Evaluation of Retrieval-Augmented Generation:
A Survey

Hao Yu[1,2], Aoran Gan[3], Kai Zhang[3], Shiwei Tong[1†], Qi Liu[3], and Zhaofeng Liu[1]

[1] Tencent Company
[2] McGill University
[3] State Key Laboratory of Cognitive Intelligence,
University of Science and Technology of China
hao.yu2@mail.mcgill.ca
gar@mail.ustc.edu.cn
{shiweitong[†],zhaofengliu}@tencent.com
{kkzhang08,qiliuql}@ustc.edu.cn

Abstract. Retrieval-Augmented Generation (RAG) has recently gained traction in natural language processing. Numerous studies and real-world applications are leveraging its ability to enhance generative models through external information retrieval. Evaluating these RAG systems, however, poses unique challenges due to their hybrid structure and reliance on dynamic knowledge sources. To better understand these challenges, we conduct *A Unified Evaluation Process of RAG* (*Auepora*) and aim to provide a comprehensive overview of the evaluation and benchmarks of RAG systems. Specifically, we examine and compare several quantifiable metrics of the Retrieval and Generation components, such as relevance, accuracy, and faithfulness, within the current RAG benchmarks, encompassing the possible output and ground truth pairs. We then analyze the various datasets and metrics, discuss the limitations of current benchmarks, and suggest potential directions to advance the field of RAG benchmarks.

1   Introduction

Retrieval-Augmented Generation (RAG) [34] efficiently enhances the performance of generative language models through integrating information retrieval techniques. It addresses a critical challenge faced by standalone generative language models: the tendency to produce responses that, while plausible, may not be grounded in facts. By retrieving relevant information from external sources, RAG significantly reduces the incidence of hallucinations [23] or factually incorrect outputs, thereby improving the content's reliability and richness. [73] This fusion of retrieval and generation capabilities enables the creation of responses that are not only contextually appropriate but also informed by the most current and accurate information available, making RAG a development in the pursuit of more intelligent and versatile language models [73,64].

[†] Corresponding Author
Paper Homepage: https://github.com/YHPeter/Awesome-RAG-Evaluation

Numerous studies of RAG systems have emerged from various perspectives since the advent of Large Language Models (LLMs) [55,45,59,42,41,69,16]. The RAG system comprises two primary components: *Retrieval* and *Generation*. The retrieval component aims to extract relevant information from various external knowledge sources. It involves two main phases, *indexing* and *searching*. Indexing organizes documents to facilitate efficient retrieval, using either inverted indexes for sparse retrieval or dense vector encoding for dense retrieval [16,12,28]. The searching component utilizes these indexes to fetch relevant documents on the user's query, often incorporating the optional rerankers [4,39,6,52] to refine the ranking of the retrieved documents. The generation component utilizes the retrieved content and question query to formulate coherent and contextually relevant responses with the prompting and inferencing phases. As the "Emerging" ability [59] of LLMs and the breakthrough in aligning human commands [42], LLMs are the best performance choices model for the generation stage. Prompting methods like Chain of Thought (CoT) [60], Tree of Thgouht [65], Rephrase and Respond (RaR) [8] guide better generation results. In the inferencing step, LLMs interpret the prompted input to generate accurate and in-depth responses that align with the query's intent and integrate the extracted information [35,9] without further finetuning, such as fully finetuning [16,1,67,68] or LoRA [21]. Appendix A details the complete RAG structure. Figure 1 illustrates the structure of the RAG systems as mentioned.
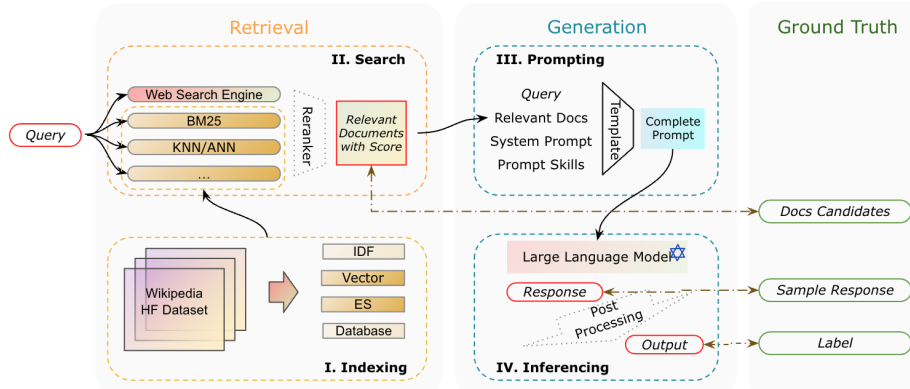

Fig. 1: The structure of the RAG system with retrieval and generation components and corresponding four phrases: indexing, search, prompting and inferencing. The pairs of "Evaluable Outputs" (EOs) and "Ground Truths" (GTs) are highlighted in read frame and green frame, with brown dashed arrows.

The importance of evaluating RAG is increasing in parallel with the advancement of RAG-specific methodologies. On the one hand, RAG is a complex system intricately tied to specific requirements and language models, resulting in various evaluation methods, indicators, and tools, particularly given the black-box LLM generation. Evaluating RAG systems involves specific components and the complexity of the overall system assessment. On the other hand, the complexity of RAG systems is further compounded

by the external dynamic database and the various downstream tasks, such as content creation or open domain question answering [16,70]. These challenges necessitate the development of comprehensive evaluation metrics that can effectively capture the interplay between retrieval accuracy and generative quality [2,7]. To clarify the elements further, we try to address the current gaps in the area, which differs from the prior RAG surveys [74,16,24] that predominantly collected specific RAG methods or data. We have compiled 12 distinct evaluation frameworks, encompassing a range of aspects of the RAG system. Following the procedure of making benchmarks, we analyze through targets, datasets and metrics mentioned in these benchmarks and summarize them into **A Unified Evaluation Process of RAG** (*Auepora*) as three corresponding phases.

For this paper, we contribute in the following aspects:

1. **Challenge of Evaluation**: This is the first work that summarizes and classifies the challenges in evaluating RAG systems through the structure of RAG systems, including three parts retrieval, generation, and the whole system.
2. **Analysis Framework**: In light of the challenges posed by RAG systems, we introduce an analytical framework, referred to as *A Unified Evaluation Process of RAG* (*Auepora*), which aims to elucidate the unique complexities inherent to RAG systems and guide for readers to comprehend the effectiveness of RAG benchmarks across various dimensions
3. **RAG Benchmark Analysis**: With the help of *Auepora*, we comprehensively analyze existing RAG benchmarks, highlighting their strengths and limitations and proposing recommendations for future developments in RAG system evaluation.

2   Challenges in Evaluating RAG Systems

Evaluating hybrid RAG systems entails evaluating retrieval, generation and the RAG system as a whole. These evaluations are multifaceted, requiring careful consideration and analysis. Each of them encompasses specific difficulties that complicate the development of a comprehensive evaluation framework and benchmarks for RAG systems.

Retrieval The retrieval component is critical for fetching relevant information that informs the generation process. One primary challenge is the dynamic and vast nature of potential knowledge bases, ranging from structured databases to the entire web. This vastness requires evaluation metrics that can effectively measure the precision, recall, and relevance of retrieved documents in the context of a given query [52,32]. Moreover, the temporal aspect of information, where the relevance and accuracy of data can change over time, adds another layer of complexity to the evaluation process [6]. Additionally, the diversity of information sources and the possibility of retrieving misleading or low-quality information pose significant challenges in assessing the effectiveness of filtering and selecting the most pertinent information [39]. The traditional evaluation indicators for retrieval, such as Recall and Precision, cannot fully capture the nuances of RAG retrieval systems, necessitating the development of more nuanced and task-specific evaluation metrics [49].

Generation The generation component, powered by LLMs, produces coherent and contextually appropriate responses based on the retrieved content. The challenge here lies in evaluating the faithfulness and accuracy of the generated content to the input data. This involves not only assessing the factual correctness of responses but also their relevance to the original query and the coherence of the generated text [75,49]. The subjective nature of certain tasks, such as creative content generation or open-ended question answering, further complicates the evaluation, as it introduces variability in what constitutes a "correct" or "high-quality" response [48].

RAG System as a Whole Evaluating the whole RAG system introduces additional complexities. The interplay between the retrieval and generation components means that the entire system's performance cannot be fully understood by evaluating each component in isolation [49,14]. The system needs to be assessed on its ability to leverage retrieved information effectively to improve response quality, which involves measuring the added value of the retrieval component to the generative process. Furthermore, practical considerations such as response latency and the ability to handle ambiguous or complex queries are also crucial for evaluating the system's overall effectiveness and usability [39,6].

Conclusion Evaluating the target shift from traditional absolute numeric metrics to multi-source and multi-target generation evaluation, along with the intricate interplay between retrieval and generation components, poses significant challenges. [5,50] Searches in a dynamic database may lead to misleading results or contradict the facts. Diverse and comprehensive datasets that accurately reflect real-world scenarios are crucial. Challenges also arise in the realm of metrics, encompassing generative evaluation criteria for distinct downstream tasks, human preferences, and practical considerations within the RAG system. Most prior benchmarks predominantly tackle one or several aspects of the RAG assessment but lack a comprehensive, holistic analysis.

3   A Unified Evaluation Process of RAG (*Auepora*)

To facilitate a deeper understanding of RAG benchmarks, we introduce *A Unified Evaluation Process of RAG* (*Auepora*), which focuses on three key questions of benchmarks: *What to Evaluate? How to Evaluate? How to Measure?* which correlated to *Target*, *Dataset*, and *Metric* respectively. We aim to provide a clear and accessible way for readers to comprehend the complexities and nuances of RAG benchmarking.

The *Target* module is intended to determine the evaluation direction. The *Dataset* module facilitates the comparison of various data constructions in RAG benchmarks. The final module, *Metrics*, introduces the metrics that correspond to specific targets and datasets used during evaluation. Overall, it is designed to provide a systematic methodology for assessing the effectiveness of RAG systems across various aspects by covering all possible pairs at the beginning between the "Evaluable Outputs" (EOs) and "Ground Truths" (GTs). In the following section, we will explain thoroughly *Auepora* and utilize it for introducing and comparing the RAG benchmarks.
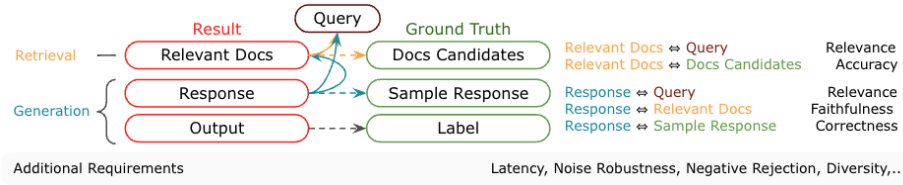
Fig. 2: The *Target* modular of the *Auepora*.

3.1 Evaluation Target (*What to Evaluate?*)

The combination of EOs and GTs in the RAG system can generate all possible targets, which is the fundamental concept of the *Auepora* (as shown in Figure 1). Once identified, these targets can be defined based on a specific pair of EOs or EO with GT, as illustrated in Figure 2, and used to analyze all aspects of current RAG benchmarks.

Retrieval The EOs are the relevant documents for evaluating the retrieval component depending on the query. Then we can construct two pairwise relationships for the retrieval component, which are *Relevant Documents ↔ Query*, *Relevant Documents ↔ Documents Candidates*.

- **Relevance** (*Relevant Documents ↔ Query*) evaluates how well the retrieved documents match the information needed expressed in the query. It measures the precision and specificity of the retrieval process.
- **Accuracy** (*Relevant Documents ↔ Documents Candidates*) assesses how accurate the retrieved documents are in comparison to a set of candidate documents. It is a measure of the system's ability to identify and score relevant documents higher than less relevant or irrelevant ones.

Generation The similar pairwise relations for the generation components are listed below. The EOs are the generated text and phrased structured content. Then we need to compare these EOs with the provided GTs and labels.

- **Relevance** (*Response ↔ Query*) measures how well the generated response aligns with the intent and content of the initial query. It ensures that the response is related to the query topic and meets the query's specific requirements.
- **Faithfulness** (*Response ↔ Relevant Documents*) evaluates if the generated response accurately reflects the information contained within the relevant documents and measures the consistency between generated content and the source documents.
- **Correctness** (*Response ↔ Sample Response*) Similar to the accuracy in the retrieval component, this measures the accuracy of the generated response against a sample response, which serves as a ground truth. It checks if the response is correct in terms of factual information and appropriate in the context of the query.

The targets of Retrieval and Generation components are introduced. Table 1 lists the relative work on improving and evaluating RAG and its benchmarks cut off in June

Table 1: The evaluating targets and corresponding metrics across various frameworks for evaluating RAG systems. The presentation distinguishes between the core areas of Retrieval and Generation considered in the evaluation. The different aspects of the evaluation are set as different colours in the table: Relevance, Accuracy of Retrieval and Faithfulness, Correctness and Relevance of Generation. The consideration of the *Additional Requirements* beyond the retrieval and generation component is also collected. Noted that quite a few of the works employed multiple methods or evaluated multiple aspects simultaneously.

Category	Framework	Time	Raw Targets	Retrieval	Generation
Tool	TruEra RAG Triad [54]	2023.10	Context Relevance Answer Relevance *Groundedness*	LLM as a Judge	LLM as a Judge
Tool	LangChain Bench. [32]	2023.11	Accuracy Faithfulness *Execution Time* *Embed. CosDistance*	Accuracy	LLM as a Judge
Tool	Databricks Eval [33]	2023.12	Correctness *Readability* *Comprehensiveness*	-	LLM as a Judge
Benchmark	RAGAs [14]	2023.09	Context Relevance Answer Relevance Faithfulness	LLM as a Judge	LLM Gen + CosSim LLM as a Judge
Benchmark	RECALL [38]	2023.11	Response Quality *Robustness*	-	BLEU, ROUGE-L
Benchmark	ARES [49]	2023.11	Context Relevance Answer Faithfulness Answer Relevance	LLM + Classifier	LLM + Classifier LLM + Classifier
Benchmark	RGB [6]	2023.12	Information Integration *Noise Robustness* *Negative Rejection* *Counterfactual Robustness*	-	Accuracy
Benchmark	MultiHop-RAG [52]	2024.01	Retrieval Quality Response Correctness	MAP, MRR, Hit@K	LLM as a Judge
Benchmark	CRUD-RAG [39]	2024.02	*CREATE*, READ *UPDATE, DELETE*	-	ROUGE, BLEU RAGQuestEval
Benchmark	MedRAG [61]	2024.02	Accuracy	-	Accuracy
Benchmark	FeB4RAG [57]	2024.02	Consistency Correctness *Clarity* *Coverage*	-	Human Evaluation Human Evaluation
Benchmark	CDQA [62]	2024.03	Accuracy	-	F1
Benchmark	DomainRAG [58]	2024.06	Correctness Faithfulness *Noise Robustness* *Structural Output*	-	F1, Exact-Match Rouge-L LLM as a Judge
Benchmark	ReEval [66]	2024.06	Hallucination	-	F1, Exact-Match LLM as a Judge Human Evaluation
Research	FiD-Light [20]	2023.07	*Latency*	-	-
Research	Diversity Reranker [4]	2023.08	*Diversity*	Cosine Distance	-

2024. Table 1 portrays this information, where each evaluation criterion is represented by a different colour. For example, FeB4RAG [57], the fourth from the last, has posited four standards based on [17] that comprise Consistency, Correctness, Clarity, and Coverage. Correctness is equivalent to accuracy in retrieval, and Consistency is tantamount to faithfulness in the generation component. While accuracy in retrieval gauges the correctness of the retrieved information, we posit that Coverage pertains to the coverage rate and is more associated with diversity. Therefore, we consider *Coverage* to be linked with diversity and an additional requirement in our proposed evaluation framework, which will be introduced subsequently. The remaining standard, *Clarity*, is also classified as an additional requirement in our proposed framework. The other tools and benchmarks are processed similarly.

Tools and benchmarks offer varying degrees of flexibility in evaluating datasets for RAG systems. Tools, which specify only evaluation targets, provide a versatile framework capable of constructing complete RAG applications and evaluation pipelines, as seen in works like [54,32,33]. Benchmarks, on the other hand, focus on different aspects of RAG evaluation with specific emphasis on either retrieval outputs or generation targets. For instance, RAGAs [14] and ARES [49] assess the relevance of retrieval documents, while RGB and MultiHop-RAG [6,52] prioritize accuracy, necessitating comparison with GTs. The [66] focuses on the Hallucination, which is a combination of faithfulness and correctness. All benchmarks consider generation targets due to their critical role in RAG systems, though their focus areas vary.

Additional Requirement In addition to evaluating the two primary components outlined, a portion of the works also addressed some additional requirements of RAG (Black and *Italics* targets in Table 2). The requirements are as follows:

- **Latency** [20,32] measures how quickly the system can find information and respond, crucial for user experience.
- **Diversity** [4,32] checks if the system retrieves a variety of relevant documents and generates diverse responses.
- **Noise Robustness** [6] assesses how well the system handles irrelevant information without affecting response quality.
- **Negative Rejection** [6] gauges the system's ability to refrain from providing a response when the available information is insufficient.
- **Counterfactual Robustness** [6] evaluates the system's capacity to identify and disregard incorrect information, even when alerted about potential misinformation.
- **More**: For more human preferences considerations, there can be more additional requirements, such as readability [57,33], toxicity, perplexity [33], etc.

For the exception, CRUD-RAG [39] introduces a comprehensive benchmark addressing the broader spectrum of RAG applications beyond question-answering, categorized into Create, Read, Update, and Delete scenarios. This benchmark evaluates RAG systems across diverse tasks, including text continuation, question answering, hallucination modification, and multi-document summarization. It offers insights for optimizing RAG technology across different scenarios. DomainRAG [58] identifies six complex abilities for RAG systems: conversational, structural information, faithfulness,

denoising, time-sensitive problem solving, and multi-doc understanding. ReEval [66] specifically targets hallucination evaluation by employing a cost-effective LLM-based framework that utilizes prompt chaining to create dynamic test cases.

Table 2: The evaluation datasets used for each benchmark. The dataset without citation was constructed by the benchmark itself.

Benchmark	Dataset
RAGAs [14]	WikiEval
RECALL [38]	EventKG [19], UJ [22]
ARES [49]	NQ [29], Hotpot [63], FEVER [53], WoW [11], MultiRC [10], ReCoRD [71]
RGB [6]	Generated (Source: News)
MultiHop-RAG [52]	Generated (Source: News)
CRUD-RAG [39]	Generated (Source: News) UHGEval [36]
MedRAG [61]	MIRAGE
FeB4RAG [57]	FeB4RAG, BEIR [26]
CDQA [62]	Generation (Source: News), Labeller
DomainRAG [58]	Generation (Source: College Admission Information)
ReEval [66]	RealTimeQA[27], NQ [15,29])

3.2   Evaluation Dataset (*How to evaluate?*)

In Table 2, distinct benchmarks employ varying strategies for dataset construction, ranging from leveraging existing resources to generating entirely new data tailored for specific evaluation aspects. Several benchmarks draw upon the part of KILT (Knowledge Intensive Language Tasks) benchmark [44] (Natural Questions [29], HotpotQA [63], and FEVER [53]) and other established datasets such as SuperGLUE [56] (MultiRC [10], and ReCoRD [71]) [49]. However, the drawback of using such datasets can't solve the challenges in dynamic real-world scenarios. A similar situation can be observed in WikiEval, from Wikipedia pages post 2022, constructed by RAGAs [14].

The advent of powerful LLMs has revolutionized the process of dataset construction. With the ability to design queries and ground truths for specific evaluation targets using these frameworks, authors can now create datasets in the desired format with ease. Benchmarks like RGB, MultiHop-RAG, CRUD-RAG, and CDQA [6,52,39,62] have taken this approach further by building their own datasets using online news articles to test RAG systems' ability to handle real-world information beyond the training data of LM frameworks. Most recently, DomainRAG [58] combines various types of QA datasets with single-doc, multi-doc, single-round, and multi-round. These datasets are generated from the yearly changed information from the college website for admission and enrollment, which forces the LLMs to use the provided and updated information.

In summary, the creation and selection of datasets are crucial for evaluating RAG systems. Datasets tailored for specific metrics or tasks improve evaluation accuracy and guide the development of adaptable RAG systems for real-world information needs.

3.3   Evaluation Metric (*How to quantify?*)

Navigating the intricate terrain of evaluating RAG systems necessitates a nuanced understanding of the metrics that can precisely quantify the evaluation targets. However, creating evaluative criteria that align with human preferences and address practical considerations is challenging. Each component within the RAG systems requires a tailored evaluative approach that reflects its distinct functionalities and objectives.

Retrieval Metrics   Various targets can be evaluated with various metrics that correspond to the given datasets. This section will introduce several commonly used metrics for retrieval and generation targets. The metrics for additional requirements can also be found in these commonly used metrics. The more specifically designed metrics can be explored in the original paper via Table 1 as a reference.
For the retrieval evaluation, the focus is on metrics that can accurately capture the relevance, accuracy, diversity, and robustness of the information retrieved in response to queries. These metrics must not only reflect the system's precision in fetching pertinent information but also its resilience in navigating the dynamic, vast, and sometimes misleading landscape of available data. The deployment of metrics like *Misleading Rate*, *Mistake Reappearance Rate*, and *Error Detection Rate* within the [38] benchmark underscores a heightened awareness of RAG systems' inherent intricacies. The integration of *MAP@K*, *MRR@K*, and *Tokenization with F1* into benchmarks like [52,62] mirrors a deepening comprehension of traditional retrieval's multifaceted evaluation. While the [17] also emphasizes that this ranking-based evaluation methodology is not unsuitable for the RAG system, and should have more RAG-specific retrieval evaluation metrics. These metrics not only capture the precision and recall of retrieval systems but also account for the diversity and relevance of retrieved documents, aligning with the complex and dynamic nature of information needs in RAG systems. The introduction of LLMs as evaluative judges, as seen in [14], further underscores the adaptability and versatility of retrieval evaluation, offering a comprehensive and context-aware approach to assessing retrieval quality.

Non-Rank Based Metrics often assess binary outcomes—whether an item is relevant or not—without considering the position of the item in a ranked list. Notice, that the following formula is just one format of these metrics, the definition of each metric may vary by the different evaluating tasks.

  - **Accuracy** is the proportion of true results (both true positives and true negatives) among the total number of cases examined.
  - **Precision** is the fraction of relevant instances among the retrieved instances,

$$\text{Precision} = \frac{TP}{TP + FP}$$

where TP represents true positives and FP represents false positives.

- **Recall** at k ($Recall@k$) is the fraction of relevant instances that have been retrieved over the total amount of relevant cases, considering only the top k results.

$$Recall@k = \frac{|RD \cap Top_{kd}|}{|RD|}$$

where RD is the relevant documents, and Top_{kd} is the top-k retrieved documents.

Rank-Based Metrics evaluate the order in which relevant items are presented, with higher importance placed on the positioning of relevant items at the ranking list.

- **Mean Reciprocal Rank (MRR)** is the average of the reciprocal ranks of the first correct answer for a set of queries.

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i}$$

where $|Q|$ is the number of queries and $rank_i$ is the rank position of the first relevant document for the i-th query.

- **Mean Average Precision (MAP)** is the mean of the average precision scores for each query.

$$MAP = \frac{1}{|Q|} \sum_{q=1}^{|Q|} \frac{\sum_{k=1}^{n}(P(k) \times rel(k))}{|\text{relevant documents}_q|}$$

where $P(k)$ is the precision at cutoff k in the list, $rel(k)$ is an indicator function equaling 1 if the item at rank k is a relevant document, 0 otherwise, and n is the number of retrieved documents.

Generation Metrics In the realm of generation, evaluation transcends the mere accuracy of generated responses, venturing into the quality of text in terms of coherence, relevance, fluency, and alignment with human judgment. This necessitates metrics that can assess the nuanced aspects of language production, including factual correctness, readability, and user satisfaction with the generated content. The traditional metrics like *BLEU*, *ROUGE*, and *F1 Score* continue to play a crucial role, emphasizing the significance of precision and recall in determining response quality. Yet, the advent of metrics such as *Misleading Rate*, *Mistake Reappearance Rate*, and *Error Detection Rate* highlights an evolving understanding of RAG systems' distinct challenges [38].

The evaluation done by humans is still a very significant standard to compare the performance of generation models with one another or with the ground truth. The approach of employing LLMs as evaluative judges [75] is a versatile and automatic method for quality assessment, catering to instances where traditional ground truths may be elusive [14]. This methodology benefits from employing prediction-powered inference (PPI) and context relevance scoring, offering a nuanced lens through which LLM output can be assessed. [49] The strategic use of detailed prompt templates ensures a guided assessment aligned with human preferences, effectively standardizing evaluations across various content dimensions [1]. This shift towards leveraging LLMs

as arbiters mark a significant progression towards automated and context-responsive evaluation frameworks, enriching the evaluation landscape with minimal reliance on reference comparisons.

- **ROUGE** Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [37] is a set of metrics designed to evaluate the quality of summaries by comparing them to human-generated reference summaries. ROUGE can be indicative of the content overlap between the generated text and the reference text. The variants of ROUGEs measure the overlap of n-grams (ROUGE-N, ROUGGE-W), word subsequences (ROUGE-L, ROUGGE-S), and word pairs between the system-generated summary and the reference summaries.
- **BLEU** Bilingual Evaluation Understudy (BLEU) [43] is a metric for evaluating the quality of machine-translated text against one or more reference translations. BLEU calculates the precision of n-grams in the generated text compared to the reference text and then applies a brevity penalty to discourage overly short translations. BLEU has limitations, such as not accounting for the fluency or grammaticality of the generated text.
- **BertScore** BertScore [72] leverages the contextual embedding from pre-trained transformers like BERT to evaluate the semantic similarity between generated text and reference text. BertScore computes token-level similarity using contextual embedding and produces precision, recall, and F1 scores. Unlike n-gram-based metrics, BertScore captures the meaning of words in context, making it more robust to paraphrasing and more sensitive to semantic equivalence.
- **LLM as a Judge** Using "LLM as a Judge" for evaluating generated text is a more recent approach. [75] In this method, LLMs are used to score the generated text based on criteria such as coherence, relevance, and fluency. The LLM can be optionally finetuned on human judgments to predict the quality of unseen text or used to generate evaluations in a zero-shot or few-shot setting. This approach leverages the LLM's understanding of language and context to provide a more nuanced text quality assessment. For instance, [1] illustrates how providing LLM judges with detailed scoring guidelines, such as a scale from 1 to 5, can standardize the evaluation process. This methodology encompasses critical aspects of content assessment, including coherence, relevance, fluency, coverage, diversity, and detail - both in the context of answer evaluation and query formulation.

Additional Requirements   These additional requirements, such as latency, diversity, noise robustness, negative rejection, and counterfactual robustness, are used to ensure the practical applicability of RAG systems in real-world scenarios aligned with human preference. This section delves into the metrics used for evaluating these additional requirements, highlighting their significance in the comprehensive assessment of RAG systems.

Latency measures the time taken by the RAG system to finish the response of one query. It is a critical factor for user experience, especially in interactive applications such as chatbots or search engines [20]. *Single Query Latency*: The mean time is taken to process a single query, including both retrieval and generating phases.

Diversity evaluates the variety and breadth of information retrieved and generated by the RAG system. It ensures that the system can provide a wide range of perspectives and avoid redundancy in responses [4]. *Cosine Similarity / Cosine Distance*: The cosine similarity/distance calculates embeddings of retrieved documents or generated responses. [30] Lower cosine similarity scores indicate higher diversity, suggesting that the system can retrieve or generate a broader spectrum of information.

Noise Robustness measures the RAG system's ability to handle irrelevant or misleading information without compromising the quality of the response [38]. The metrics *Misleading Rate* and *Mistake Reappearance Rate* are described in [38], providing detailed descriptions tailored to the specific dataset and experimental setup. [58]

Negative Rejection evaluates the system's capability to withhold responses when the available information is insufficient or too ambiguous to provide an accurate answer [6]. *Rejection Rate*: The rate at which the system refrains from generating a response.

Counterfactual Robustness Counterfactual robustness assesses the system's ability to identify and disregard incorrect or counterfactual information within the retrieved documents [39]. *Error Detection Rate*: The ratio of counterfactual statements detected in retrieved information.

4 Discussion

For RAG systems, traditional Question Answering (QA) datasets and metrics remain a common format for interaction. [14,49,38,6,61,62,58,66] While these provide a basic verification of RAG's capabilities, it becomes challenging to distinguish the impact of retrieval components when faced with strong Language Models (LLMs) capable of excelling in QA benchmarks. To comprehensively evaluate the performance of entire RAG systems, there is a need for diverse and RAG-specific benchmarks. Several papers offer guidance on improving QA format benchmarks, including variations in question types: from simple Wikipedia filling questions to multi-hop [52], multi-document questions [66] and single-round to multi-round dialogue [39,58]. For answers, aspects such as structural output [58], content moderation [6,54], and hallucination [66] can be considered when evaluating relevance, faithfulness, and correctness. In addition to these, RAG systems require additional requirements such as robustness to noisy documents, language expression, latency, and result diversity. [32,33,38,6,39,57,58,20,4] Furthermore, research is needed on performance changes involving intermediate outputs and retrieved documents, as well as the relationship and analysis between retrieval metrics and final generation outputs.

Regarding *datasets*, creating a universal dataset was challenging due to the target-specific nature of different RAG benchmarks. Tailored datasets [14,38,49,39,57] are necessary for a thorough evaluation, but this approach increases the effort and resources required. Moreover, the diversity of datasets, from news articles to structured databases [66], reflects the adaptability required of RAG systems but also poses a barrier to streamlined evaluation. Recently, with the cutting-edge performance of LLMs, complex data processing and automatic QA pair generation can be automated to achieve daily or finer-grained time resolution, preventing LLMs from cheating and evaluating the robustness of RAG systems in rapidly changing data. [6,52,39,62,58,66]

When it comes to *metrics*, the use of LLMs as automatic evaluative judges signifies a burgeoning trend, promising versatility and depth in generative outputs with reasoning on a large scale compared to human evaluation. However, using "LLMs as a Judge" [75] for responses presents challenges in aligning with human judgment, establishing effective grading scales, and applying consistent evaluation across varied use cases. Determining correctness, clarity, and richness can differ between automated and human assessments. Moreover, the effectiveness of example-based scoring can vary, and there's no universally applicable grading scale and prompting text, complicating the standardization of "LLM as a Judge". [33]

In addition to the challenges mentioned above, it is important to consider the resource-intensive nature [76] of using Large Language Models (LLMs) for data generation and validation. RAG benchmarks must balance the need for thorough evaluation with the practical constraints of limited computational resources. As such, it is desirable to develop evaluation methodologies that can effectively assess RAG systems using smaller amounts of data while maintaining the validity and reliability of the results.

5 Conclusion

This survey systematically explores the complexities of evaluating RAG systems, highlighting the challenges in assessing their performance. Through the proposed *A Unified Evaluation Process of RAG*, we outline a structured approach to analyzing RAG evaluations, focusing on targets, datasets and measures. Our analysis emphasizes the need for targeted benchmarks that reflect the dynamic interplay between retrieval accuracy and generative quality and practical considerations for real-world applications. By identifying gaps in current methodologies and suggesting future research directions, we aim to contribute to more effective, and user-aligned benchmarks of RAG systems.

References

1. Balaguer, A., Benara, V., Cunha, R.L.d.F., Filho, R.d.M.E., Hendry, T., Holstein, D., Marsman, J., Mecklenburg, N., Malvar, S., Nunes, L.O., Padilha, R., Sharp, M., Silva, B., Sharma, S., Aski, V., Chandra, R.: RAG vs Fine-tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture. Tech. rep. (Jan 2024), `http://arxiv.org/abs/2401.08406`, arXiv:2401.08406 [cs] type: article

2. Barnett, S., Kurniawan, S., Thudumu, S., Brannelly, Z., Abdelrazek, M.: Seven failure points when engineering a retrieval augmented generation system (Jan 2024). `https://doi.org/10.48550/ARXIV.2401.05856`

3. Besta, M., Blach, N., Kubicek, A., Gerstenberger, R., Podstawski, M., Gianinazzi, L., Gajda, J., Lehmann, T., Niewiadomski, H., Nyczyk, P., Hoefler, T.: Graph of thoughts: Solving elaborate problems with large language models. Proceedings of the AAAI Conference on Artificial Intelligence 2024 (AAAI'24) (Aug 2023). `https://doi.org/10.48550/ARXIV.2308.09687`

4. Blagojevic, V.: Enhancing RAG Pipelines in Haystack: Introducing DiversityRanker and LostInTheMiddleRanker (Aug 2023), `https://towardsdatascience.com/enhancing-rag-pipelines-in-haystack-45f14e2bc9f5`

5. Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., et al.: A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology **15**(3), 1–45 (2024)

6. Chen, J., Lin, H., Han, X., Sun, L.: Benchmarking large language models in retrieval-augmented generation (Sep 2023). `https://doi.org/10.48550/ARXIV.2309.01431`

7. Cuconasu, F., Trappolini, G., Siciliano, F., Filice, S., Campagnano, C., Maarek, Y., Tonellotto, N., Silvestri, F.: The power of noise: Redefining retrieval for rag systems (Jan 2024). `https://doi.org/10.48550/ARXIV.2401.14887`

8. Deng, Y., Zhang, W., Chen, Z., Gu, Q.: Rephrase and respond: Let large language models ask better questions for themselves (Nov 2023). `https://doi.org/10.48550/ARXIV.2311.04205`

9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). `https://doi.org/10.18653/v1/N19-1423`, `https://aclanthology.org/N19-1423`

10. DeYoung, J., Jain, S., Rajani, N.F., Lehman, E., Xiong, C., Socher, R., Wallace, B.C.: Eraser: A benchmark to evaluate rationalized nlp models

11. Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., Weston, J.: Wizard of Wikipedia: Knowledge-powered conversational agents. In: Proceedings of the International Conference on Learning Representations (ICLR) (2019)

12. Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.E., Lomeli, M., Hosseini, L., Jégou, H.: The faiss library (2024)

13. DuckDuckGo: DuckDuckGo — Privacy, simplified. (2024), `https://duckduckgo.com//home`

14. Es, S., James, J., Espinosa-Anke, L., Schockaert, S.: Ragas: Automated evaluation of retrieval augmented generation (Sep 2023). `https://doi.org/10.48550/ARXIV.2309.15217`

15. Fisch, A., Talmor, A., Jia, R., Seo, M., Choi, E., Chen, D.: MRQA 2019 shared task: Evaluating generalization in reading comprehension. In: Fisch, A., Talmor, A., Jia, R., Seo, M., Choi, E., Chen, D. (eds.) Proceedings of the 2nd Workshop on Machine Reading for Question Answering. pp. 1–13. Association for Computational Linguistics, Hong Kong, China (Nov 2019). `https://doi.org/10.18653/v1/D19-5801`, `https://aclanthology.org/D19-5801`

16. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Guo, Q., Wang, M., Wang, H.: Retrieval-Augmented Generation for Large Language Models: A Survey. Tech. rep. (Jan 2024), `http://arxiv.org/abs/2312.10997`, arXiv:2312.10997 [cs] type: article

17. Gienapp, L., Scells, H., Deckers, N., Bevendorff, J., Wang, S., Kiesel, J., Syed, S., Fröbe, M., Zuccon, G., Stein, B., Hagen, M., Potthast, M.: Evaluating Generative Ad Hoc Information Retrieval. Tech. rep. (Nov 2023), `http://arxiv.org/abs/2311.04694`, arXiv:2311.04694 [cs] type: article

18. Google: Programmable Search Engine | Google for Developers (2024), `https://developers.google.com/custom-search`

19. Gottschalk, S., Demidova, E.: Eventkg: A multilingual event-centric temporal knowledge graph (Apr 2018). `https://doi.org/10.48550/ARXIV.1804.04526`

20. Hofstätter, S., Chen, J., Raman, K., Zamani, H.: FiD-Light: Efficient and Effective Retrieval-Augmented Text Generation. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1437–1447. SIGIR '23, Association for Computing Machinery, New York, NY, USA (Jul 2023). `https://doi.org/10.1145/3539618.3591687`, `https://doi.org/10.1145/3539618.3591687`

21. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-Rank Adaptation of Large Language Models. Tech. rep. (Oct 2021). `https://doi.org/10.48550/arXiv.2106.09685`, `http://arxiv.org/abs/2106.09685`, arXiv:2106.09685 [cs] type: article

22. Huang, J., Shao, H., Chang, K.C.C., Xiong, J., Hwu, W.m.: Understanding jargon: Combining extraction and generation for definition modeling. In: Proceedings of EMNLP (2022)

23. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., Liu, T.: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions (Nov 2023). `https://doi.org/10.48550/ARXIV.2311.05232`

24. Huang, Y., Huang, J.: A survey on retrieval-augmented text generation for large language models (Apr 2024). `https://doi.org/10.48550/ARXIV.2404.10981`

25. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with GPUs. IEEE Transactions on Big Data **7**(3), 535–547 (2019)

26. Kamalloo, E., Thakur, N., Lassance, C., Ma, X., Yang, J.H., Lin, J.: Resources for brewing beir: Reproducible reference models and an official leaderboard (2023)

27. Kasai, J., Sakaguchi, K., Takahashi, Y., Bras, R.L., Asai, A., Yu, X., Radev, D., Smith, N.A., Choi, Y., Inui, K.: Realtime qa: What's the answer right now? (Jul 2022). `https://doi.org/10.48550/ARXIV.2207.13332`, `https://arxiv.org/abs/2207.13332`

28. Khattab, O., Zaharia, M.: Colbert: Efficient and effective passage search via contextualized late interaction over bert (Apr 2020). `https://doi.org/10.48550/ARXIV.2004.12832`

29. 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.: Natural questions: A benchmark for question

answering research. Transactions of the Association for Computational Linguistics **7**, 453–466 (2019). `https://doi.org/10.1162/tacl_a_00276`, `https://doi.org/10.1162/tacl_a_00276`

30. Lahitani, A.R., Permanasari, A.E., Setiawan, N.A.: Cosine similarity to determine similarity measure: Study case in online essay assessment. In: 2016 4th International Conference on Cyber and IT Service Management. pp. 1–6 (2016). `https://doi.org/10.1109/CITSM.2016.7577578`

31. Lanchantin, J., Toshniwal, S., Weston, J., Szlam, A., Sukhbaatar, S.: Learning to reason and memorize with self-notes (May 2023). `https://doi.org/10.48550/ARXIV.2305.00833`

32. LangChain: Evaluating rag architectures on benchmark tasks (Nov 2023), `https://langchain-ai.github.io/langchain-benchmarks/notebooks/retrieval/langchain_docs_qa.html`

33. Leng, Q., Uhlenhuth, K., Polyzotis, A.: Best Practices for LLM Evaluation of RAG Applications (Dec 2023), `https://www.databricks.com/blog/LLM-auto-eval-best-practices-RAG`

34. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. pp. 9459–9474. NIPS'20, Curran Associates Inc., Red Hook, NY, USA (Dec 2020)

35. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Tech. rep. (Apr 2021), `http://arxiv.org/abs/2005.11401`, arXiv:2005.11401 [cs] type: article

36. Liang, X., Song, S., Niu, S., Li, Z., Xiong, F., Tang, B., Wy, Z., He, D., Cheng, P., Wang, Z., Deng, H.: Uhgeval: Benchmarking the hallucination of chinese large language models via unconstrained generation. arXiv preprint arXiv:2311.15296 (2023)

37. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (Jul 2004), `https://aclanthology.org/W04-1013`

38. Liu, Y., Huang, L., Li, S., Chen, S., Zhou, H., Meng, F., Zhou, J., Sun, X.: Recall: A benchmark for llms robustness against external counterfactual knowledge (Nov 2023). `https://doi.org/10.48550/ARXIV.2311.08147`

39. Lyu, Y., Li, Z., Niu, S., Xiong, F., Tang, B., Wang, W., Wu, H., Liu, H., Xu, T., Chen, E., Luo, Y., Cheng, P., Deng, H., Wang, Z., Lu, Z.: Crud-rag: A comprehensive chinese benchmark for retrieval-augmented generation of large language models (Jan 2024). `https://doi.org/10.48550/ARXIV.2401.17043`

40. Microsoft: Web Search API | Microsoft Bing, `https://www.microsoft.com/en-us/bing/apis/bing-web-search-api`

41. 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., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H.W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S.P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T.,

Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S.S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Kamali, A., Kanitscheider, I., Keskar, N.S., Khan, T., Kilpatrick, L., Kim, J.W., Kim, C., Kim, Y., Kirchner, J.H., Kiros, J., Knight, M., Kokotajlo, D., Kondraciuk, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C.M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S.M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., Peres, F.d.A.B., Petrov, M., Pinto, H.P.d.O., Michael, Pokorny, Pokrass, M., Pong, V.H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F.P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M.B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J.F.C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J.J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., Zoph, B.: GPT-4 Technical Report (Mar 2023). https://doi.org/10.48550/ARXIV.2303.08774

42. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback. Tech. rep. (Mar 2022). https://doi.org/10.48550/arXiv.2203.02155, http://arxiv.org/abs/2203.02155, arXiv:2203.02155 [cs] type: article

43. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Isabelle, P., Charniak, E., Lin, D. (eds.) Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (Jul 2002). https://doi.org/10.3115/1073083.1073135, https://aclanthology.org/P02-1040

44. Petroni, F., Piktus, A., Fan, A., Lewis, P., Yazdani, M., De Cao, N., Thorne, J., Jernite, Y., Karpukhin, V., Maillard, J., Plachouras, V., Rocktäschel, T., Riedel, S.: 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. pp. 2523–2544. Association for Computational Linguistics, Online (Jun 2021). https://doi.org/10.18653/v1/2021.naacl-main.200, https://aclanthology.org/2021.naacl-main.200

45. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)

46. Ramos, J., et al.: Using tf-idf to determine word relevance in document queries. In: Proceedings of the first instructional conference on machine learning. vol. 242, pp. 29–48. Citeseer

(2003)

47. Robertson, S., Zaragoza, H., et al.: The probabilistic relevance framework: Bm25 and beyond. Foundations and Trends® in Information Retrieval **3**(4), 333–389 (2009)
48. Rosset, C., Chung, H.L., Qin, G., Chau, E.C., Feng, Z., Awadallah, A., Neville, J., Rao, N.: Researchy questions: A dataset of multi-perspective, decompositional questions for llm web agents (Feb 2024). `https://doi.org/10.48550/ARXIV.2402.17896`
49. Saad-Falcon, J., Khattab, O., Potts, C., Zaharia, M.: Ares: An automated evaluation framework for retrieval-augmented generation systems (Nov 2023). `https://doi.org/10.48550/ARXIV.2311.09476`
50. Sai, A.B., Mohankumar, A.K., Khapra, M.M.: A survey of evaluation metrics used for nlg systems. ACM Computing Surveys (CSUR) **55**(2), 1–39 (2022)
51. Shahabi, C., Kolahdouzan, M.R., Sharifzadeh, M.: A road network embedding technique for k-nearest neighbor search in moving object databases. In: Proceedings of the 10th ACM international symposium on advances in geographic information systems. pp. 94–100 (2002)
52. Tang, Y., Yang, Y.: Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries (Jan 2024). `https://doi.org/10.48550/ARXIV.2401.15391`
53. Thorne, J., Vlachos, A., Christodoulopoulos, C., Mittal, A.: FEVER: a large-scale dataset for fact extraction and VERification. In: NAACL-HLT (2018)
54. TruLens: TruLens (2023), `https://www.trulens.org/trulens_eval/getting_started/quickstarts/quickstart/`
55. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (Jun 2017). `https://doi.org/10.48550/ARXIV.1706.03762`
56. Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: SuperGLUE: A stickier benchmark for general-purpose language understanding systems. arXiv preprint 1905.00537 (2019)
57. Wang, S., Khramtsova, E., Zhuang, S., Zuccon, G.: Feb4rag: Evaluating federated search in the context of retrieval augmented generation (Feb 2024). `https://doi.org/10.48550/ARXIV.2402.11891`
58. Wang, S., Liu, J., Song, S., Cheng, J., Fu, Y., Guo, P., Fang, K., Zhu, Y., Dou, Z.: Domainrag: A chinese benchmark for evaluating domain-specific retrieval-augmented generation (Jun 2024). `https://doi.org/10.48550/ARXIV.2406.05654`
59. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E.H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., Fedus, W.: Emergent abilities of large language models (Jun 2022). `https://doi.org/10.48550/ARXIV.2206.07682`
60. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models (Jan 2022). `https://doi.org/10.48550/ARXIV.2201.11903`
61. Xiong, G., Jin, Q., Lu, Z., Zhang, A.: Benchmarking retrieval-augmented generation for medicine (Feb 2024). `https://doi.org/10.48550/ARXIV.2402.13178`
62. Xu, Z., Li, Y., Ding, R., Wang, X., Chen, B., Jiang, Y., Zheng, H.T., Lu, W., Xie, P., Huang, F.: Let llms take on the latest challenges! a chinese dynamic question answering benchmark (Feb 2024). `https://doi.org/10.48550/ARXIV.2402.19248`
63. Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W.W., Salakhutdinov, R., Manning, C.D.: HotpotQA: A dataset for diverse, explainable multi-hop question answering. In: Conference on Empirical Methods in Natural Language Processing (EMNLP) (2018)
64. Yao, J.Y., Ning, K.P., Liu, Z.H., Ning, M.N., Yuan, L.: Llm lies: Hallucinations are not bugs, but features as adversarial examples. arXiv preprint arXiv:2310.01469 (2023)
65. Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T.L., Cao, Y., Narasimhan, K.: Tree of Thoughts: Deliberate problem solving with large language models (2023)

66. Yu, X., Cheng, H., Liu, X., Roth, D., Gao, J.: ReEval: Automatic hallucination evaluation for retrieval-augmented large language models via transferable adversarial attacks. In: Duh, K., Gomez, H., Bethard, S. (eds.) Findings of the Association for Computational Linguistics: NAACL 2024. pp. 1333–1351. Association for Computational Linguistics, Mexico City, Mexico (Jun 2024), `https://aclanthology.org/2024.findings-naacl.85`

67. Zhang, K., Liu, Q., Qian, H., Xiang, B., Cui, Q., Zhou, J., Chen, E.: Eatn: An efficient adaptive transfer network for aspect-level sentiment analysis. IEEE Transactions on Knowledge and Data Engineering **35**(1), 377–389 (2021)

68. Zhang, K., Zhang, H., Liu, Q., Zhao, H., Zhu, H., Chen, E.: Interactive attention transfer network for cross-domain sentiment classification. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 5773–5780 (2019)

69. Zhang, K., Zhang, K., Zhang, M., Zhao, H., Liu, Q., Wu, W., Chen, E.: Incorporating dynamic semantics into pre-trained language model for aspect-based sentiment analysis. arXiv preprint arXiv:2203.16369 (2022)

70. Zhang, Q., Chen, S., Xu, D., Cao, Q., Chen, X., Cohn, T., Fang, M.: A Survey for Efficient Open Domain Question Answering. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 14447–14465. Association for Computational Linguistics, Toronto, Canada (Jul 2023). `https://doi.org/10.18653/v1/2023.acl-long.808`, `https://aclanthology.org/2023.acl-long.808`

71. Zhang, S., Liu, X., Liu, J., Gao, J., Duh, K., Van Durme, B.: Record: Bridging the gap between human and machine commonsense reading comprehension (Oct 2018). `https://doi.org/10.48550/ARXIV.1810.12885`

72. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BERTScore: Evaluating Text Generation with BERT. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net (2020), `https://openreview.net/forum?id=SkeHuCVFDr`

73. Zhang, Y., Khalifa, M., Logeswaran, L., Lee, M., Lee, H., Wang, L.: Merging Generated and Retrieved Knowledge for Open-Domain QA. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 4710–4728. Association for Computational Linguistics, Singapore (Dec 2023). `https://doi.org/10.18653/v1/2023.emnlp-main.286`, `https://aclanthology.org/2023.emnlp-main.286`

74. Zhao, P., Zhang, H., Yu, Q., Wang, Z., Geng, Y., Fu, F., Yang, L., Zhang, W., Cui, B.: Retrieval-augmented generation for ai-generated content: A survey (Feb 2024). `https://doi.org/10.48550/ARXIV.2402.19473`

75. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J.E., Stoica, I.: Judging llm-as-a-judge with mt-bench and chatbot arena (Jun 2023). `https://doi.org/10.48550/ARXIV.2306.05685`

76. Zhou, Y., Lin, X., Zhang, X., Wang, M., Jiang, G., Lu, H., Wu, Y., Zhang, K., Yang, Z., Wang, K., Sui, Y., Jia, F., Tang, Z., Zhao, Y., Zhang, H., Yang, T., Chen, W., Mao, Y., Li, Y., Bao, D., Li, Y., Liao, H., Liu, T., Liu, J., Guo, J., Zhao, X., WEI, Y., Qian, H., Liu, Q., Wang, X., Kin, W., Chan, Li, C., Li, Y., Yang, S., Yan, J., Mou, C., Han, S., Jin, W., Zhang, G., Zeng, X.: On the opportunities of green computing: A survey (Nov 2023)

77. Zhu, F., Lei, W., Wang, C., Zheng, J., Poria, S., Chua, T.S.: Retrieving and Reading: A Comprehensive Survey on Open-domain Question Answering. Tech. rep. (May 2021), `http://arxiv.org/abs/2101.00774`, arXiv:2101.00774 [cs] type: article

A   Structure of RAG System

A.1   Retrieval Component

The retrieval component of RAG systems in Figure 1 can be categorized into three types: sparse retrieval, dense retrieval [77], and web search engine. The standard for evaluation is the output of *relevant documents* with numerical scores or rankings.

Before the introduction of neural networks, *sparse retrievals* are widely used for retrieving relative text content. Methods like TF-IDF [46] and BM25 [47] rely on keyword matching and word frequency but may miss semantically relevant documents without keyword overlap.

By leveraging deep learning models such as BERT [9], *dense retrieval* can capture the semantic meaning of texts, which allows them to find relevant documents even when keyword overlap is minimal. This is crucial for complex queries that require a contextual understanding to retrieve accurate information. With advanced fusion structure for queries and documents [28] and the more efficient implementation of K-Nearest Neighbors (KNN) [51], Approximate Nearest Neighbor (ANN) [12,25] search techniques, dense retrieval methods have become practical for large-scale use.

Web search engine employs the complex online search engine to provide relevant documents, such as Google Search [18], Bing Search [40], DuckDuckGo [13]. RAG systems can traverse the web's extensive information, potentially returning a more diverse and semantically relevant set of documents via the API of the search provider. The black box of the search engine and the expense of large-scale search are not affordable sometimes.

It is observed that dense retrieval techniques, particularly those leveraging embeddings, stand out as the preferred choice within the RAG ecosystem. These methods are frequently employed in tandem with sparse retrieval strategies, creating a hybrid approach that balances precision and breadth in information retrieval. Moreover, the adoption of sophisticated web search engines for benchmark assessment underscores their growing significance in enhancing the robustness and comprehensiveness of evaluations.

Indexing   The indexing component processes and indexes document collections, such as HuggingFace datasets or Wikipedia pages. Chunking before indexing can improve retrieval by limiting similarity scores to individual chunks, as semantic embedding is less accurate for long articles, and desired content is often brief [32]. Index creation is designed for fast and efficient search. For example, the inverted index for sparse retrieval and the ANN index for dense retrieval.

Sparse Retrieval involves calculating IDF for each term and storing values in a database for quick look-up and scoring when queried.

Dense Retrieval encodes documents into dense vectors using a pre-trained language model like BERT. These vectors are then indexed using an Approximate Nearest Neighbor (ANN) search technique, like graph-based Hierarchical Navigable Small World (HNSW) or Inverted File Index (IVF) [12]. This process allows for the efficient retrieval of "closed" items by given predefined distance metrics.

Search   This step is responsible for retrieving relevant documents based on a given query. Queries are submitted using the respective API to retrieve relevant documents for web search engine retrieval. For local resources, the query component is responsible for formatting the query in the format required by different sparse or dense retrieval methods. Then, the query is submitted to the retrieval system, which returns a set of relevant documents along with their scores.

In both local and web-based scenarios, an optional reranker can be employed to refine the ranking of retrieved documents further. The reranker usually comprises a more complex and larger model that considers additional features of the documents and the given query. These additional features often include the semantic relationship between the query and the document content, document importance or popularity, and other custom measures specific to the information need at hand.

A.2   Generation Component

The evaluable output for the generation component is the *response* of LLMs and the *structured or formatted output* from the phrased response.

Prompting   The generation process critically hinges on prompting, where a query, retrieval outcomes, and instructions converge into a single input for the language model. Research showcases various strategic prompting tactics such as the Chain of Thought (CoT) [60], Tree of Thought (ToT) [3], and Self-Note [31], each significantly shaping the model's output. These methods, especially the step-by-step approach, are pivotal in augmenting LLMs for intricate tasks.

Prompting innovations have introduced methods like Rephrase and Respond (RaR) [8], enhancing LLMs by refining queries within prompts for better comprehension and response. This technique has proven to boost performance across diverse tasks. The latest RAG benchmarks [61,62] in the specific domains start to evaluate the robustness of various prompting engineering skills, including CoT, RaR, etc.

Inference   The final input string prepared in the prompting step is then passed on to the LLMs as input, which generates the output. The inference stage is where the LLM operates on the input derived from the retrieval and the prompting stages in the pipeline to generate the final output. This is usually the answer to the initial query and is used for downstream tasks.

Depending on the specifics of the task or expected output structure, a post-processing step may be implemented here to format the generated output suitably or extract specific information from the response. For example, the classification problems (multi-choice questions) or if the task requires the extraction of specific information from the generated text, this step could involve additional named entity recognition or parsing operations.