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]