The Effectiveness of Retrieval Metrics in Evaluating the Results of Retrieval-Augmented Generation in the Question-Answering Domain

Keisuke Kamahori, Khyati Morparia, Vidya Srinivas, Yue Wu

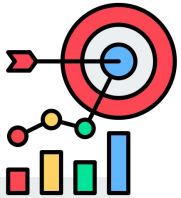
Introduction

Large Language Models (LLMs) show impressive capabilities but have limitations (e.g., outdated knowledge, hallucinations).
While retrieval systems have shown performance enhancements, it still remains a challenge to effectively and accurately evaluate the performance of these models as a human would.

Key Challenges:

- ➢ Evaluating retrieval systems effectively in question-answering scenarios.
- ➢ Syntax-based metrics (e.g. F1-Score) often fail to capture the true accuracy of generated answers.

Goals



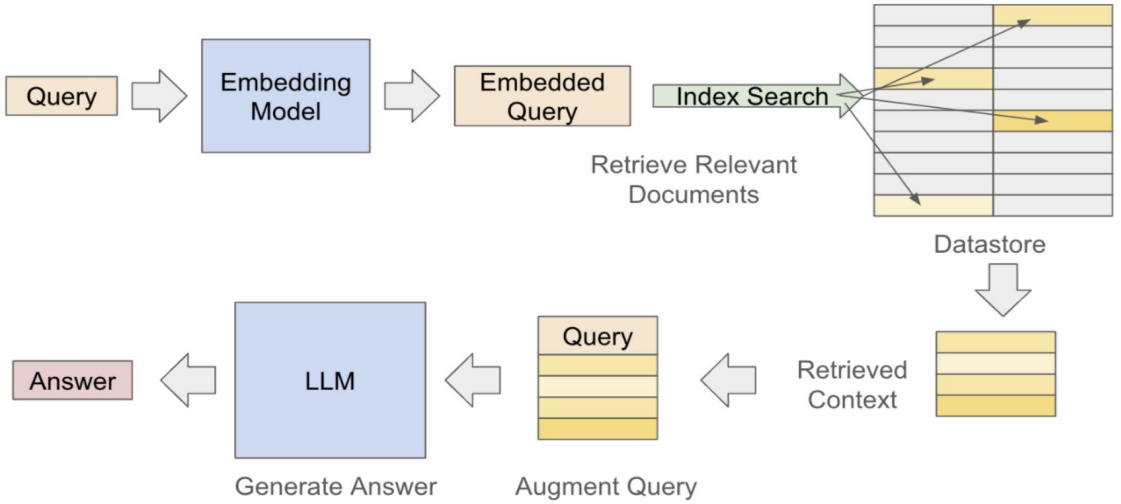
- Identify scenarios where syntax-based metrics fail to capture the correctness of the answer provided by a retrieval system
- Assess metric performance across different question types
- Compare performance of retrieval when given relevant context v.s. not
- Evaluate the deviation of the estimates of performance provided by the metrics v.s those calculated from human annotated samples

Implementation

Baseline Model:

Used Meta-Llama-3-8B-Instruct from Hugging Face.

This model is developed by Meta as a member of Meta Llama 3 family of large language models (LLMs), a collection of pretrained and instruction tuned generative text models in 8 and 70B sizes.



Dataset

We used the dev-distractor partition from the **HotpotQA** dataset

- Each question-answer datapoint is represented as a dictionary, whose keys include
  - _id: unique identifier for each datapoint
  - answer: a string of ground truth answer
  - question: question string
  - context: list of paragraphs which may be helpful for answering the question
  - supporting facts: list of sentences that support the answer of the question

Data Annotation

We evaluated on 1250 of these data points, and used GPT-3.5 to annotate the questions into 6 types:

- Fine-grained (FG): question contains additional information about the entity being asked
- Coarse-grained (CG):questions asks about an entity without additional information
- Yes/No (YN): answer is “yes”, “no”, or one word
- Multiple choice (MC): choose from multiple options
- Time Frame (TF): asks for a date or a range of dates or times
- Comparisons (CMP): question is a comparison that requires context about both choices

Experiments

On Baseline Model Llama3-8B:

- Compare when only given the question (LLM) v.s. given both question and context (LLM+Gold, we simply refer to as Gold)
- Compare results using the following three prompts under Gold configuration
- Evaluate on each of the 6 question types with LLM v.s. Gold
- Manually annotate 100 data points, compare human annotated results with LLM and Gold

Prompt 1 - “You are an assistant for question-answering tasks. Use the following pieces of retrieved context to answer the question. If you don’t know the answer, just say that you don’t know. Answer the query with a succinct response within 10 tokens."

Prompt 2 - “Only answer every question with 10 words, does not need to be full sentences. Use the following pieces of retrieved context and your existing knowledge to answer the question with 10 words max. If the context does not help, ignore it."

Prompt 3 - “You are a qa test machine, you need to answer the [Question] from given the [Context], you only need to come out the correct answer without any other words."

Evaluation

Preprocessing Prediction and Ground Truth Strings:

- Normalize to avoid false negatives (e.g. remove punctuation, all lower case letters, white space fixing)

Metrics:

- Exact Match (EM), F1-Score, Precision, Recall

Baseline Evaluation

Name	EM	F1-Score	Precision	Recall
LLM	0.1912	0.2698	0.2964	0.2619
Gold	0.3672	0.4857	0.5206	0.4808

Insight:

- Gold significantly outperforms LLM, confirming that additional context improves the performance of the purely parametric model

Evaluation on three different prompts

Name	EM	F1-Score	Precision	Recall
Prompt 1	0.0100	0.1560	0.1042	0.4705
Prompt 2	0.0100	0.2212	0.1508	0.5445
Prompt 3	0.3672	0.4857	0.5206	0.4808

Insight:

- Concise prompts yield better performance, with the most effective prompt (Prompt 3) achieving the highest scores due to how syntax-based metrics are calculated
- More verbose prompts lead to longer and often incorrect responses with hallucinations, reducing overall performance

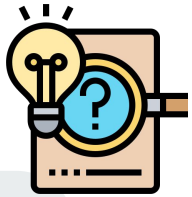
Evaluation by question annotation

Name	EM	F1-Score	Precision	Recall
Fine-grained (LLM)	0.1513	0.2399	0.2676	0.2313
Fine-grained (Gold)	0.3395	0.4773	0.5159	0.5159
Coarse-grained (LLM)	0.1920	0.2685	0.2950	0.2597
Coarse-grained (Gold)	0.3696	0.4858	0.5197	0.4818
Yes/no (LLM)	0.1804	0.2578	0.2826	0.2501
Yes/no (Gold)	0.3618	0.4864	0.5236	0.4793
Multiple choice (LLM)	0.6000	0.6000	0.6000	0.6000
Multiple choice (Gold)	0.6000	0.7333	0.7000	0.8000
Time frame (LLM)	0.1446	0.2278	0.2545	0.2196
Time frame (Gold)	0.3363	0.4693	0.5095	0.4614
Comparisons (LLM)	0.1896	0.2682	0.2948	0.2603
Comparisons (Gold)	0.3688	0.4871	0.5220	0.4822

Insight

- All question types benefit from context retrieval, but the improvements are consistent across types, contrary to the hypothesis that open-ended questions would benefit more.
- The improvement for multiple choice is comparatively lower, but the sample size is too small (smallest)

Human Annotation and Case Study

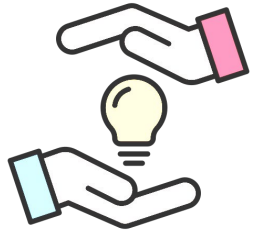


To find out why Gold retrieval doesn't show variations in improvement across question types, we manually annotated 100 samples. Based on case studies we find

- Performance varies greatly depending on the completeness and relevance of the context.
- Full context leads to higher accuracy (46% increase in accuracy) compared to partial context (33% increase).
- Syntax-based metrics are inclined to fail when questions are badly formatted (incorrect grammar, repeated words.

Name	Accuracy
Gold correct on questions with full context	0.7674
LLM correct on questions with full context	0.3023
Gold correct on questions with partial context	0.4912
LLM correct on questions with partial context	0.1579

Conclusion & Future Work



- Syntax-Based Metrics Limitations:
  a. Fall short in assessing verbose answers, synonyms, or semantically similar responses.
  b. Do not adequately reflect variations across different question types, necessitating more nuanced evaluation methods.
- Retrieval Performance:
  a. Significantly better when the provided context is good and comprehensive.
  b. Syntax-based metrics underestimate retrieval performance and fail to account for the quality and correctness of retrieved passages.
- Future Research:
  a. Developing robust, accurate metrics for auto-evaluation that align closely with human annotators remains a crucial research area