Evaluation of RAG: A Survey

Hao Yu[1,2], Aoran Gan[3], Kai Zhang[3], Shiwei Tong[1†], Qi Liu[3], and Zhaofeng Liu[1]

[1] Tencent Company
[2] McGill University
[3] State Key Laboratory of Cognitive Intelligence,
University of Science and Technology of China

Background - Structure of RAG

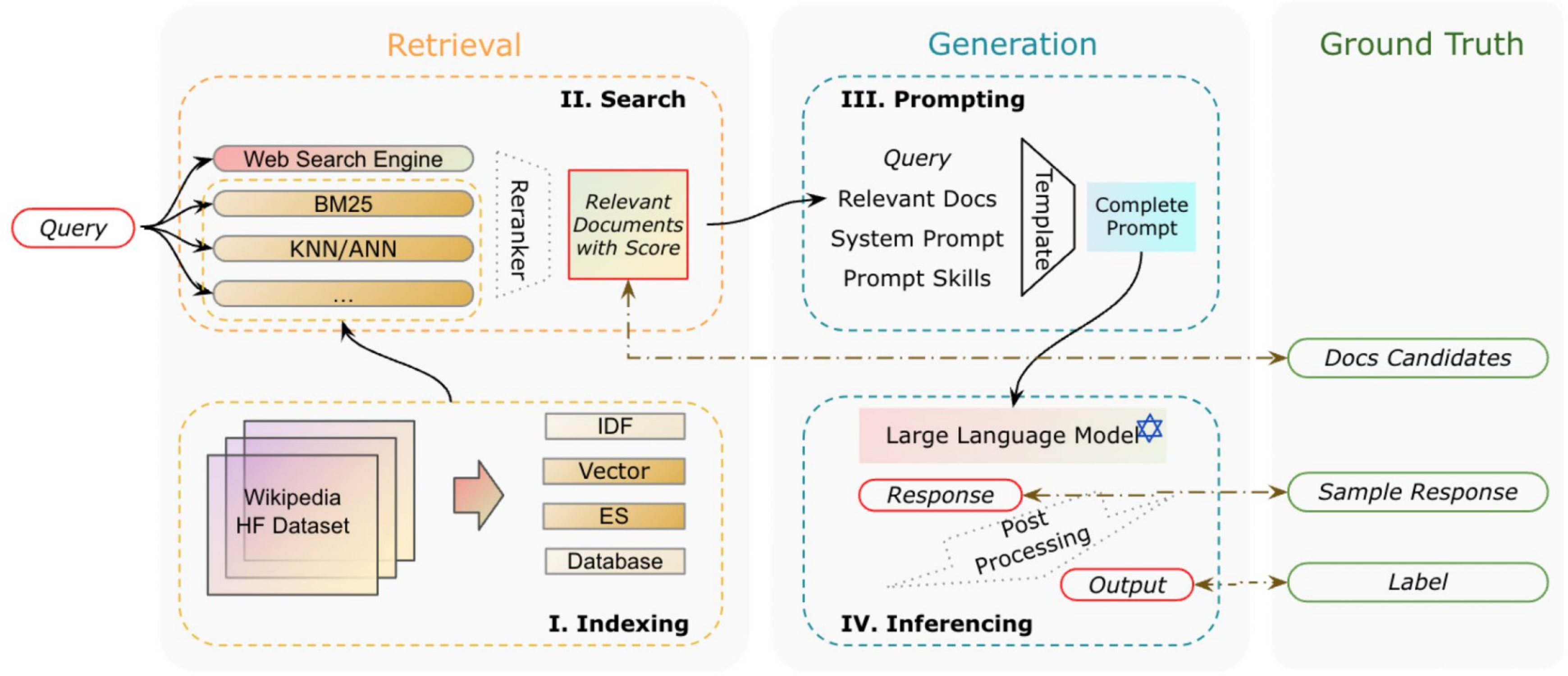

Fig. 1: The structure of the RAG system with retrieval and generation components and corresponding four phrases: indexing, search, prompting and inferencing. The pairs of "Evaluable Outputs" (EOs) and "Ground Truths" (GTs) are highlighted in read frame and green frame, with brown dashed arrows.

Background - Evaluation Challenge

Retrieval:

- Dynamic Knowledge Base
- Outdated over Time
- Misleading or Low-Quality Information/Source
- Metrics for retrieved content for RAG

Generation:

- Various Text Generation Evaluation
- Faithfulness of Retrieval Content

RAG System as a Whole:

- Interplay between Retrieval and Generation
- Practical Aspects: Latency, Robustness, ...

A Unified Evaluation Process of RAG (*Auepora*)

Auepora.I - Target (*What to Evaluate?*)

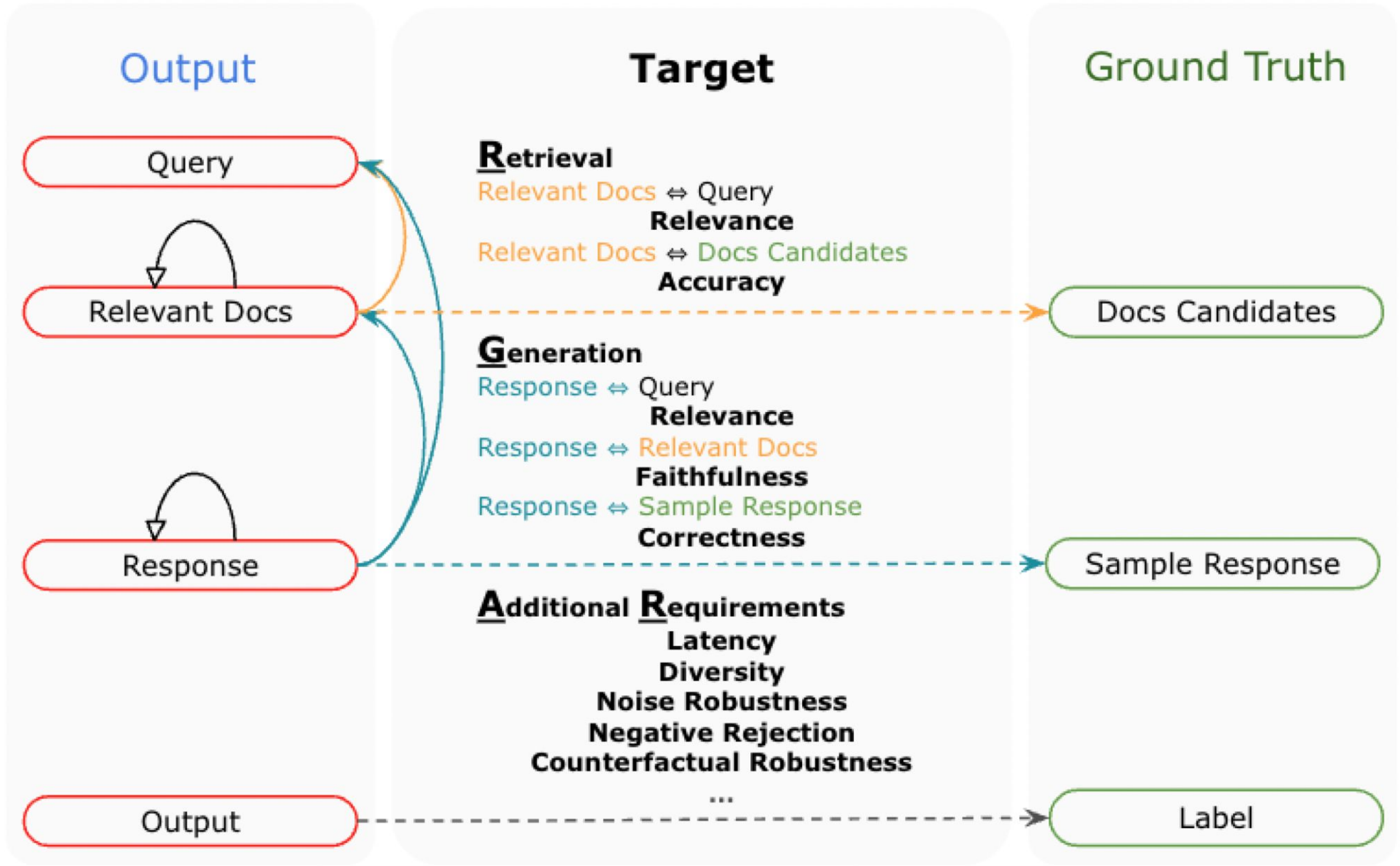

Fig. 2: The *Target* modular of the *Auepora*.

Auepora.II - Dataset (*How to Evaluate?*)

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])

arXiv
https://arxiv.org/abs/2405.07437

Github
https://github.com/YHPeter/Awesome-RAG-Evaluation

Auepora.III - Metrics (*How to Quantify?*)

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	-