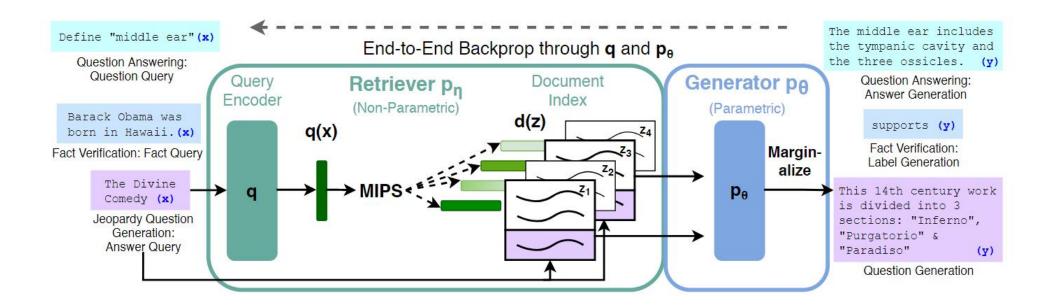
Evaluating Retrieval Quality in Retrieval-Augmented Generation

2024.9.24

唐明昊

Background

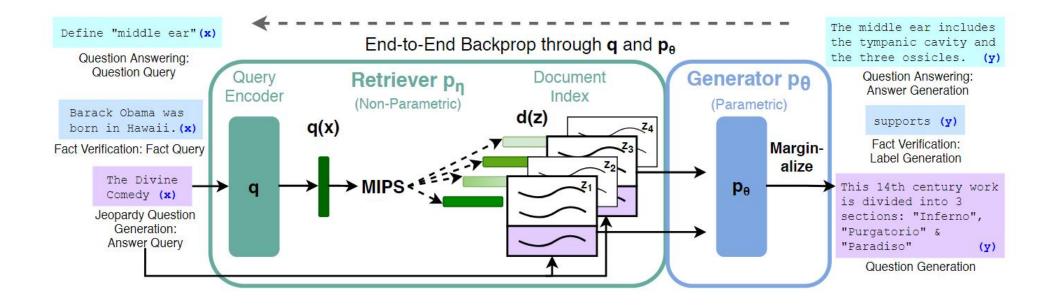
retrieval-augmented generation (RAG) has emerged as a prominent approach in natural language processing, combining the strengths of retrieval and generation models



Background

how to evaluate RAG systems: end-to-end assessment

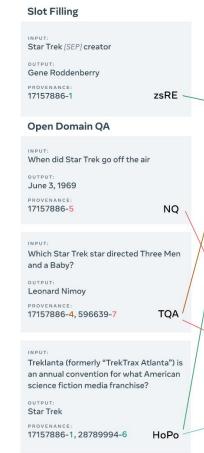
- lack of transparency
- high resource consumption
- complex system pipeline
- hard to optimize



Motivation

evaluating retrievers in RAG:

- human annotation
 costly, often impractical, human preferences
- downstream ground truth sometimes impractical, partial
- LLM annotation
 LLM preference, computational cost





Knowledge source: 5.9 Million Wikipedia pages

Star Trek 17157886

Star Trek is an American media franchise based on the science fiction

television series created by Gene Roddenberry. [...] It followed the interstellar adventures of Captain James T. Kirk (William Shatner) and his crew aboard the starship USS "Enterprise", a space exploration vessel built by the United

Federation of Planets in the 23rd century. The "Star Trek" canon includes "The Original Series", an animated series, five spin-off television series, the film franchise, and further adaptations in several media.

[...] The original 1966–69 series featured William Shatner as Captain James T.

Kirk, Leonard Nimoy⁴ as Spock, DeForest Kelley as Dr. Leonard "Bones" McCoy, James Doohan as Montgomery "Scotty" Scott, Nichelle Nichols as Uhura, George Takei as Hikaru Sulu, and Walter Koenig as Pavel Chekov. During the series' first run, it earned several nominations for the Hugo Award for Best Dramatic Presentation, and won twice. [...]

NBC canceled the show after three seasons; the last original episode aired on June 3, 1969^{5} . [...]

Three Men and a Baby 596639

Three Men and a Baby is a 1987 American comedy film directed by <u>Leonard Nimoy</u> and starring Tom Selleck, Steve Guttenberg, Ted Danson and Nancy Travis. [...]

Treklanta 28789994

Treklanta is an annual "Star Trek" convention based in Atlanta, Georgia that places special emphasis on fan-based events, activities, programming and productions. 6 [...]

Dialogue

INDICE

I am a big fan of Star Trek, the American franchise created by Gene Roddenberry. I don't know much about it. When did the first episode air?
It debuted in 1996 and aired for 3

seasons on NBC.
What is the plot of the show?

OUTPUT:

William Shatner plays the role of Captain Kirk. He did a great job.

17157886-2

WoW

Fact Checking

NPUT:

Star Trek had spin-off television series.

Supports

PROVENANCE: 17157886-3

FEV

Entity Linking

INPUT

[...] Currently the site offers five movie collections ranging from \$149 for 10 [START_ENT] Star Trek [END_ENT] films to \$1,125 for the eclectic Movie Lovers' Collection of 75 movies. [...]

Star Trek

PROVENANCE

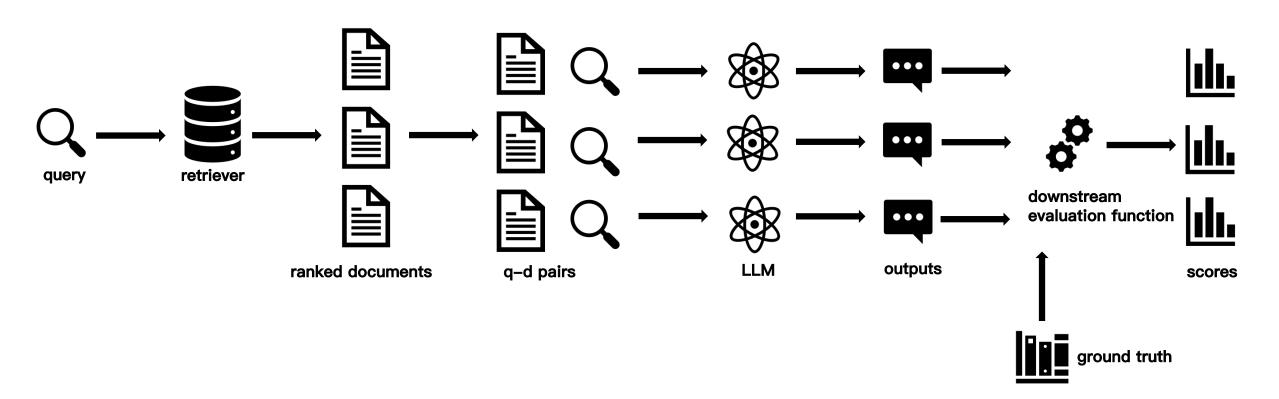
17157886

CnWn

eRAG

utilize the LLM in RAG system itself as the arbiter for generating labels to evaluate the retrieval model

$$\mathcal{G}_q[d] = \mathcal{E}_{\mathcal{M}}(\mathcal{M}(q, \{d\}), y) : \forall d \in \mathbf{R}_k$$



eRAG

utilize the LLM in RAG system itself as the arbiter for generating labels to evaluate the retrieval model

$$\mathcal{G}_q[d] = \mathcal{E}_{\mathcal{M}}(\mathcal{M}(q, \{d\}), y) : \forall d \in \mathbf{R}_k$$

given the ranked documents list and relevance score for each document, use an evaluation metric to get a specific score:

- Precision (P)
- Recall (R)
- Mean Average Precision (MAP)
- Mean Reciprocal Rank (MRR)
- Normalized Discounted Cumulative Gain (NDCG)
- Hit Ratio

Experiments Configuration

- Datasets: KILT (NQ, TriviaQA, HotpotQA, FEVER, WoW)
- Downstream metrics: EM for NQ, TriviaQA, HotpotQA. Accuracy for FEVER. F1 for WoW
- Retriever: BM25, Contriever
- LLM: T5-small with Fusion-in-Decoder
- LLM annotator: Mistral 7B

Findings

Table 1: The correlation between each evaluation approach and the downstream performance of the LLM. T5-small with FiD with 50 retrieved documents is used. We do not report correlation for the Answers method for FEVER and WOW datasets because the answers to queries do not exist in the document since FEVER is a classification dataset and WoW is long-text generation. For the WoW dataset, we only report correlation on Precision and Hit Ratio because other metrics do not support non-integer relevance labels. Tau is Kendall's tau and rho is Spearman's rho.

Relevance Annotation	Metric	BM25								Contriever											
		NQ		TriviaQA		HotpotQA		FEVER		WoW		NQ		TriviaQA		HotpotQA		FEVER		WoW	
		tau	rho	tau	rho	tau	rho	tau	rho	tau	rho	tau	rho	tau	rho	tau	rho	tau	rho	tau	rho
Containing the Answer	MAP	0.349	0.417	0.298	0.364	0.359	0.423	7 -	-		-	0.303	0.366	0.265	0.325	0.379	0.429	-	-	-	-
	MRR	0.361	0.417	0.313	0.340	0.398	0.449	72	120	12	20	0.301	0.353	0.257	0.292	0.384	0.430	-	-	141	82
	NDCG	0.357	0.427	0.298	0.365	0.370	0.435	-	-	-	120	0.313	0.378	0.270	0.331	0.385	0.437	-		-	-
	P	0.353	0.411	0.276	0.333	0.396	0.454	3. 5	(53)	-	1700 E	0.346	0.403	0.283	0.340	0.406	0.449	-	. =	-	- -
	R	0.325	0.325	0.232	0.232	0.375	0.375	-		S=	-	0.319	0.319	0.215	0.215	0.401	0.401	-	-	-	-
	Hit Ratio	0.325	0.325	0.232	0.232	0.375	0.375	7		19	-	0.319	0.319	0.215	0.215	0.401	0.401	-	=	4	72
KILT Provenance	MAP	0.181	0.218	0.142	0.172	0.007	0.009	0.026	0.032	0.015	0.021	0.161	0.196	0.113	0.137	0.128	0.155	0.045	0.056	0.055	0.080
	MRR	0.177	0.205	0.151	0.175	0.074	0.080	0.036	0.040	0.013	0.017	0.152	0.173	0.120	0.136	0.151	0.169	0.045	0.049	0.059	0.081
	NDCG	0.179	0.216	0.142	0.172	0.021	0.026	0.029	0.036	0.013	0.019	0.159	0.193	0.115	0.140	0.134	0.162	0.045	0.056	0.056	0.081
	P	0.163	0.192	0.140	0.165	0.139	0.164	0.043	0.051	0.011	0.015	0.131	0.157	0.108	0.130	0.181	0.215	0.033	0.040	0.045	0.064
	R	0.216	0.216	0.187	0.187	0.113	0.113	0.050	0.050	0.019	0.023	0.157	0.157	0.135	0.135	0.163	0.163	0.038	0.038	0.056	0.068
	Hit Ratio	0.216	0.216	0.187	0.187	0.113	0.113	0.050	0.050	0.019	0.023	0.157	0.157	0.135	0.135	0.163	0.163	0.038	0.038	0.056	0.068
Relevance Annotation with LLM (Mistral 7B)	MAP	0.045	0.055	0.176	0.216	0.034	0.042	0.018	0.022	-0.005	-0.008	0.032	0.039	0.174	0.213	0.051	0.063	0.021	0.026	-0.002	-0.003
	MRR	0.060	0.062	0.189	0.196	0.001	0.001	-0.021	-0.022	-0.008	-0.011	0.048	0.050	0.143	0.151	0.034	0.038	-0.007	-0.007	0.004	0.005
	NDCG	0.049	0.060	0.178	0.218	0.032	0.039	0.018	0.022	-0.006	-0.009	0.036	0.044	0.175	0.214	0.049	0.060	0.022	0.028	0.000	0.000
	P	0.028	0.034	0.137	0.166	-0.004	-0.006	0.021	0.025	-0.005	-0.008	0.002	0.003	0.138	0.167	0.010	0.013	0.014	0.017	-0.006	-0.010
	R	0.014	0.014	0.032	0.032	-0.016	-0.016	0.019	0.019	0.003	0.003	0.000	0.000	0.039	0.039	-0.042	-0.042	-0.017	-0.017	0.017	0.021
	Hit Ratio	0.014	0.014	0.032	0.032	-0.016	-0.016	0.019	0.019	0.003	0.003	0.000	0.000	0.039	0.039	-0.042	-0.042	-0.017	-0.017	0.017	0.021
eRAG Annotations	MAP	0.492	0.575	0.474	0.578	0.610	0.694	0.386	0.463	-	-	0.467	0.544	0.427	0.519	0.634	0.705	0.399	0.479	-	18
	MRR	0.503	0.577	0.486	0.553	0.629	0.695	0.592	0.611	80	27	0.466	0.537	0.424	0.495	0.639	0.698	0.481	0.504	(2)	
	NDCG	0.505	0.590	0.486	0.592	0.612	0.697	0.404	0.484	7	-	0.481	0.560	0.440	0.536	0.635	0.705	0.403	0.484		-
	P^a	0.529	0.598	0.484	0.577	0.594	0.663	0.329	0.391	0.504	0.669	0.522	0.586	0.482	0.571	0.633	0.695	0.378	0.449	0.540	0.712
	R	0.519	0.519	0.426	0.426	0.619	0.619	0.301	0.301	124	-	0.488	0.488	0.408	0.408	0.631	0.631	0.299	0.299	-	1-
	Hit Ratio ^b	0.519	0.519	0.426	0.426	0.619	0.619	0.301	0.301	0.390	0.532	0.488	0.488	0.408	0.408	0.631	0.631	0.299	0.299	0.414	0.561

 $^{^{}a}$ For non-integer relevance labels, precision is equal to average of the relevance labels.

^b For non-integer relevance labels, hit ratio is equal to maximum of the relevance labels.

Findings

How do different retrieval evaluation methods in RAG perform as the **size of retrieval results** increases?

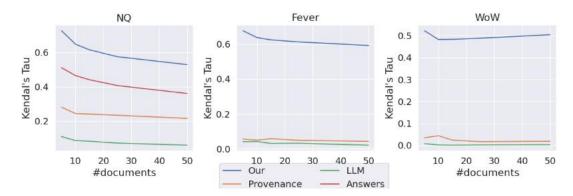


Figure 1: The correlation between evaluation approaches and the LLM's downstream performance varying number of retrieved documents by BM25. T5-small with FiD is used. The metric with the highest correlation in Table 1 is used.

How does eRAG correlate with the downstream RAG performance as the size of large language models increases

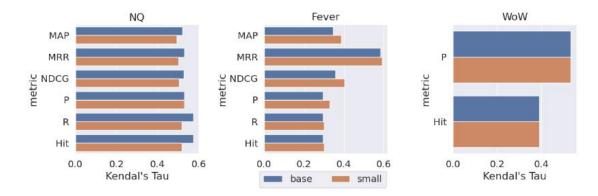


Figure 2: The correlation between eRAG and the downstream performance of different LLM sizes. In this experiment, T5-small (60M parameters) and T5-base (220M parameters) with FiD are used. The documents are retrieved using BM25.

Findings

How does different retrieval-augmentation approaches affect the correlation between eRAG and the downstream RAG performance?

How much more **efficient** is eRAG compared to the end-toend evaluation?

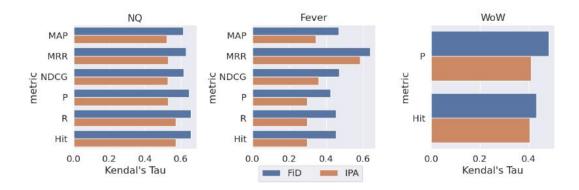


Figure 3: The correlation between eRAG and the downstream performance of FiD and IPA LLMs. T5-small with 10 documents retrieved by BM25 is used. The number of documents is chosen based on the limitations of the input size in IPA.

Table 2: The runtime and memory consumption of eRAG in comparison with end-to-end evaluation. T5-small with FiD, consuming 50 documents is used.

Dataset	Runtim	e (GPU)	Memory Consumption (GPU)						
Dataset	E2E	eRAG	E2E	eRAG-Query	eRAG-Document				
NQ	918 sec	351 sec	75.0 GB	4.9 GB	1.5 GB				
TriviaQA	1819 sec	686 sec	46.2 GB	5.4 GB	1.5 GB				
HotpotQA	1844 sec	712 sec	52.4 GB	5.5 GB	1.5 GB				
FEVER	3395 sec	1044 sec	66.5 GB	4.1 GB	1.5 GB				
WoW	912 sec	740 sec	47.9 GB	6.5 GB	1.5 GB				