








Benchmarking of Retrieval Augmented Generation: A Comprehensive Systematic Literature Review on Evaluation Dimensions, Evaluation Metrics and Datasets

Simon Knollmeyer^{1,*}^a, Oğuz Caymazer^{2,*}^b, Leonid Koval¹^c, Muhammad Uzair Akmal¹^d,
Saara Asif¹^e, Selvine G. Mathias¹^f and Daniel Großmann¹^g

¹Technische Hochschule Ingolstadt, Almotion Bavaria, Esplanade 10, Ingolstadt, Germany

²University of Münster, Department of Information Systems, Münster, Germany

{Simon.Knollmeyer, Leonid.Koval, MuhammadUzair.Akmal, Saara.Asif, SelvineGeorge.Mathias,


Keywords: Large Language Model, Retrieval Augmented Generation, Evaluation Dimensions, Evaluation Metrics, Datasets, Systematic Literature Review.


Abstract: Despite the rapid advancements in the field of Large Language Models (LLM), traditional benchmarks have proven to be inadequate for assessing the performance of Retrieval Augmented Generation (RAG) systems. Therefore, this paper presents a comprehensive systematic literature review of evaluation dimensions, metrics, and datasets for RAG systems. This review identifies key evaluation dimensions such as context relevance, faithfulness, answer relevance, correctness, and citation quality. For each evaluation dimension, several metrics and evaluators are proposed on how to assess them. This paper synthesizes the findings from 12 relevant papers and presents a concept matrix that categorizes each evaluation approach. The results provide a foundation for the development of robust evaluation frameworks and suitable datasets that are essential for the effective implementation and deployment of RAG systems in real-world applications.


1 INTRODUCTION


The rapid evolution of Artificial Intelligence (AI) especially in the field of Large Language Models (LLMs) attracts widespread attention due to their groundbreaking achievements in solving complex problems even surpassing the performance of humans in certain fields (Benbya et al., 2024; Bubeck et al., 2023; OpenAI et al., 2023). The speed of AI development in the area of LLMs outpaced methods to assess their performance and accuracy, leading to major flaws in existing traditional benchmarks to evaluate LLMs output through reliable metrics impeding their adoption (Hammond, 2024).


Despite their potential, LLMs face substantial challenges, particularly in fully grasping contextual factors such as unique technical requirements within a specific industries, yet understanding these factors is essential for effective decision-making (Benbya et al., 2024). Even the most powerful models such as GPT-4 struggle with hallucinations, lack of the ability to update itself, and limited context (Bubeck et al., 2023; OpenAI et al., 2023). Several researchers point out that these LLMs seem to rather memorize frequently occurring information encountered during their pre-training and struggle with infrequent information, i.e., which would typically occur in a specific industry (Kandpal et al., 2022; Mallen et al., 2022).


^a  <https://orcid.org/0009-0002-1429-6992>


^b  <https://orcid.org/0009-0003-4096-3784>

^c  <https://orcid.org/0000-0003-4845-6579>

^d  <https://orcid.org/0009-0007-3961-1174>

^e  <https://orcid.org/0009-0006-1284-5635>

^f  <https://orcid.org/0000-0002-6549-0763>

^g  <https://orcid.org/0000-0002-7388-5757>

* co-authors of the paper, contributed equally

Most promising and common to solve this problem is to augment LLM with non-parametric less common knowledge by providing them retrieved text chunks from an external database (Asai et al., 2024; Y. Gao et al., 2023; Mallen et al., 2022; Wang et al., 2023; Zhang et al., 2023). This approach is known as Retrieval Augmented Generation (RAG), and there are several different RAG paradigms, ranging from so-called naïve RAG to more advanced ones (Asai et al., 2023; Y. Gao et al., 2023; Ma et al., 2023). Research results suggest that LLM RAGs outperform LLMs, particularly in long-tail knowledge questions (Asai et al., 2023; Izacard et al., 2022; Ma et al., 2023; Mallen et al., 2022; Wang et al., 2023).

However, there is still uncertainty about the accuracy of such approaches due to the lack of comprehensive evaluation frameworks to provide evaluation dimensions and metrics to assess RAG LLMs (Y. Gao et al., 2023; Wang et al., 2023). Thus, we address the following research questions (RQ):

- RQ1: How to evaluate a RAG-enhanced LLM comprehensively across different dimensions and metrics?
- RQ2: What type of datasets are available for applying the dimensions and metrics?

Therefore, the research contribution of this paper lies in addressing the current research gap by providing a systematic overview of how to evaluate RAG pipelines comprehensively, offering insights into the development of robust evaluation dimensions, metrics and possible datasets (Y. Gao et al., 2023; Hammond, 2024; Wang et al., 2023).

The paper is structured as follows: The next section introduces RAG. Section three explains the chosen research method. Section four presents the evaluation framework. Finally, the last section provides the conclusions.

2 RETRIEVAL AUGMENTED GENERATION

Traditional pre-trained LLMs such as GPT and BERT encode knowledge within their parameters, but struggle with tasks requiring specific factual knowledge which is not present in their parameters (Y. Gao et al., 2023; Lewis et al., 2020). This problem is evident in the fact that even the most powerful models such as GPT-4 struggle with made-up facts known as hallucinations, a lack of ability to self-update and limited context (Bubeck et al., 2023;

OpenAI et al., 2023). Even further increasing the size of their parameters, i.e., the training dataset, in which the knowledge appears to be stored to include more information will be likely insufficient to address the issue of long-tail knowledge (Kandpal et al., 2022; Mallen et al., 2022).

Therefore, Lewis et al. (2020) proposed RAG by combining non-parametric memory with parametric memory that uses a dense vector index of external documents such as Wikipedia articles that can be dynamically accessed using a retriever. Research results comparing the performance of RAG LLM with standalone LLM, suggest for the former superior performance (Asai et al., 2023; Izacard et al., 2022; Lewis et al., 2020; Ma et al., 2023; Mallen et al., 2022; Wang et al., 2023).

A naïve RAG pipeline involves three main steps. Firstly, the documents containing specific information are indexed (Lewis et al., 2020). The most common method is to split the documents into smaller sections so-called chunks and store their embedding in a vector database (Y. Gao et al., 2023). In the second step, a given input query is likewise embedded and then compared with the passages in the vector database by calculating the similarity, returning a set of top-ranked chunks that are most relevant for the query (Y. Gao et al., 2023; Karpukhin et al., 2020). In the final step, the retrieved content and the query are combined and prompted into an LLM so that it can provide a coherent answer (Y. Gao et al., 2023; Lewis et al., 2020).

This naïve setup can be modified by applying different advanced methods relating to pre- or post-retrieval (Asai et al., 2023; Y. Gao et al., 2023; Ma et al., 2023). For instance, Ma et al. (2023) propose query rewriting as an advanced pre-retrieval method and report performance improvements.

Despite the use of advanced methods, the RAG approach can still be divided into the outlined steps of a naïve RAG for evaluation. However, there is a lack of evaluation dimensions and metrics on how to analyse and assess such systems, e.g., which evaluation dimensions to consider and what kind of metrics to calculate for which step (Y. Gao et al., 2023).

3 RESEARCH METHOD

The Systematic Literature Review (SLR) is a well-known and established research method within Information Systems (IS) research for reviewing scientific articles based on a search process (Bell et al., 2019; Paré et al., 2016). The term "systematic" means that the research steps should be

understandable, reproducible, and grounded in a structured process that minimizes potential researcher bias by providing a clear audit trail for decisions and conclusions (Bell et al., 2019). SLR is especially valuable when summarizing and comparing fragmented knowledge on a certain topic (Bell et al., 2019). Existing literature reveals a significant gap in comprehensive evaluation frameworks for assessing RAG systems (Y. Gao et al., 2023; Wang et al., 2023). Therefore, given the emerging and unexplored research on evaluation dimensions for RAG, the SLR is particularly appropriate.

This paper ensures rigorous documentation by using the literature search process proposed by Brocke et al. (2009), extended with the recommendations of Paré et al. (2016). It also follows the recommendation of Webster and Watson (2002) to use a concept matrix for structuring and comparing the results. The complete adopted literature search process is illustrated in Figure 1.

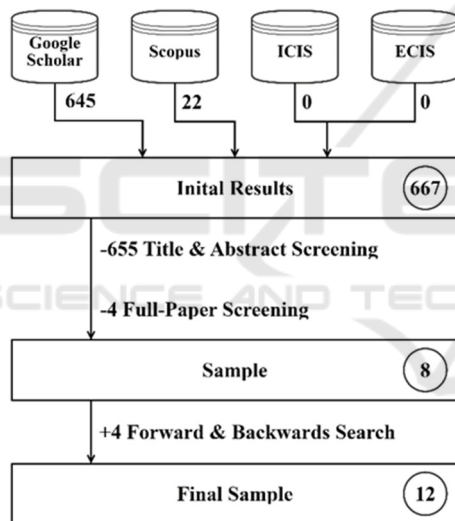


Figure 1: Systematic Literature Review process.

The search process included conducting a pre-study on Google Scholar by a detailed screening of titles, abstracts and full texts to identify papers specifically addressing evaluation metrics for RAG systems (Y. Gao et al., 2023; Wang et al., 2023). These insights resulted in the following search string:

- TITLE-ABS-KEY (“Retrieval Augmented Generation” AND “Evaluation Metric”)

The following inclusion and exclusion criteria were applied to filter relevant papers from the search results in Google Scholar, Scopus and the IS conferences ICIS and ECIS:

Inclusion Criteria:

- Papers from academic journals, conferences, or gray literature.
- Published in English.

Exclusion Criteria:

- Duplicates across databases.
- Minimal relevance to evaluation metrics, or lack of focus on RAG systems.
- No guidelines on metric implementation or application at the different RAG steps.

The search process commenced with an abstract screening of the initial results, followed by a full-text review of the selected papers. Forward and backward citation searches were subsequently performed on relevant studies to identify additional literature. This comprehensive approach yielded a final sample of 12 papers, as illustrated in Figure 1.

In the final step, overarching categories from the final sample were synthesized into a concept matrix (Webster & Watson, 2002). Relevant concepts on evaluation dimensions were identified and mapped to the RAG steps (cf. Section 2) to provide an accurate overview on how to evaluate each RAG phase.

The concept matrix is shown in Table 1. It categorizes the sampled papers based on predictive evaluation criteria and dataset characteristics. The former includes the columns "Retrieval" and "Generation" relating to the RAG steps. The "Evaluator" column indicates whether lexical matching, semantic similarity or LLM as a judge is used for evaluation. The dataset characteristics include single-hop and multi-hop reasoning tasks, synthetic datasets (triples), and open-domain question answering (QA). Each paper is marked ("X") to show the criteria it addresses, providing a comprehensive overview of their focus areas. In addition, the frequency of occurrence in the literature can be used to determine how widespread or accepted a metric is. Each proposed metric for the evaluation dimensions depending on the evaluator and the requirements for the data to calculate it are summarized in Table 2.

4 RESULTS

This section starts with examining the first column of the concept matrix: **predictive evaluation**, which involves assessing the performance of the RAG system in retrieving accurate context and effectively utilizing it to generate responses (Guinet et al., 2024).

Table 1: Concept Matrix with Evaluation Dimensions and Datasets.

Reference	Predictive Evaluation								Dataset			
	Retrieval Context Relevance	Generation				Evaluator			Retrieval		Generation	
		Faithfulness	Answer Relevance	Correctness	Citation Quality	Lexical Matching	Semantic Similarity	LLM as a Judge	Single- Hop	Multi- Hop	Synthetic Dataset (Triple)	Open- Domain QA
(Adlakha et al., 2023)		X	X		X	X	X	X	X			X
(Es et al., 2023)	X	X					X	X	(X)	X		
(T. Gao et al., 2023)			X	X	X			X	X			X
(Guinet et al., 2024)			X		X			X		X		
(Hu et al., 2024)		X			X	X		X	X			X
(Min et al., 2023)		X						X				(X)
(Rackauckas et al., 2024)		X			X	X		X	X			X
(Rau et al., 2024)					X	X		X	X			X
(Ravi et al., 2024)		X						X	X			X
(Saad-Falcon et al., 2023)	X	X						X		X		
(Yu et al., 2024)	X	X	X		X							
(Zhang et al., 2024)	X		X		X			X				

Table 2: Summary of proposed Evaluation Dimensions, the corresponding Metrics and Dataset Requirements.

Evaluation Dimension	Definition	Evaluator	Evaluation Metric	Dataset Requirement	Reference
Context Relevance	Relates to the retrieval step and measures the extent to which the retrieved context contains only the information required to answer the query and as little irrelevant information as possible.	Lexical Matching	$Recall@k = \frac{ \text{Relevant Passages} \cap k\text{-Passages} }{ \text{Relevant Passages} }$	Context, Reference Retrieval (Golden-Retrieval)	(Yu et al., 2024)
			$MRR@k = \frac{1}{ Q } \sum_{i=1}^k \frac{1}{rank_i}$		
Faithfulness	Refers to the generation step and measures the degree to which the LLM response is grounded in the retrieved context.	LLM as a judge	$CRS = \frac{\text{Number of extracted sentences } S_{ext}}{\text{Total Number of Sentences in } c(q)}$	Query-Context	(Es et al., 2023)
			$K\text{-Precision} = \frac{\text{Matched Tokens}}{\text{Response Tokens}}$		
			$FC = \frac{\sum_{i=1}^N \text{Faithfulness Score}_i}{\text{Total Number of Responses}}$		
Answer Relevance	Assesses whether the LLM response is directly addressing the actual query.	LLM as a judge	$FS = \frac{ V }{ S }$	Triple (Query-Context-Response)	(Es et al., 2023)
			$ARS = \frac{1}{n} \sum_{i=1}^n \text{sim}(q, q_i)$		
Correctness	Refers to the generation step and evaluates whether the LLM response answers the query accurately by matching with the expected answer, often referred to as “golden-passage” provided by human annotators.	Lexical Matching	$Recall = \frac{\text{Relevant Tokens in Response}}{\text{Total Relevant Tokens in Gold-Passage}}$	Reference Answer (Golden-Passage), Response	(Adlakha et al., 2023)
			$CC = \frac{\sum_{i=1}^N \text{Correctness Score}_i}{\text{Total Number of Responses}}$		
Citation Quality	focuses on assessing whether an LLM correctly cites its sources when generating text.	LLM as a judge	Citation recall and citation precision calculated by the judging LLM	Context-Response	(T. Gao et al., 2023)

The proposed evaluation pipeline outlines (cf. Figure 2) the process of assessing the RAG approach across various evaluation dimensions and focuses on the retrieval and generation stages of a typical RAG system. The evaluation process starts with the **retrieval step**, emphasizing **context relevance** as a critical dimension to assess how effectively relevant information is retrieved.

Subsequently, the focus then shifts to the **generation step**, examining the evaluation dimensions of **answer relevance**, **correctness**, **faithfulness**, and **citation quality** to determine the accuracy and reliability of the generated responses. Each evaluation dimension is carefully defined, and quantifying metrics are proposed according to the sampled papers.

How and what type of metric is ultimately used to calculate the respective evaluation dimension depends on the chosen **evaluator**. **Lexical matching** metrics focus on exact word matching and simple statistical calculations, such as keyword frequency or position-based measures like Mean Reciprocal Rank (MRR), i.e., these metrics assess how closely the words in the documents match the query without considering deeper meanings.

Semantic similarity metrics, on the other hand, go beyond surface-level text comparison to understand the underlying meanings and concepts by comparing their semantical similarity based on context and conceptual relationships between words and sentences. This approach captures the intent of the query and the documents, evaluating relevance through semantic similarity rather than just keyword occurrence.

Finally, **LLM as a judge** uses a LLM to evaluate content by making it context-aware, prompting the

model to consider the coherence, factuality, and relevance of the information based on its comprehensive understanding of language and knowledge. Therefore, the choice of the evaluator determines the type of metric applied to assess each evaluation dimension, depending on whether the focus is on exact word matching, conceptual similarity, or a nuanced, context-aware judgment by an advanced LLM.

The upcoming sub-sections follows the rationale of first explaining the evaluation dimension and then the respective evaluator by detailing how to calculate the metric proposed in the sampled papers.

4.1 Context Relevance

The evaluation dimension of **context relevance** pertains to the retrieval step and assesses the degree to which the retrieved context contains only the necessary information to answer the query, reducing computational costs and improving efficiency by minimizing irrelevant content (Es et al., 2023; Saad-Falcon et al., 2023; Yu et al., 2024). Additionally, when retrieved passages are too long, LLMs often struggle to effectively utilize the information, particularly if the relevant details are embedded in the middle of the passage (Es et al., 2023). Hence, concise query-relevant passages significantly improve the LLM generation quality (Es et al., 2023; Yu et al., 2024).

Recall@k and **MRR@k** are key **lexical metrics** for evaluating the retrieval performance in RAG systems (Rackauckas et al., 2024; Yu et al., 2024). Each metric provides a different perspective on the effectiveness of the retrieval process. **Recall@k**

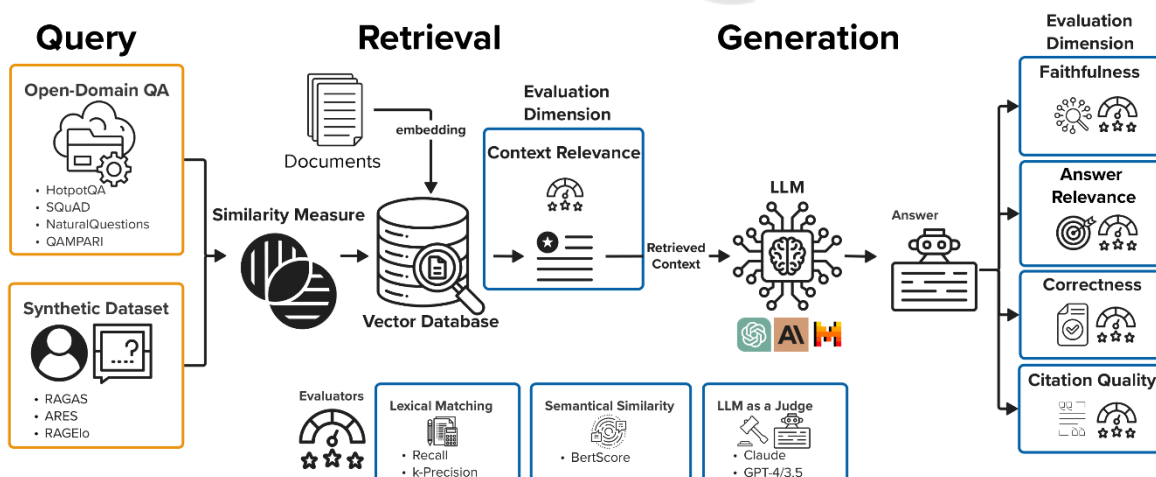


Figure 2: Evaluation Pipeline with Evaluation Dimensions in a naïve RAG setup.

measures how many relevant passages are captured within the top k retrieved chunks, even if some irrelevant ones are included. The formula is as follows:

$$\text{Recall@k} = \frac{|\text{Relevant Passages} \cap \text{k-Passages}|}{|\text{Relevant Passages}|} \quad (1)$$

MRR@k calculates context relevance by emphasizing the rank of the first relevant passage across multiple queries (Rackauckas et al., 2024). If a relevant passage appears in the top k results, its contribution to the **MRR@k** score is the inverse of its rank, e.g., a passage ranked two contributes as $\frac{1}{2}$ with **MRR@5** (Rackauckas et al., 2024). If no relevant passage is found in the top k the contribution is zero. The formula for **MRR@k** is:

$$\text{MRR@k} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i} \quad (2)$$

By using the **LLM as a judge**, it is possible to calculate an estimated context relevance score. Given a query q and its retrieved context $c(q)$, the LLM is prompted to extract a subset of sentences (S_{ext}) from $c(q)$ that are relevant to answering q by using the following *prompt*:

"Please extract relevant sentences from the provided context that can potentially help answer the following question. If no relevant sentences are found, or if you believe the question cannot be answered from the given context, return the phrase 'Insufficient Information'. While extracting candidate sentences, you're not allowed to make any changes to sentences from the given context." (Es et al., 2023)

The prompt instructs the LLM to select only the sentences that it considers relevant to q without changing the content. The **Context Relevance Score (CRS)** is calculated by dividing the relevant sentences extracted (S_{ext}) from the context $c(q)$ by the total number of sentences. This can be expressed with the following formula (Es et al., 2023):

$$\text{CRS} = \frac{\text{Number of extracted sentences } S_{ext}}{\text{Total number of sentences in } c(q)} \quad (3)$$

The CRS indicates the proportion of the context that is relevant. A higher score indicates that a greater proportion of the retrieved context is focused and relevant for answering the query, while a lower score indicates that much of the retrieved context contains irrelevant information (Es et al., 2023; Saad-Falcon et al., 2023; Yu et al., 2024).

4.2 Faithfulness

The evaluation dimensions **faithfulness** refers to the generation step and evaluates how well an LLM's response is grounded in the retrieved context, i.e., all information in the response can be directly inferred from it (Adlakha et al., 2023; Es et al., 2023; Hu et al., 2024). This evaluation dimension is crucial to identify possible hallucinations in the answer of LLMs ensuring factual correctness (Adlakha et al., 2023; Es et al., 2023; Ravi et al., 2024). For instance, Adlakha et al. (2023) found that GPT-4 had the highest agreement with human annotations, followed by GPT-3.5 and k -precision.

As the **primary lexical metric** Adlakha et al. (2023) propose **k -precision** to evaluate the degree of faithfulness since it has the highest agreement with human judgements. It can be calculated as the proportion of tokens in the LLMs response that are present in the retrieved context, i.e., it is the overlap of matching tokens with the retrieved context divided by the total number of tokens in the response (Adlakha et al., 2023). Hence, the formula is as follows:

$$\text{K-Precision} = \frac{\text{Matched Tokens}}{\text{Response Tokens}} \quad (4)$$

Another way of calculating the faithfulness is by using **LLMs** such as GPT-4/3.5 as **evaluators**. The LLMs are prompted to judge whether the response and the retrieved context match from an ordinal scale of "fully", "partially", or "not at all" (Adlakha et al., 2023). In the same way, Ravi et al. (2024) propose to assess the responses through an evaluator LLM whether the responses are supported, contradicted or not supported by the retrieved context. In order to quantify this ordinal scale, we can assign numerical scores depending on the degree of support and take the average of all individual response scores, i.e., assign 1 for "fully", 0.5 for "partially", and 0 for "not at all". The formula for calculating the **Faithfulness Coefficient (FC)** is as follows:

$$\text{FC} = \frac{\sum_{i=1}^N \text{Faithfulness Score}_i}{\text{Total number of responses}} \quad (5)$$

A similar method is used by Hu et al. (2024), who propose a framework that extracts "claim triplets" (subject, predicate, object) to represent fine-grained knowledge assertions within the LLM response. The purpose of this extraction is to break down the answer into specific atomic claims that can be checked individually (Hu et al., 2024; Min et al., 2023). A judging LLM evaluates each triplet as "entailment"

(supported), “contradiction” (contradicted), or “neutral” (unsupported) (Hu et al., 2024). Other authors also follow this approach of breaking down the statements from the LLM response into atomic facts to obtain a fine-grained measure of the faithfulness degree (Min et al., 2023).

Es et al. (2023) propose a process in which the answer $a_s(q)$ is considered faithful to the context $c(q)$ if the statements in the LLM response can be directly inferred from the retrieved context. The process begins by using a **LLM as a judge** to decompose the LLM response into a set of statements, $S(a_s(q))$, which involves breaking down longer sentences into shorter ones (Es et al., 2023). For each statement s_i in S , the judging LLM verifies if it can be inferred from the given context $c(q)$ using a verification function $v(s_i, c(q))$ (Es et al., 2023). The judging LLM assesses whether each statement is supported by the information in the retrieved context and provides a “yes” or “no” verdict for each statement (Es et al., 2023). **The Faithfulness Score (FS)** is then calculated as the ratio of the number of supported statements V to the total number of statements S , which can be expressed as follows:

$$FS = \frac{|V|}{|S|} \quad (6)$$

4.3 Answer Relevance

The evaluation dimensions **answer relevance** refers to the generation step and assesses whether the LLM response is directly addressing the query (Es et al., 2023; Rackauckas et al., 2024; Saad-Falcon et al., 2023; Yu et al., 2024). This evaluation dimension penalizes incomplete or redundant answers, **regardless of factuality** (Es et al., 2023; Rackauckas et al., 2024).

By using **LLM as a judge** it is possible to calculate an estimation of the answer relevance (Es et al., 2023; Rackauckas et al., 2024; Saad-Falcon et al., 2023). Given a generated answer $a_s(q)$, the judging LLM is prompted to generate n potential questions q_i that could be answered by using $a_s(q)$ (Es et al., 2023). This is done using the following prompt:

*Generate a question for the given answer:
answer:[answer] (Es et al., 2023)*

Subsequently, text embeddings for all generated questions q_i and the original query q are created to calculate in the next step the cosine similarity between their embeddings (Es et al., 2023). The **Answer Relevance Score (ARS)** is obtained by

averaging the similarity between the generated questions q_i and the original query q using the following formula:

$$ARS = \frac{1}{n} \sum_{i=1}^n \text{sim}(q, q_i) \quad (7)$$

where $\text{sim}(q, q_i)$ represents the cosine similarity between the embeddings of the generated questions q_i and the original query q (Es et al., 2023). This metric effectively measures how well the generated answer matches the intent and content of the original question (Es et al., 2023).

4.4 Correctness

The evaluation dimension **correctness** refers to the generation step and evaluates whether the LLM’s response accurately matches the “golden passage” provided by human annotators (Adlakha et al., 2023; T. Gao et al., 2023). This metric **focuses on the factual accuracy** of the information by comparing the LLM response with a reference answer (Adlakha et al., 2023; T. Gao et al., 2023; Guinet et al., 2024; Rackauckas et al., 2024).

As a **primary lexical metric** Adlakha et al. (2023) propose using **recall** as it correlates well with human annotations. Traditional metrics like Exact Match (EM), F1, and ROUGE are often too strict due to their focus on exact word matching (Adlakha et al., 2023). Recall measures how much of the reference answer’s essential content is captured in the model’s response without penalizing additional information (Adlakha et al., 2023).

Some authors find that **semantic similarity** metrics like BERTScore is less effective than recall for correctness due to lower alignment with human annotations (Adlakha et al., 2023; T. Gao et al., 2023). These metrics do not account for factual accuracy or logical consistency, as responses can be textually similar but factually incorrect (Adlakha et al., 2023; T. Gao et al., 2023).

By using **LLMs** such as GPT-3.5 and GPT-4 as **evaluators to judge** the correctness of responses by prompting them with the question, the reference answer, and the LLMs response to determine whether the model response is correct, partially correct, or incorrect (Adlakha et al., 2023; Rackauckas et al., 2024). This approach seems to yield the highest agreement with human annotations (Adlakha et al., 2023; Rackauckas et al., 2024). Correctness can be quantified by assigning scores: 1 for “fully correct”, 0.5 for “partially”, and 0 for “incorrect”. The

Correctness Coefficient (CC) is the average of all response scores. Therefore, the formula is as follows:

$$CC = \frac{\sum_{i=1}^N \text{Correctness Score}_i}{\text{Total number of responses}} \quad (8)$$

4.5 Citation Quality

The evaluation dimension **citation quality** focuses on assessing whether an LLM correctly cites its sources when generating text (T. Gao et al., 2023). Citation quality is calculated using two metrics: **citation recall** and **citation precision**. Citation recall ensures that every piece of information in the generated response is fully supported by the cited passages, while citation precision checks whether all cited passages are relevant and necessary for the statements made (T. Gao et al., 2023).

To perform this evaluation automatically, the **LLM acts as a judge**, which is prompted with a chain-of-thought method instead of simple lexical matching (T. Gao et al., 2023). The LLM checks if the concatenated text from the cited passages semantically supports the generated statements (T. Gao et al., 2023). For citation recall, it evaluates whether all generated statements are substantiated by the citations. A statement receives a recall score of 1 if it is fully supported by at least one citation, otherwise, it receives a score of 0 (T. Gao et al., 2023). For citation precision, the LLM identifies any "irrelevant" citations, i.e., those that do not independently support a statement or are not necessary when other citations already provide full support (T. Gao et al., 2023). A citation receives a precision score of 1 if it is relevant and contributes to the statement's support and 0 if it is irrelevant (T. Gao et al., 2023).

The robustness of these metrics is validated by their **strong correlation** with **human judgements** (T. Gao et al., 2023). By averaging the citation recall and precision scores across all statements and citations in the generated response, an overall citation quality score can be calculated, providing a comprehensive measure of how accurately and appropriately an LLM uses citations in its output.

5 DATASETS & APPLICATION

Building on the concept matrix's evaluation dimensions presented above, it's essential to consider how different datasets relate to the evaluators of these dimensions and its corresponding evaluation metrics. The choice between open-domain QA datasets and

synthetic datasets like RAGAS or ARES, along with the type of reasoning required (single-hop vs. multi-hop), plays a crucial role in ensuring the robustness and reliability of these evaluations.

Open-domain QA datasets such as SQuAD2.0, HotpotQA, and Natural Questions are based on Wikipedia articles. These are mainly suitable for **lexical matching** enable comparisons across RAG systems (Kwiatkowski et al., 2019; Rajpurkar et al., 2018; Yang et al., 2018). These datasets often include "golden passages" which make them ideal for evaluating **correctness** and **faithfulness** by providing a factual reference (Kwiatkowski et al., 2019; Rajpurkar et al., 2018; Yang et al., 2018). For instance, SQuAD2.0 includes unanswerable queries requiring RAG systems to recognize when there is insufficient information to provide a valid answer (Rajpurkar et al., 2018; Rau et al., 2024). Also calculating **context relevance** is straightforward because the datasets provide clear ground truth in terms of which passages are relevant, making them ideal for recall-oriented metrics like **Recall@k** or **MRR@k** (Adlakha et al., 2023; Hu et al., 2024). However, evaluating **answer relevance** and **citation quality** is more challenging with open-domain QA datasets since these typically focus on finding a single correct answer rather than assessing nuanced citation practices or multi-source relevance (Kwiatkowski et al., 2019; Rajpurkar et al., 2018; Yang et al., 2018).

Synthetic datasets such as RAGAS and ARES are specifically designed to evaluate the effectiveness of RAG systems by minimizing reliance on human annotations (Es et al., 2023; Saad-Falcon et al., 2023). These frameworks often use synthetic datasets that only require query-context-response triples, making them suitable to evaluate every evaluation dimension (Es et al., 2023; Hu et al., 2024; Min et al., 2023; Saad-Falcon et al., 2023). Synthetic datasets combined with LLM judges align well with human annotations, outperforming lexical and semantic similarity metrics (Adlakha et al., 2023; Saad-Falcon et al., 2023). Additionally, this approach is model-agnostic, allowing flexible use across different LLMs and setups (Es et al., 2023; Saad-Falcon et al., 2023).

This adaptability ensures that RAG systems can be effectively assessed and fine-tuned for diverse and complex queries, enhancing their performance in practical, real-world setting. Table 3 summarizes the key differences between the datasets by comparing them.

In terms of reasoning, **single-hop** and **multi-hop** queries require different approaches and datasets **Single-hop** reasoning involves deriving an answer from a single piece of evidence, i.e., one retrieved

Table 3: Comparing Open Domain QA Datasets with Synthetic Datasets for evaluating RAG.

Aspect	Open Domain QA Dataset	Synthetic Dataset
Examples	SQuAD2.0, HotpotQA, Natural Questions, QAMPARI	RAGAS, ARES
Reasoning Type	Single-hop (e.g., SQuAD2.0, Natural Questions) Multi-hop (e.g., HotpotQA, QAMPARI)	Single-hop and adaptable to Multi-hop
Evaluation Dimensions	Correctness, Faithfulness, Context Relevance (basic)	Correctness, Faithfulness, Context Relevance (multi-retrieval), Answer Relevance, Citation Quality
Evaluators	Lexical Matching	Semantic Similarity, LLM as a Judge
Strengths	High reliability for correctness and faithfulness due to golden-passages	Model and vendor agnostic, adaptable for various queries and rapid evaluation of RAG without the need for human annotations or gold-passages
Limitations	Less effective in evaluating multi-source citations, complex context relevance, and answer relevance in multi-hop	Associated costs related to token usage and potential latency due to LLM judging, different performances depending on the employed LLM model

passage, and is well-suited for the evaluation dimensions **correctness**, **faithfulness**, and **basic context relevance**. Popular single-hop datasets are Natural Questions (Kwiatkowski et al., 2019) and SQuAD2.0 (Rajpurkar et al., 2018), where the relevant information is contained within a single passage, allowing the calculation of metrics predominantly with lexical matching or simple semantic similarity, e.g., focusing on metrics such as precision and recall (Adlakha et al., 2023; Ravi et al., 2024). In contrast, **multi-hop** reasoning requires to connect multiple pieces of retrieval, usually from different documents or distant parts of the same document, to be combined in order to obtain a correct answer. This approach is better suited for evaluating **context relevance** for multiple retrieval, **answer relevance**, i.e., how the combined context informs the answer, and **citation quality** that correctly attributes information to multiple sources (Adlakha et al., 2023; Es et al., 2023; T. Gao et al., 2023; Hu et al., 2024). Open-domain QA datasets like HotpotQA (Yang et al., 2018) or QAMPARI (Amouyal et al., 2022) are specifically designed for multi-hop reasoning. These require synthesizing information from multiple retrieved contexts, which involves understanding complex connections and contextual relevance that go beyond surface-level comparisons. Employing LLMs as a judge for this evaluation is the most suitable option since LLMs can comprehend the combination of multiple contexts better by making more nuanced judgement than simple lexical matching or semantic similarity of concepts (Es et al., 2023; T. Gao et al., 2023; Min et al., 2023; Saad-Falcon et al., 2023).

6 PRACTICAL APPLICATION

In order to apply evaluation dimensions and select the appropriate datasets, it is necessary to understand dataset requirements. For instance, the evaluation dimension faithfulness requires data regarding the retrieved passages of the RAG to be tested and a reference retrieval (golden retrieval), but these are often not available in real-world operations. Therefore, synthetic datasets that are applicable reference-free would be more suitable for practical operation, and they also have a high alignment with human annotations (Es et al., 2023; Saad-Falcon et al., 2023).

It is also necessary to select suitable metrics according to the objective of optimizing the RAG system. For this purpose, a distinction should first be made as to whether the performance of the retrieval or generation step should be considered. A suitable metric is then selected according to a specific problem. For example, if the factuality of the answer is to be increased, correctness is a more suitable metric than faithfulness.

Finally, it is important to build an automated, robust and reliable evaluation pipeline that can be used to evaluate the RAG system (Es et al., 2023).

7 CONCLUSION

This paper proposes a comprehensive evaluation framework specifically for RAG by conducting an SLR and providing an extensive overview of currently existing evaluation approaches. Since the introduction of RAG in 2020 (Lewis et al., 2020), it has taken considerable time for methods to be

developed and established in the literature for evaluating these approaches. Despite the rigorous methodology, it remains possible that some papers were overlooked due to the rapid pace of developments in the field.

With reference to RQ1, our evaluation framework introduces robust evaluation dimensions and metrics to assess the different steps within RAG. Moreover, this paper advances the understanding of RAG evaluation by providing reliable dimensions and metrics (Y. Gao et al., 2023; Wang et al., 2023). Furthermore, Section 5 gives a comprehensive summary about RQ2 by providing an overview of the available datasets to apply the proposed evaluation dimensions and metrics and what kind of requirements to consider.

As a future avenue of research, a practical application of these metrics should be conducted to validate their use and alignment with human preferences. In addition, the provision of all the metrics presented could be examined within a framework to simplify practical application.

ACKNOWLEDGEMENTS

The presented paper was produced as part of the research project MoFaPro. This project is funded by the AUDI AG. The present approach was developed within the institute "AIMotion Bavaria" at the Technische Hochschule Ingolstadt. This work is part of Oğuz Caymazer's master's thesis and was conducted during his internship at the AUDI AG.

REFERENCES

- Adlakha, V., Behnamghader, P., Lu, x. H., Meade, n., & Reddy, S. (2023). *Evaluating correctness and faithfulness of instruction-following models for question answering*. <https://doi.org/10.48550/arxiv.2307.16877>
- Amouyal, S. J., Wolfson, T., Rubin, O., Yoran, O., Herzig, J., & Berant, J. (2022). *QAMPARI: An Open-domain Question Answering Benchmark for Questions with Many Answers from Multiple Paragraphs*. <https://doi.org/10.48550/arXiv.2205.12665>
- Asai, A., Wu, Z., Wang, Y [Yizhong], Sil, A., & Hajishirzi, H. (2023). *Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection*. <https://doi.org/10.48550/arXiv.2310.11511>
- Asai, A., Zhong, Z., Chen, D [Danqi], Koh, P. W., Zettlemoyer, L., Hajishirzi, H., & Yih, W. (2024). *Reliable, Adaptable, and Attributable Language Models with Retrieval*. <https://doi.org/10.48550/arXiv.2403.03187>
- Bell, E., Bryman, A., & Harley, B. (2019). *Business research methods* (Fifth edition). Oxford University Press.
- Benbya, H., Strich, F., & Tamm, T. (2024). Navigating Generative Artificial Intelligence Promises and Perils for Knowledge and Creative Work. *Journal of the Association for Information Systems*, 25(1), 23–36. <https://doi.org/10.17705/1jais.00861>
- Brocke, J., Simons, A., Niehaves, B., Riemer, K., Plattfaut, R., & Cleven, A. (2009). Reconstructing the Giant: On the Importance of Rigour in Documenting the Literature Search Process. In European Conference on Information Systems (Chair), *17th European Conf. on Information Systems*. <https://www.wi.uni-muenster.de/research/publications/3069>
- Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y [Yi]. (2023). *Sparks of Artificial General Intelligence: Early experiments with GPT-4*. <https://doi.org/10.48550/arXiv.2303.12712>
- Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2023). *RAGAS: Automated Evaluation of Retrieval Augmented Generation*. <https://doi.org/10.48550/arXiv.2309.15217>
- Gao, T., Yen, H., Yu, J., & Chen, D [Danqi]. (2023). *Enabling Large Language Models to Generate Text with Citations*. <https://doi.org/10.48550/arXiv.2305.14627>
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y [Yi], Sun, J., Wang, M., & Wang, H. (2023). *Retrieval-Augmented Generation for Large Language Models: A Survey*. <https://doi.org/10.48550/arXiv.2312.10997>
- Guinet, G., Omidvar-Tehrani, B., Deoras, A., & Callot, L. (2024). *Automated Evaluation of Retrieval-Augmented Language Models with Task-Specific Exam Generation*. <https://github.com/amazon-science/auto-rag-eval> <https://doi.org/10.48550/arXiv.2405.13622>
- Hammond, G. (2024, April 10). Speed of AI development stretches risk assessments to breaking point. *Financial Times*. <https://www.ft.com/content/499c8935-f46e-4ec8-a8e2-19e07e3b0438>
- Hu, X [Xiangkun], Ru, D., Qiu, L., Guo, Q., Zhang, T., Xu, Y., Luo, Y., Liu, P., Zhang, Y [Yue], & Zhang, Z [Zheng]. (2024). *RefChecker: Reference-based Fine-grained Hallucination Checker and Benchmark for Large Language Models*. <https://doi.org/10.48550/arXiv.2405.14486>
- Izacard, G., Lewis, P., Lomeli, M., Hosseini, L., Petroni, F., Schick, T., Dwivedi-Yu, J., Joulin, A., Riedel, S., & Grave, E. (2022). *Atlas: Few-shot Learning with Retrieval Augmented Language Models*. <https://doi.org/10.48550/arXiv.2208.03299>
- Kandpal, N., Deng, H., Roberts, A., Wallace, E., & Raffel, C. (2022). *Large Language Models Struggle to Learn Long-Tail Knowledge*. <https://doi.org/10.48550/arXiv.2211.08411>

- Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D [Danqi], & Yih, W. (2020). *Dense Passage Retrieval for Open-Domain Question Answering*. <https://doi.org/10.48550/arXiv.2004.04906>
- Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.-W., Dai, A. M., Uszkoreit, J., Le, Q., & Petrov, S. (2019). Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, 7, 453–466. https://doi.org/10.1162/tacl_a_00276
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. <https://doi.org/10.48550/arXiv.2005.11401>
- Ma, X., Gong, Y., He, P., Zhao, H., & Duan, N. (2023). *Query Rewriting for Retrieval-Augmented Large Language Models*. <https://doi.org/10.48550/arXiv.2305.14283>
- Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., & Hajishirzi, H. (2022). *When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories*. <https://doi.org/10.48550/arXiv.2212.10511>
- Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W., Koh, P., Iyyer, M., Zettlemoyer, L., & Hajishirzi, H. (2023). FAActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 12076–12100). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.emnlp-main.741>
- OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., . . . Zoph, B. (2023). *GPT-4 Technical Report*. <https://doi.org/10.48550/arXiv.2303.08774>
- Paré, G., Tate, M., Johnstone, D., & Kitsiou, S. (2016). Contextualizing the twin concepts of systematicity and transparency in information systems literature reviews. *European Journal of Information Systems*, 25(6), 493–508. <https://doi.org/10.1057/s41303-016-0020-3>
- Rackauckas, Z., Câmara, A., & Zavrel, J. (2024). *Evaluating RAG-Fusion with RAGElo: an Automated Elo-based Framework*. <https://doi.org/10.48550/arXiv.2406.14783>
- Rajpurkar, P., Jia, R., & Liang, P. (2018). *Know What You Don't Know: Unanswerable Questions for SQuAD*. <https://doi.org/10.48550/arXiv.1806.03822>
- Rau, D., Déjean, H., Chirkova, N., Formal, T., Wang, S., Nikoulina, V., & Clinchant, S. (2024). *BERGEN: A Benchmarking Library for Retrieval-Augmented Generation*. <https://doi.org/10.48550/arXiv.2407.01102>
- Ravi, S. S., Mielczarek, B., Kannappan, A., Kiela, D., & Qian, R. (2024). *Lynx: An Open Source Hallucination Evaluation Model*. <https://arxiv.org/abs/2407.08488>
- Saad-Falcon, J., Khattab, O., Potts, C., & Zaharia, M. (2023). *ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems*. <https://doi.org/10.48550/arXiv.2311.09476>
- Wang, C., Liu, X., Yue, Y., Tang, X., Zhang, T., Jiayang, C., Yao, Y., Gao, W., Hu, X [Xuming], Qi, Z., Wang, Y [Yidong], Yang, L., Wang, J [Jindong], Xie, X., Zhang, Z [Zheng], & Zhang, Y [Yue]. (2023). *Survey on Factuality in Large Language Models: Knowledge, Retrieval and Domain-Specificity*. <https://doi.org/10.48550/arXiv.2310.07521>
- Webster, J., & Watson, R. T. (2002). Analyzing the Past to Prepare for the Future: Writing a Literature Review. *MIS Quarterly*, 26(2), xiii–xxiii. https://www.jstor.org/stable/4132319?seq=1#metadata_info_tab_contents
- Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W. W., Salakhutdinov, R., & Manning, C. D. (2018). *HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering*. <https://doi.org/10.48550/arXiv.1809.09600>
- Yu, H., Gan, A., Zhang, K., Tong, S., Liu, Q., & Liu, Z. (2024). *Evaluation of Retrieval-Augmented Generation: A Survey*. <https://doi.org/10.48550/arXiv.2405.07437>
- Zhang, Z [Zihan], Fang, M., & Chen, L. (2024). *RetrievalQA: Assessing Adaptive Retrieval-Augmented Generation for Short-form Open-Domain Question Answering*. <https://doi.org/10.48550/arXiv.2402.16457>
- Zhang, Z [Zihan], Fang, M., Chen, L., Namazi-Rad, M.-R., & Wang, J [Jun]. (2023). *How Do Large Language Models Capture the Ever-changing World Knowledge? A Review of Recent Advances*. <https://doi.org/10.48550/arXiv.2310.07343>