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.