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.