negative impact of distracting ones. Building on this foundation, we propose UDCG (Utility and Distraction-aware Cumulative Gain), a metric using an LLM-oriented positional discount to directly optimize the correlation with the end-to-end answer accuracy. Experiments on five datasets and six LLMs demonstrate that UDCG improves correlation by up to 36% compared to traditional metrics. Our work provides a critical step toward aligning IR evaluation with LLM consumers and enables more reliable assessment of RAG components[1].

1 Introduction

The emergence of Retrieval Augmented Generation (RAG) systems has fundamentally transformed how we approach knowledge-intensive natural language processing tasks (Gao et al., 2024). By combining the parametric knowledge of Large Language Models (LLMs) with the dynamic retrieval of relevant information from external corpora, RAG systems have demonstrated remarkable capabilities across diverse applications, including question answering (Lewis et al., 2020; Izacard and Grave, 2021) and fact-checking (Khaliq et al., 2024) and beyond. However, as these systems have gained prominence, a critical gap has emerged between the evaluation methodologies inherited from traditional information retrieval (IR) and the unique characteristics of LLM-based retrieval consumers.

Traditional offline IR evaluation metrics, including Normalized Discounted Cumulative Gain (nDCG) (Järvelin and Kekäläinen, 2002), Mean Average Precision (MAP), and Mean Reciprocal Rank (MRR), were developed with human users in mind (Manning et al., 2008). In other words, they assume that the ultimate judge of the quality of the results is a human. These assumptions cause two major limitations in the generative AI and agentic AI eras, when LLMs or more generally *machines* consume the search results.

Human vs. Machine Position Discount. Humans sequentially examine retrieved documents; therefore, human-centric IR metrics assign monotonically decreasing weights to lower-ranked documents, reflecting the intuitive notion that documents appearing further down the list are less likely to be examined and therefore less valuable to the user. Conversely, in the case of RAG systems, LLMs process all retrieved documents as a whole within their input context. Recent research has revealed that LLMs exhibit systematic positional biases when processing prompts that include multiple documents (Liu et al., 2024; Hutter et al., 2025). These biases can manifest in various patterns depending on the model and context, potentially causing LLMs to attend differently to information based on its position in the prompt. This suggests that the monotonic position-based discount embedded

*These authors contributed equally to this work.
†trappolini@diag.uniroma1.it
‡Work conducted while FC being a research intern at TII.
§cuconasu@diag.uniroma1.it
[1]Code and data at https://github.com/GiovanniTRA/UDCG

in classical IR metrics may not accurately reflect how LLMs leverage retrieved content, and how it should be valued.

Human Relevance vs. Machine Utility. A second limitation is that traditional IR metrics operate under a simplistic binary (relevant vs. irrelevant) or ordinal (e.g., 0-5 relevance scale) relevance framework, treating all non-relevant documents as equivalent. This assumption fails to capture two crucial aspects of the RAG setting: (*i*) LLMs might fail to use relevant passages (e.g., because they are convoluted); and (*ii*) irrelevant documents not only do not contribute value, they actually have a negative impact as they act as distractors that can mislead the LLM and degrade the system performance (Cuconasu et al., 2024; Jin et al., 2025). The heterogeneous nature of these distracting effects means that different irrelevant documents can have vastly different impacts on the quality of the final generation (Amiraz et al., 2025). Yet, current metrics provide no mechanism to distinguish between them.

These two limitations of existing IR relevance metrics extend beyond theoretical concerns. They have practical implications for RAG system development and deployment, since IR components are optimized using these metrics, which may actually harm end-to-end system performance. We argue here that RAG systems face today a fundamental challenge: *how can they be reliably evaluated and improved, when their retrieval component is optimized for metrics designed for an entirely different use case?* This paper addresses this challenge through both empirical analysis and methodological innovation.

We begin by providing concrete demonstrations of how traditional IR metrics fail to predict RAG performance, using constructed examples that highlight the disconnect between metric scores and actual system effectiveness. To address this disconnect, we propose a novel passage annotation strategy that fundamentally reconceptualizes relevance assessment for LLM consumers. Rather than relying on discrete relevance judgments designed for human interpretation, we introduce a continuous utility scoring framework that captures both the beneficial impact of relevant passages on LLM generation quality and the distracting effects of irrelevant passages that can mislead the model. This dual consideration of utility and distraction enables a more nuanced understanding of how individual passages contribute to or detract from end-to-end system performance. We demonstrate the practical value of this annotation approach by showing that ranking retrieved passages according to our utility scores and selecting the top-k results yields significantly better LLM accuracy compared to selection based on traditional relevance annotations. This finding provides empirical evidence that optimizing for utility-aware rankings directly translates to improved RAG system performance.

Building on these utility-based annotations, we introduce a novel evaluation metric specifically designed for RAG systems that we refer to as Utility and Distraction-aware Cumulative Gain (UDCG). UDCG leverages our continuous scoring framework to substitute human relevance with machine utility. Moreover, it applies an LLM-oriented positional discount that directly optimizes the correlation between the IR metric and the end-to-end answer accuracy. Through comprehensive experiments across six different LLMs and five diverse benchmarks, we demonstrate that UDCG exhibits substantially stronger correlation with end-to-end RAG accuracy than traditional IR metrics, with increments up to 36%. These results establish UDCG as a more reliable indicator of retrieval component quality for RAG systems, enabling practitioners to optimize their systems toward metrics that actually predict downstream performance.

2 Related Work

2.1 Retrieval Augmented Generation

Recently, RAG has emerged as a new application of IR systems: LLMs by accessing information retrieved with IR systems can incorporate external knowledge into their outputs, improving the response quality (Lewis et al., 2020; Gao et al., 2024; Fan et al., 2024). This approach addresses critical limitations of standalone language models, particularly their tendency toward hallucination and their inability to access information beyond their training cutoff dates. When operating in the RAG setting, LLMs have two major limitations related to their ability to capitalize on the input context. First, the effect of irrelevant documents, defined in Cuconasu et al. (2024) as documents not containing the answer to the question. Humans are not really confused by their presence in the result list since they can quickly skim and dismiss them. On the contrary, for LLMs in RAG systems, irrelevant documents pose a more serious problem: they act as noise that can dilute attention, trigger hallucinations, or cause the model to synthesize incorrect

information by attempting to find connections between the query and unrelated content. Cuconasu et al. (2024) show that while random passages unrelated to the question do not affect answer quality, distracting passages, i.e., passages semantically related to the query but not containing the answer, do. Attempts to improve the LLM's robustness mitigate but do not fully solve the LLM's susceptibility to irrelevant content (Amiraz et al., 2025; Jin et al., 2025; Lin et al., 2024; Yoran et al., 2024; Yu et al., 2024). Amiraz et al. (2025) expand this line of research and propose a continuous score to measure the distracting effect of irrelevant passages. Formally, they compute the distracting effect $\text{DE}_q(p)$ of an irrelevant passage p for question q as the probability of the LLM not abstaining when receiving only passage p and question q in its prompt:

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

In our paper, we will include this formulation as a core component of our proposed metric for evaluating IR systems serving the RAG scenario. A second limitation of the LLM's reading process is the positional bias: the capability of identifying relevant content depends on its location in the prompt. In particular, Liu et al. (2024) discovered the lost-in-the-middle effect, where the LLMs tend to ignore information in the middle of the prompt, and Hutter et al. (2025) demonstrate that different LLMs exhibit distinct positional bias patterns. Cuconasu et al. (2025) extend this work by showing how the positional bias also affects the impact of distracting documents.

2.2 IR Evaluation for RAG

Most works in the RAG literature evaluate the IR components jointly with the generation components, by assessing the quality of the RAG end-to-end response (Yu et al., 2025). This approach, in addition to obscuring the retrieval's distinct contribution, is computationally demanding since it requires generating a response for each new tested context, including contexts that contain the same passages but in a different order.

When focusing on specifically evaluating the retrieved content, most practitioners still rely on traditional IR metrics, including Precision (Prec), HITS, Mean Average Precision (MAP), Mean Reciprocal Rank (MRR), and normalized Discounted Cumulative Gain (nDCG)[2]. These metrics have been designed for human users, and have relevance as a central notion, which represents the degree to which a retrieved document satisfies a user's information need. Also, they generally apply a discount function to reduce the contribution of documents at lower ranks, reflecting the realistic assumption that users are less likely to examine documents further down in the ranking.

A few papers propose IR evaluation strategies directly designed for the RAG setting. Salemi and Zamani (2024) propose eRAG: they score each document based on the quality of the LLM's response when receiving only that document as a context in a RAG setting; then they show that using these scores instead of standard relevance labeling inside traditional IR metrics (e.g., MAP, MRR, nDCG) improves their correlation with the end-to-end RAG performance. Dai et al. (2025) propose Semantic Perplexity, a metric that captures the LLM's internal belief about the correctness of the generated answer, by quantifying the variation of perplexity in the LLM response when answering with and without the retrieved content. The intuition is that useful content leads to increased confidence in the LLM answer.

Similarly to these approaches, in this paper, we employ an LLM-oriented definition of relevance, which we integrate with the notion of distracting effect into a holistic IR metric for the RAG scenario.

3 Flaws of Traditional IR Metrics in RAG

Traditional IR systems and evaluation metrics have been fundamentally designed with human users as the primary consumers of retrieved information, reflecting assumptions about human cognitive processes, information processing capabilities, and search behaviors. These metrics inherently assume that users will manually examine ranked result lists, with higher-ranked documents receiving greater attention–a design philosophy rooted in human browsing patterns (Granka et al., 2004; Cutrell and Guan, 2007).

When the consumers of IR results are LLMs instead of humans, many of these assumptions do not hold. First, the human-centric notion of document relevance does not necessarily align with the document utility in enabling an LLM to successfully execute its task (Tian et al., 2025).

Moreover, there is a fundamental difference in how humans and LLMs handle irrelevant documents, as LLMs lack the intuitive filtering mecha-

[2]A detailed description of these metrics is in Appendix A.

Model	Rel in WD	Rel in HD	Δ (WD - HD)
Llama 3B	81.11	75.56	5.55
Llama 8B	89.44	82.22	7.22
Llama 70B	90.32	83.50	6.82
Gemma 4B	84.00	78.89	5.11
Mistral 7B	88.33	82.00	6.33
Qwen 7B	87.78	78.67	9.11

Table 1: Answer accuracy of LLMs prompted with one relevant passage and four Weak Distractors (WD, DE<0.2) or Hard Distractors (HD, DE>0.8). All differences are statistically significant (Wilcoxon, $p < 0.05$).

nisms that allow humans to cleanly separate signal from noise. Jin et al. (2025) provide evidence of this by showing that irrelevant passages retrieved by high-performing retrieval systems are more distracting, and therefore they cause more harm, than irrelevant passages retrieved by weaker retrieval systems. Consequently, we argue here that IR systems that are trained and evaluated over traditional IR metrics are likely to optimize for the wrong objective when the consumer of results is an LLM, a machine, rather than a human.

Table 1 exemplifies this flaw by comparing the average answer accuracy on NQ[3] of several LLMs when prompted with a single relevant passage surrounded by only weakly or only hardly distracting passages. All traditional IR metrics[4] would score these two scenarios equally, although their end-to-end accuracy changes significantly, exhibiting differences of up to 9 accuracy points.

Another aspect to consider when evaluating IR systems for the RAG scenario is that, unlike humans who sequentially scan through retrieved documents one by one until finding relevant information or abandoning their search, LLMs process the entire prompt holistically, consuming all retrieved documents as a whole. Furthermore, the LLMs' ability to leverage the input context is affected by the positional bias. This contradicts the human-centric reward schema characterizing traditional IR metrics such as MRR, nDCG, or MAP, where relevant documents provide a contribution that decreases with their position in the ranking.

Figure 1 illustrates this in a controlled experiment where a single relevant passage is moved in a context of irrelevant ones. Experiments use Qwen 7B on 500 questions from PopQA, NQ, and TriviaQA. LLM accuracy exhibits the U-shaped

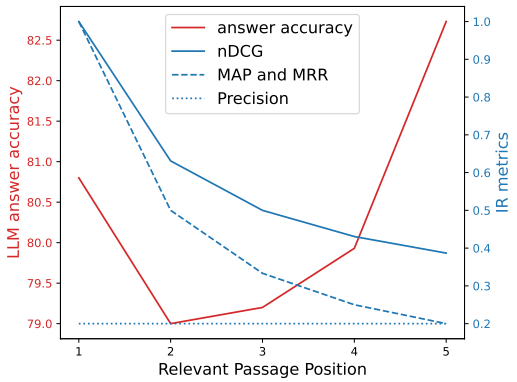

Figure 1: Misalignment between traditional IR metrics (blue) and LLM answer accuracy (red) when a single relevant passage is shifted among irrelevant passages.

lost-in-the-middle effect, while nDCG, MAP, and MRR decrease monotonically[5]; Precision, being position-unaware, remains constant.

4 An LLM-oriented Annotation Schema

Building on the discussion in the previous section, we propose a novel RAG-oriented annotation schema that replaces the human-centric notion of passage relevance with passage utility for an LLM. Similarly to how the distracting effect is computed, we assess the utility of a passage p for a question q as the probability of the LLM not abstaining when receiving that passage. If the passage is relevant, i.e., if it contains the answer, we interpret this probability as the likelihood that the LLM infers such an answer from the passage[6]. When the passage is irrelevant, the non-abstention probability corresponds to the distracting effect (Eq. 1) formalized in Amiraz et al. (2025), and we interpret it as a negative utility. More formally, given a question q and a passage p, the passage utility u is defined as:

$$u(q,p) = R(q,p)\Big(1 - p^{\text{LLM}}(\text{NO-RESPONSE}|q,p)\Big) \quad (2)$$

Where $R(q,p)$ is 1 if the passage p is relevant to the question q, -1 otherwise. Such a relevance label can be human-annotated or machine-generated using the LLM-as-a-judge approach.

Note that for a ranked list of k passages, obtaining utility scores for the k individual passages is

[3]Refer to Section 6 for details on LLMs, benchmarks, and answer evaluation.

[4]Henceforth, we assume all metrics have a cutoff at k.

[5]With one relevant passage, MRR equals MAP.

[6]We assume that relevant passages do not include distracting information. Therefore, when prompted with a lone relevant passage, the LLM either answers correctly or abstains, and never produces wrong answers. Although potentially possible, we never observed this case in preliminary experiments.

computationally cheaper than applying an end-to-end evaluation, i.e., running the LLM generator on the entire list and assessing the quality of its response. This is because the runtime cost of a vanilla transformer (Vaswani et al., 2017) scales quadratically with its input length. Consequently, if we assume that each passage has an average length of w tokens, and each answer consists on average of g tokens, generating the LLM answer using the full list would cost $\mathcal{O}(gk^2w^2)$; conversely, to obtain the utility of each passage[7], we would need to invoke k times the LLM with a shorter input and only generate the probability distribution of the first token, for a total cost of $\mathcal{O}(kw^2)$.

5 Defining a RAG-tailored Metric

Given a question q, the k retrieved passages $\mathcal{C} = [p_1, \ldots, p_k]$ and their utilities $u_i = u(q, p_i)$ to an LLM, we formalize **Utility and Distraction-aware Cumulative Gain** (UDCG), a metric to score how the entire context $\mathcal{C}$ is expected to help the LLM to answer q. We propose two variants of this metric. The former is a learnable version (UDCG_θ) that holistically accounts for the concepts of passage relevance, distracting effect, and LLM positional bias and is defined as follows:

$$\text{UDCG}_\theta(q, \mathcal{C}) = \sigma\left(\sum_{i=1}^{k} \alpha_i u_i^+ + \sum_{i=1}^{k} \beta_i u_i^-\right) \quad (3)$$

where u^+ and u^- are the positive and negative parts of u, respectively, and isolate the relevance and distracting effect of the passages; α_i and β_i are positional weights associated with the relevance and the distracting effect of the passages, respectively; finally $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function, which is used to bound the range of the metric in [0, 1] and facilitates the average across multiple questions. In short, UDCG_θ calculates a weighted sum of the relevance and the distracting effect of the passages, where the weights take into account the LLM positional bias. For contexts having k passages, the metric has $2k$ parameters, i.e., the $\alpha_1, \ldots \alpha_k$ and the $\beta_1, \ldots \beta_k$, which can be learned using a linear model on feature vectors containing $[u_1^+, \ldots u_k^+, u_1^-, \ldots u_k^-]$ trained to predict the end-to-end answer accuracy.

UDCG_θ has two drawbacks: (*i*) it requires a training process to learn its $2k$ parameters; and (*ii*) it strictly depends on the number of retrieved passages k, which means that each value of k requires a dedicated training process. To facilitate the metric applicability, we propose its training-free version, named just UDCG for simplicity:

$$\text{UDCG}(q, \mathcal{C}) = \sigma\left(\frac{1}{k}\sum_{i=1}^{k} u_i^+ + \frac{\gamma}{k}\sum_{i=1}^{k} u_i^-\right) \quad (4)$$

This version neglects the LLM's positional bias and performs a linear combination of the average passage relevance and the average distracting effect. There is a single hyperparameter $\gamma \in [0, 1]$ that balances the contribution of these two terms. It can be tuned for each scenario, but in our experiments, we show that $\gamma = 1/3$ provides good results on multiple LLMs and datasets.

6 Experimental Setting

6.1 Datasets and Models

We run experiments using five question-answering datasets spanning general knowledge and specialized domains. For general knowledge evaluation, we use PopQA (Mallen et al., 2023), the English subset of NoMIRACL (Thakur et al., 2024), and the KILT versions (Petroni et al., 2021) of Natural Questions (NQ) (Kwiatkowski et al., 2019) and TriviaQA (Joshi et al., 2017). To evaluate performance in specialized domains, we additionally include the factoid subset of BioASQ (Tsatsaronis et al., 2015) for biomedical question answering. For the Wikipedia-based datasets, we use the KILT corpus[8]; for BioASQ, we use PubMed[9]. We index the corpora using BGE-large-en-v1.5 embedding model (Chen et al., 2024a), with the only exception of NoMIRACL, which provides pre-retrieved passages with relevance annotations.

As LLMs for answer generation, we use the instruction-tuned versions of Llama-3.2-3B (L-3B), Llama-3.1-8B (L-8B), Llama-3.3-70B (L-70B) (Grattafiori et al., 2024), Mistral-7B (M-7B) (Jiang et al., 2023), Gemma-3-4B (G-4B) (Team et al., 2025), and Qwen-2.5-7B (Q-7B) (Yang et al., 2025), spanning different model sizes and families[10]. We employ greedy decoding for reproducibility.

[7] We assume passage relevance annotations are provided (common in IR datasets). Otherwise, they can be computed via LLM-as-a-judge at a similar cost to the abstention probability.

[8] https://huggingface.co/datasets/facebook/kilt_wikipedia

[9] https://pubmed.ncbi.nlm.nih.gov/

[10] For brevity, we omit version numbers when referring to models (e.g., Qwen 7B refers to Qwen-2.5-7B).

6.2 Evaluation Strategy

Following established practices in RAG evaluation (Zheng et al., 2023; Gu et al., 2025; Rahmani et al., 2024), we assess passage relevance and answer quality using the LLM-as-a-judge approach. For passage relevance annotation, we use Claude 3.7 Sonnet via AWS Bedrock to classify each retrieved passage as relevant or irrelevant by providing both the query and the reference answer to the model (see Figure 4). We consider a passage relevant only if it contains the necessary information to answer the query. For NoMIRACL, we utilize the relevance annotations directly provided in the dataset. We then apply our utility-based scoring mechanism to quantify the effects of different passages on LLM effectiveness, as described in Section 4.

For the quality evaluation of the responses, we follow previous work on the evaluation of high-fidelity RAG (Thakur et al., 2024; Chen et al., 2024b) and adopt the setting where LLMs are explicitly asked to abstain when sufficient information is not present in the input context (see Figure 3). In this framework, a response is correct only if the input context to the LLM contains at least a relevant passage and its response matches the reference answer (see Figure 5); we evaluate the semantic equivalence of the answer with Gemini 2.0 Flash.

This abstention requirement serves multiple purposes: it aligns with real-world deployment scenarios where RAG systems must balance informativeness with reliability, particularly in high-stakes domains (e.g., biomedicine) where incorrect responses can be more harmful than acknowledged uncertainty (Sun et al., 2025); it ensures that responses are only provided when supported by the retrieved content, thereby enhancing faithfulness and user transparency (Wallat et al., 2025); and it provides clearer attribution of effectiveness to the retrieval component by preventing LLMs from relying solely on their parametric knowledge when retrieved context is insufficient (Xu et al., 2024).

7 Experimental Results

7.1 Ideal Rankings of Different Annotations

To show the potential of our annotation schema, we sample 1000 questions from each dataset and retrieve the top 25 passages using BGE-large-en-v1.5. Then we annotate them using (*i*) binary relevance labeling, (*ii*) the eRAG approach (Salemi and Zamani, 2024), which assigns a utility score to each passage by computing the ROUGE-L (Lin, 2004) similarity between the reference answer and the answer generated by the LLM when prompted only with that passage, and (*iii*) our utility-based annotation schema according to Eq. 2. For each schema, we assume to have an oracle re-ranker that selects the best five passages based on their annotations[11] and prompt the LLM to generate the response based on such passages.

For 87% of the queries across the five benchmarks, there is at least a relevant passage among the 25 we retrieved; by selecting passages based on the binary relevance labeling or based on our annotation schema, relevant passages are always ranked better than irrelevant ones; therefore, a perfect LLM would provide 87% correct responses and abstain in the remaining 13% of the cases. Conversely, the passage selection according to the eRAG annotation schema sometimes favors irrelevant passages over relevant ones, resulting in only 71% to 77% of the contexts that include at least a relevant passage among the selected five, depending on the LLM; this significantly lowers the upper-bound answer correctness achievable with eRAG.

	Binary Rel.			eRAG			Our Method		
Model	**C↑**	**A**	**W↓**	**C↑**	**A**	**W↓**	**C↑**	**A**	**W↓**
Llama 3B	71.6	13.4	15.0	60.5	23.0	16.5	**72.9**	19.0	**8.2**
Llama 8B	73.5	15.9	10.6	64.4	23.6	12.0	**75.1**	20.0	**4.9**
Llama 70B	78.4	12.1	9.5	65.9	23.7	10.3	**80.2**	16.7	**3.0**
Gemma 4B	74.2	7.6	18.2	66.6	11.7	21.7	**76.0**	13.1	**10.9**
Mistral 7B	76.2	10.5	13.3	62.7	20.8	16.4	**76.9**	16.5	**6.6**
Qwen 7B	72.1	18.5	9.4	62.1	26.3	11.5	**73.5**	21.8	**4.8**

Table 2: Percentage of correct (C) answers, abstentions (A), and wrong (W) answers when LLMs are prompted with 5 ideal passages based on Binary Relevance, eRAG, and our utility-based annotations. Results averaged across all tested datasets. All differences are statistically significant (Wilcoxon, $p < 0.05$).

Table 2 reports the average results across the five benchmarks. Our annotation schema always improves results significantly, demonstrating that achieving perfect traditional IR metrics is insufficient for optimal RAG performance, calling for the development of new evaluation frameworks that better align with the objectives of modern retrieval-augmented generation systems. In particular, our annotation schema, by selecting all possible relevant passages and completing the context with the

[11] For traditional relevance labeling, we order the passages based on their binary relevance labels, with relevant passages appearing before irrelevant ones. For passages sharing the same relevance status (either all relevant or all irrelevant), we applied secondary sorting based on their retrieval scores.

	Llama 3B					Llama 8B					Llama 70B				
Method	**NQ**	**Trivia**	**PopQA**	**BioASQ**	**NoMIR**	**NQ**	**Trivia**	**PopQA**	**BioASQ**	**NoMIR**	**NQ**	**Trivia**	**PopQA**	**BioASQ**	**NoMIR**
NDCG	0.458	0.545	0.438	0.618	0.326	0.515	0.623	0.499	0.699	0.332	0.560	0.708	0.669	0.752	0.509
MRR	0.452	0.543	0.436	0.617	0.315	0.510	0.619	0.501	0.698	0.323	0.560	0.704	0.667	0.751	0.503
MAP	0.454	0.543	0.438	0.617	0.320	0.513	0.620	0.499	0.697	0.329	0.558	0.707	0.667	0.751	0.506
PREC	0.482	0.600	0.462	0.631	0.373	0.546	0.671	0.532	0.714	0.369	0.591	0.734	0.682	0.769	0.539
HITS	0.411	0.559	0.444	0.590	0.358	0.509	0.638	0.514	0.716	0.359	0.570	0.741	0.662	0.804	0.550
eRAG	0.324	0.219	-0.022	0.290	0.017	0.356	0.265	0.007	0.300	0.031	0.308	0.140	0.010	0.306	-0.053
UDCG	**0.536**	0.641	**0.525**	0.665	0.482	0.611	0.711	0.599	0.718	0.503	**0.665**	0.803	0.707	0.802	**0.662**
UDCG$_\theta$	0.532	**0.641**	0.512	**0.668**	**0.487**	**0.615**	**0.713**	**0.610**	**0.743**	**0.503**	0.661	**0.806**	**0.714**	**0.811**	0.653
	Gemma 4B					**Mistral 7B**					**Qwen 7B**				
Method	**NQ**	**Trivia**	**PopQA**	**BioASQ**	**NoMIR**	**NQ**	**Trivia**	**PopQA**	**BioASQ**	**NoMIR**	**NQ**	**Trivia**	**PopQA**	**BioASQ**	**NoMIR**
NDCG	0.557	0.572	0.514	0.728	0.531	0.548	0.720	0.438	0.734	0.567	0.471	0.478	0.236	0.665	0.292
MRR	0.558	0.575	0.518	0.732	0.524	0.546	0.721	0.440	0.735	0.560	0.475	0.474	0.231	0.668	0.283
MAP	0.555	0.570	0.513	0.728	0.527	0.544	0.718	0.437	0.733	0.564	0.469	0.474	0.232	0.663	0.288
PREC	0.572	0.595	0.560	0.759	0.561	0.589	0.752	0.503	0.754	0.600	0.510	0.552	0.296	0.699	0.329
HITS	0.554	0.592	0.524	0.758	0.576	0.542	0.742	0.456	0.768	0.606	0.463	0.459	0.201	0.689	0.305
eRAG	0.408	0.316	-0.130	0.444	0.023	0.361	0.318	-0.039	0.321	0.017	0.316	0.226	-0.046	0.343	-0.027
UDCG	0.587	0.681	**0.607**	0.776	**0.626**	0.629	**0.786**	0.542	0.767	0.609	**0.605**	0.649	0.397	0.752	**0.562**
UDCG$_\theta$	**0.593**	**0.682**	0.598	**0.793**	0.618	**0.634**	0.783	**0.556**	**0.777**	**0.632**	0.601	**0.650**	**0.406**	**0.755**	0.550

Table 3: Spearman correlation between IR metrics and end-to-end RAG accuracy. Bold indicates best per model-dataset combination. UDCG improvements are statistically significant (bootstrap, $p < 0.05$).

irrelevant ones that minimize the distracting effect, facilitates the LLM task of inferring the right answer when present in the passage context, leading to up to 2% more correct answers than the binary relevance labeling. Additionally, when no relevant passage is present, the absence of hard distracting passages helps the LLM to abstain, reducing by half the wrong answers. The improvement w.r.t. eRAG is even more evident, and we argue that the main reasons for this difference are (*i*) eRAG neglects the damaging impact of distracting documents, and (*ii*) eRAG uses ROUGE-L, which is a weak estimator of the answer quality, as other token-level answer equivalence measures (Bulian et al., 2022).

7.2 Correlation between IR Metrics and Answer Accuracy

To validate the effectiveness of our proposed metrics, we conducted comprehensive experiments examining their correlation with the end-to-end answer accuracy of the RAG system. Specifically, given a question q, we create 10 different contexts $\mathcal{C}_1, \ldots, \mathcal{C}_{10}$, each one containing 5 passages, which we use to generate the corresponding LLM answers $a_1, \ldots, a_{10}$. Then we assess the correctness of each answer and sort the contexts so that contexts leading to correct answers are positioned first, followed by contexts leading to abstation, followed by contexts leading to wrong answers. We compare this ideal context ranking with the one induced by a metric and compute their Spearman correlation. We create contexts by randomly sampling passages from the top-25 retrieved ones. We balanced the resulting examples so that 50% of the contexts contain at least a relevant passage, while the remaining 50% contain only irrelevant passages.

To learn the $\theta = \{\alpha_1, \ldots \alpha_k, \beta_1, \ldots \beta_k\}$ parameters of UDCG$_\theta$, we use a linear SVM^{rank} model (Joachims, 2002) with default hyperparameters[12]. As a training set, we use 1200 query-context examples derived from NQ, and tested on 500 query-context examples extracted from a disjoint set of NQ questions. Furthermore, we assess whether the learned metric generalizes to other datasets, by testing on 500 query-context samples of TriviaQA, PopQA, NoMIRACL, and BioASQ.

Table 3 reports the results of different LLMs. Our proposed metrics, UDCG$_\theta$ and its training-free version UDCG, consistently achieve the best correlation with the end-to-end RAG accuracy, across all LLMs and benchmarks, with UDCG$_\theta$ achieving a 36% (+10pts) increase in terms of Spearman correlation when compared to NDCG (averaging across all considered LLMs and datasets). UDCG$_\theta$, although trained only on NQ, maintains competitive results on all benchmarks, demonstrating its generalized applicability.

Surprisingly, the difference between UDCG and UDCG$_\theta$ is minor; conversely to UDCG, UDCG$_\theta$ accounts for the passage positions in the prompt, but we argue that this information is not very im-

[12] https://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

Metric	L-3B	L-8B	L-70B	G-4B	M-7B	Q-7B
UDCG$_\theta$	**0.532**	**0.615**	**0.661**	**0.593**	**0.634**	**0.601**
(rel-only)	0.508	0.594	0.625	0.567	0.607	0.553
(binary)	0.473	0.542	0.589	0.554	0.584	0.498
UDCG	**0.536**	**0.611**	**0.665**	**0.587**	**0.629**	**0.605**
(rel-only)	0.515	0.599	0.638	0.572	0.606	0.553

Table 4: Ablation study. Spearman correlation with RAG accuracy. UDCG consistently outperforms relevance-only baselines.

portant to estimate the end-to-end accuracy. Indeed, as observed by Cuconasu et al. (2025), the effect of positional bias, which is often amplified in controlled settings, is marginal in real scenarios.

7.3 Ablation Study for UDCG

We examine the contribution of each component in our proposed metrics through an ablation study on the NQ dataset across all LLMs. Table 4 reports the Spearman correlation between metric scores and end-to-end RAG accuracy under different configurations. For UDCG$_\theta$, we train three variants independently: the full model (see Eq. 3) incorporating both relevance utility features (u_i^+) and distracting effect features (u_i^-); a relevance-only version, where we disable the distracting effect features; and a binary version where we additionally constrain the relevance utility features $u_i^+ = 1$ to be binary values reflecting the traditional binary relevance labeling. Removing the distracting effect (rel-only) causes consistent degradation across all models, with correlation drops ranging from 3.4% to 8%. The binary variant shows even larger degradation, with drops of up to 17%, highlighting the importance of our fine-grained utility-based annotations over binary relevance judgments. For the training-free version of UDCG, we compare the full model (see Eq. 4) against a relevance-only version ($\gamma = 0$), which completely removes the contribution of distracting passages[13]. Consistent with UDCG$_\theta$, removing distracting effects leads to an average correlation decrease of 4%. These results confirm that our utility-based scoring, which includes the distracting effects, is essential for accurately predicting RAG system effectiveness.

7.4 Sensitivity to Context Length

We evaluate whether our proposed metrics maintain consistent predictive power as the number of retrieved documents k varies. We conduct this study using Qwen 7B on NQ, systematically varying k from 1 to 10 passages. We limit our analysis to $k \leq 10$ because additional documents at lower ranks would likely consist of distractors with a negative impact on the generation quality (Cuconasu et al., 2025). Figure 2 presents the results. UDCG demonstrates remarkable stability across different context sizes, maintaining correlations between 0.554 and 0.704, while consistently outperforming all baseline metrics across all values of k. In contrast, traditional IR metrics show substantially more volatility, with NDCG and Precision exhibiting standard deviations that are respectively 1.6 and 1.38 times higher than UDCG[14].

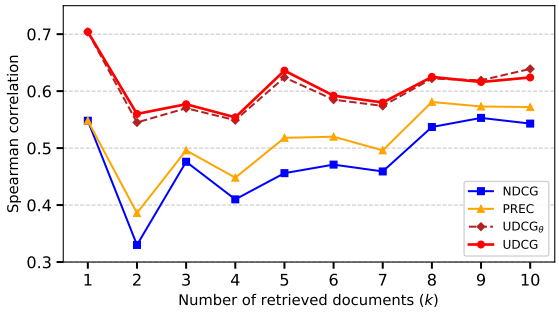

Figure 2: IR metric robustness across context sizes for Qwen 7B. UDCG maintains a stable correlation with RAG accuracy for all k.

8 Conclusions

This paper addresses a fundamental misalignment between traditional information retrieval evaluation metrics and the requirements of modern RAG systems. We demonstrated that classical IR metrics like nDCG, MAP, and MRR fail to adequately predict RAG performance because they assume monotonically decreasing document utility with rank position and ignore the heterogeneous distracting effects of irrelevant documents on LLM generation quality. To address these limitations, we proposed a novel RAG-oriented annotation schema that replaces the human-centric notion of passage relevance with positive or negative passage utility for an LLM. Building on this foundation, we defined UDCG, a metric using an LLM-oriented positional discount to directly optimize the correlation with the end-to-end answer accuracy. Notably, we have shown that removing positional discounting entirely achieves nearly identical performance.

[13] Note that further substituting continuous utility with binary relevance yields Precision (see Table 3).

[14] UDCG$_\theta$ shows similar stability but requires training for each k, making it less practical for dynamic context sizes.

Our extensive experiments on five QA datasets and six LLMs demonstrate that both metrics consistently outperform traditional IR metrics and existing RAG-oriented approaches in predicting end-to-end system performance by up to 36%.

Limitations

Our work has some limitations that suggest directions for future research. First, our experiments focus on the question-answering task where queries have specific and verifiable answers. The applicability of our metrics to other tasks, such as multi-hop question answering or fact verification, remains unexplored. Second, our utility-based annotation schema requires access to the LLM's output probabilities to compute abstention likelihood. This requirement limits our approach to settings where logit access is available, excluding black-box commercial APIs that only return generated text. Third, our evaluation is conducted exclusively on English-language datasets. The effectiveness of our utility-based scoring and the UDCG metric in multilingual or cross-lingual RAG settings remains to be investigated, as language-specific characteristics may influence both passage utility and distracting effects.

References

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)*, pages 18228–18258, Vienna, Austria. Association for Computational Linguistics.

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*, pages 291–305, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. BGE M3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024b. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

Florin Cuconasu, Simone Filice, Guy Horowitz, Yoelle Maarek, and Fabrizio Silvestri. 2025. Do rag systems really suffer from positional bias? *Preprint*, arXiv:2505.15561.

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, page 719–729, New York, NY, USA. Association for Computing Machinery.

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, page 407–416, New York, NY, USA. Association for Computing Machinery.

Lu Dai, Yijie Xu, Jinhui Ye, Hao Liu, and Hui Xiong. 2025. Seper: Measure retrieval utility through the lens of semantic perplexity reduction. In *The Thirteenth International Conference on Learning Representations*.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. 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*, pages 6491–6501.

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. *Preprint*, arXiv:2312.10997.

Laura A. Granka, Thorsten Joachims, and Geri Gay. 2004. Eye-tracking analysis of user behavior in www search. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, page 478–479, New York, NY, USA. Association for Computing Machinery.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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. *Preprint*, arXiv:2411.15594.

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, Lucca, Italy, April 6–10, 2025, Proceedings, Part I*, page 247–261, Berlin, Heidelberg. Springer-Verlag.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

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*.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, page 133–142, New York, NY, USA. Association for Computing Machinery.

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)*, pages 1601–1611.

Mohammed Abdul Khaliq, Paul Yu-Chun Chang, Mingyang Ma, Bernhard Pflugfelder, and Filip Miletić. 2024. RAGAR, your falsehood radar: RAG-augmented reasoning for political fact-checking using multimodal large language models. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 280–296, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1).

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Richard James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, Luke Zettlemoyer, and Scott Yih. 2024. RA-DIT: Retrieval-augmented dual instruction tuning. In *The Twelfth International Conference on Learning Representations*.

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*, 12:157–173.

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 *The 61st Annual Meeting Of The Association For Computational Linguistics*.

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

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*, pages 2523–2544, Online. Association for Computational Linguistics.

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 model for evaluation in information retrieval (llm4eval 2024) at sigir 2024. *Preprint*, arXiv:2408.05388.

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*, SIGIR '24, page 2395–2400, New York, NY, USA. Association for Computing Machinery.

Xin Sun, Jianan Xie, Zhongqi Chen, Qiang Liu, Shu Wu, Yuehe Chen, Bowen Song, Zilei Wang, Weiqiang Wang, and Liang Wang. 2025. Divide-then-align: Honest alignment based on the knowledge boundary of RAG. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11461–11480, Vienna, Austria. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

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*, pages 12508–12526, Miami, Florida, USA. Association for Computational Linguistics.

Fangzheng Tian, Debasis Ganguly, and Craig Macdonald. 2025. Is relevance propagated from retriever to generator in rag? *Preprint*, arXiv:2502.15025.

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, and 3 others. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16:138.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jonas Wallat, Maria Heuss, Maarten de Rijke, and Avishek Anand. 2025. Correctness is not faithfulness in retrieval augmented generation attributions. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, ICTIR '25, page 22–32, New York, NY, USA. Association for Computing Machinery.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, Miami, Florida, USA. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context. In *The Twelfth International Conference on Learning Representations*.

Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. 2025. *Evaluation of Retrieval-Augmented Generation: A Survey*, page 102–120. Springer Nature Singapore.

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Peixin Cao, Kaixin Ma, Jian Li, Hongwei Wang, and Dong Yu. 2024. Chain-of-note: Enhancing robustness in retrieval-augmented language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14672–14685, Miami, Florida, USA. Association for Computational Linguistics.

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*.

A Traditional IR Metrics

The offline evaluation of information retrieval systems relies heavily on a well-established set of metrics that measure different aspects of retrieval performance. These traditional metrics have relevance as a central notion, which represents the degree to which a retrieved document satisfies a user's information need. The conventional approach to relevance in IR has been predominantly binary, where documents are classified as either relevant or non-relevant to a given query. The IR metrics proposed in the literature mostly differ in how they account

for the presence and positioning of relevant documents in the retrieved document rank.

Precision and Recall serve as the foundational metrics in IR evaluation. Precision measures the fraction of retrieved documents that are relevant, while recall measures the fraction of relevant documents that are successfully retrieved. For modern (web-scale) information retrieval, recall is no longer a meaningful metric, as many queries have thousands of relevant documents, and few users will be interested in reading all of them. Precision@k ($P@k$), by measuring precision at fixed cutoff points, reflects practical constraints where users typically examine only the top-k results.

Mean Average Precision (MAP) extends previous metrics by incorporating ranking information. MAP computes the average precision at each relevant document position and then averages across all queries in a test collection. This metric effectively rewards systems that rank relevant documents higher in the result list, making it particularly suitable for evaluating systems where users examine results sequentially. More formally:

$$\text{MAP} = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{R_q} \sum_{k=1}^{n} P@k \times r(k)$$

where Q is the set of test queries, R_q is the number of existing relevant documents for query q, n is the number of retrieved documents, and $r(k)$ is an indicator function equaling 1 if the item at rank k is a relevant document, zero otherwise.

Mean Reciprocal Rank (MRR) focuses specifically on the rank position of the first relevant document, making it particularly appropriate for navigational queries where users seek a single correct answer. MRR computes the reciprocal of the rank at which the first relevant document appears, namely rank_i, and averages this value across all queries:

$$\text{MRR} = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{\text{rank}_i}$$

Normalized Discounted Cumulative Gain (nDCG) enables the usage of graded relevance judgments rather than binary values. nDCG recognizes that documents can have varying degrees of relevance and that highly relevant documents should contribute more to the overall evaluation score. The metric applies a logarithmic discount function to reduce the contribution of documents at lower ranks, reflecting the realistic assumption that users are less likely to examine documents further down in the ranking. The DCG accumulated at a particular rank position p is defined as:

$$\text{DCG}_p = \sum_{i=1}^{p} \frac{rel_i}{\log_2(i+1)}$$

where rel_i is the graded relevance of the result at position i retrieved for the query q. This value is typically normalized to facilitate comparison across queries with different numbers of relevant documents. To this end, we sort documents of a result list by relevance, producing an ideal DCG at position p (IDCG_p), and use this value as a normalization factor:

$$\text{nDCG}_p = \frac{\text{DCG}_p}{\text{IDCG}_p}$$

The nDCG values for all queries can be averaged to obtain a measure of the average performance of a search engine's ranking algorithm.

Finally, HITS measures whether there is at least a relevant passage in the top-k retrieved ones.

B Open-source Alternative for Relevance

In Section 6.2, we describe our method to compute the relevance of a passage with an LLM-as-a-Judge approach. In our analysis, we used Claude 3.7 Sonnet; however, we also tested Llama-3.3-70B Instruct as a strong open-source alternative for our evaluation tasks. We sampled 300 queries from our datasets and evaluated relevance decisions for the top-5 passages retrieved by BGE using both models, analyzing a total of 1500 passages. For passage relevance assessment, the Cohen's Kappa coefficient between Claude and Llama yielded a score of 0.85, indicating substantial agreement according to standard interpretation guidelines (Landis and Koch, 1977). These results suggest that Llama-3.3-70B Instruct can serve as a reliable open-source alternative for passage relevance assessment in RAG experiments.

```
You are given a question and you must respond based on the provided documents. Respond directly
without providing any premise or explanation. If none of the documents contain the answer, please
respond with NO-RESPONSE. Do not try to respond based on your own knowledge.

Documents:
<document>

Question:
<question>

Answer:
```

Figure 3: Prompt for evaluating the utility of a document and for answer generation.

```
Determine if a document is RELEVANT or IRRELEVANT for answering a question. A document is
RELEVANT if it contains information that directly supports at least one acceptable answer.

RELEVANT examples:
- Q: "where does the story the great gatsby take place" | Answers: ['Long Island of 1922']
  Doc: "The Great Gatsby...follows characters living in West Egg on Long Island in summer of 1922"
  → Contains the exact answer

- Q: "when did korn's follow the leader come out" | Answers: ['August 18, 1998', 'Summer 1998']
  Doc: "Follow the Leader...was released on August 18, 1998"
  → Contains the exact release date

IRRELEVANT examples:
- Q: "where does the story the great gatsby take place" | Answers: ['Long Island of 1922']
  Doc: "While Long Island features prominently in American literature, the socioeconomic
  dynamics..."
  → Mentions Long Island but not as the story's setting

- Q: "who played bobby byrd in get on up" | Answers: ['Nelsan Ellis']
  Doc: "Critics praised the casting of Bobby Byrd and the chemistry between the main characters...
  → Discusses casting but doesn't name the specific actor

Evaluation steps:
1. Find the specific information needed to match an acceptable answer. Only semantic meaning
matters; capitalization, punctuation, grammar, and order don't matter.
2. Check if the document contains this information directly or through clear inference
3. Check for these common errors:
    - The document contains similar keywords/themes but not the actual answer
    - The document contains partial information that would need to be combined with external
    knowledge
    - The document discusses related topics but doesn't specifically answer the question

Here is a new example. Don't apologize or correct yourself if there was a mistake; we are just
trying to evaluate the relevance of the document.
```
Question: {question}
Acceptable answers list: {answers}
Document: {document}
```

Evaluate the document for this new question as one of:
A: RELEVANT
B: IRRELEVANT

Return a JSON object with the following format:
{
    "motivation": "Brief explanation (max 2 sentences)",
    "grade": "A" or "B"
}
```

Figure 4: Prompt template for document relevance assessment using Claude 3.7 Sonnet as judge.

```
Your job is to grade a predicted answer as CORRECT, INCORRECT, or ABSTAINED against a list of acceptable answers.

## Grading Rules

A predicted answer is CORRECT if:
- It contains the essential information from at least one acceptable answer
- It contains no contradictory information
- When the acceptable answer is a list, at least one item from that list is present

A predicted answer is INCORRECT if:
- It contradicts any acceptable answer
- It contains factually wrong information (even with hedging like "I think" or "possibly")

A predicted answer is ABSTAINED if:
- It explicitly refuses to answer or states uncertainty without providing substantive information
- It contains phrases like "I don't know", "I'm not sure", "I cannot answer", "NO-RESPONSE", "I'm uncertain"

## Examples

**CORRECT answers:**
```
Question: What are Barack Obama's children's names?
Acceptable: ['Malia Obama and Sasha Obama', 'Natasha Marian and Malia Ann']
"Malia and Sasha" (essential info present)
"I believe it's Malia and Sasha Obama" (hedging with correct info)
```

**INCORRECT answers:**
```
Question: What are Barack Obama's children's names?
Acceptable: ['Malia Obama and Sasha Obama']
"Only Malia" (incomplete when both required)
"Malia, Sasha, and Susan" (contains wrong info)
"He has no children" (contradicts acceptable answer)
```

**ABSTAINED answers:**
```
Question: What are Barack Obama's children's names?
"I don't know the names of his children"
"I'm not sure about this information"
"I cannot provide that information"
"I'm uncertain and don't want to guess"
"NO-RESPONSE"
```

## Important Notes
- Ignore capitalization, punctuation, grammar, and order.
- Don't penalize omitted information that's clearly inferred from the question.
- Don't penalize minor typos in names.
- Grade using only the question and acceptable answers provided, even if you know the answer is factually wrong.

---

Here is a new example. Don't apologize or correct yourself if there was a mistake; we are just trying to grade the answer.
```
Question: {question}
Acceptable answers list: {target}
Predicted answer: {predicted_answer}
```

Grade the predicted answer of this new question as one of:
A: CORRECT
B: INCORRECT
C: ABSTAINED

Return a JSON object with the following format:
{
 "motivation": "Your concise motivation for the grade here. Use maximum 2 sentences.",
 "grade": "A", "B", or "C"
}
```

Figure 5: Prompt template for answer correctness assessment using Gemini 2.0 Flash as judge.