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