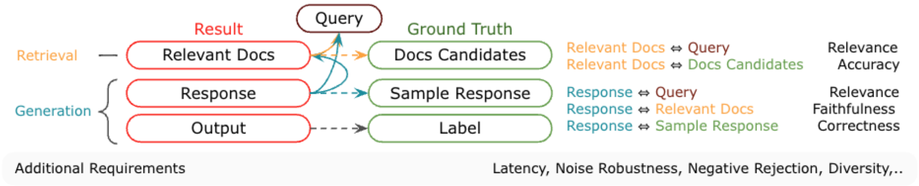

Fig. 2: The *Target* modular of the *Auepora*.

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, Exacct-Match LLM as a Judge Human Evaluation
Research	FiD-Light [20]	2023.07	*Latency*	-	-
Research	Diversity Reranker [4]	2023.08	*Diversity*	Cosine Distance	-