

VERA: Validation and Evaluation of Retrieval-Augmented systems

Tianyu Ding
Amazon
Arlington, VA, USA
tianyd@amazon.com

Adi Banerjee
Amazon
Boston, MA, USA
adibaner@amazon.com

Yunhong Li
Amazon
Seattle, WA, USA
yunhonl@amazon.com

Laurent Mombaerts
Amazon
Luxembourg, Luxembourg
lmomb@amazon.lu

Tarik Borogovac
Amazon
Boston, MA, USA
tarikbo@amazon.com

Juan Pablo De la Cruz
Weinstein
Amazon
Seattle, WA, USA
jcruam@amazon.com

ABSTRACT

The increasing use of Retrieval-Augmented Generation (RAG) systems in various applications necessitates stringent protocols to ensure RAG systems' accuracy, safety, and alignment with user intentions. In this paper, we introduce VERA (Validation and Evaluation of Retrieval-Augmented Systems), a framework designed to enhance the transparency and reliability of outputs from large language models (LLMs) that utilize retrieved information. VERA improves the way we evaluate RAG systems in two important ways: (1) it introduces a cross-encoder based mechanism that encompasses a set of multidimensional metrics into a single comprehensive ranking score, addressing the challenge of prioritizing individual metrics, and (2) it employs Bootstrap statistics on LLM-based metrics across the document repository to establish confidence bounds, ensuring the repository's topical coverage and improving the overall reliability of retrieval systems. Through several use cases, we demonstrate how VERA can strengthen decision-making processes and trust in AI applications. Our findings not only contribute to the theoretical understanding of LLM-based RAG evaluation metric but also promote the practical implementation of responsible AI systems, marking a significant advancement in the development of reliable and transparent generative AI technologies.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence; Natural Language Processing.**

KEYWORDS

Large Language Models, Retrieval Augmented System, Evaluation

ACM Reference Format:

Tianyu Ding, Adi Banerjee, Yunhong Li, Laurent Mombaerts, Tarik Borogovac, and Juan Pablo De la Cruz Weinstein. 2024. VERA: Validation and Evaluation of Retrieval-Augmented systems. In *Proceedings of August 25–26,*

2024 (GenAI Evaluation KDD 2024). ACM, New York, NY, USA, 13 pages.
<https://doi.org/XXXXXXX.XXXXXXX>

1 INTRODUCTION

The integration of Retrieval-Augmented Generation (RAG) systems with Large Language Models (LLMs) has significantly advanced the field of natural language processing, particularly enhancing capabilities in areas such as open-domain question answering, fact-checking, and customer service support. These systems combine extensive data repositories with sophisticated generative capabilities to produce responses that are both relevant and informative [13, 18].

Despite recent advancements, RAG systems rely on LLMs and hence face similar challenges, such as untraceable reasoning processes, supporting evidence is not provided as part of the answers, the production of "hallucinated" responses and answers that are coherent but factually incorrect or irrelevant [22]. Furthermore, integrating these systems with additional databases presents unique challenges. Since these databases are static, they can have limited coverage on topics and can lead to outdated responses. Additionally, their large volumes can result in high computational costs.

Traditional methods for evaluating RAG systems involve extensive manual annotations and continuous human monitoring, both of which are resource-intensive [44]. To address these challenges, we have developed VERA, a scalable RAG evaluation method that utilizes LLM-based evaluation mechanisms and statistical estimators to provide annotations and evaluation tools suitable for production environments.

VERA efficiently evaluates both the retrieval and generation phases of RAG systems by measuring retrieval precision and recall to ensure optimal information retrieval and assessing the faithfulness and relevance of generated answers. Additionally, VERA enhances its evaluation by leveraging a cross-encoder that incorporates these retrieval and generation metrics to yield a single comprehensive score that can be used to rank RAG systems against each other. This singular score enables users to quickly ascertain the performance of their RAG systems, as well as make any engineering decisions related to it. For instance, whether to roll-back a deployment that caused an unforeseen change to their RAG performance [30].

Furthermore, VERA introduces an innovative method that utilizes bootstrap estimators to validate and assess the topicality of

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

GenAI Evaluation KDD 2024, Barcelona, Spain,

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM

<https://doi.org/XXXXXXX.XXXXXXX>

document repositories, which is essential for both industry and academic applications, particularly as synthetic data proliferates rapidly in the GPT era. Document repository topicality for given topics refers to the degree to which the documents stored in a repository are relevant and exclusively related to the specified topics, without contamination by unrelated or off-topic content. For example, in a repository dedicated to "Cloud Computing Sales and Marketing," topicality would be measured by the proportion of documents that focus precisely on strategies, trends, and analytics specific to selling and marketing cloud computing services, while excluding unrelated topics such as healthcare management, traditional retail marketing, or general IT infrastructure. This method evaluates the topicality of a repository by examining the relevance of the documents it contains to specific topics. For example, VERA can assess the extent to which the documents in a given repository are pertinent to a designated topic or set of queries.

2 RELATED WORK

Traditionally, RAG models were evaluated based on their performance in specific downstream tasks, utilizing established metrics like EM and F1 scores for entity or sentiment classification, BERTScore and MoverScore for question answering, or accuracy for fact-checking [12, 21, 31, 36, 41, 42]. Tools like RALLE automate this process using task-specific metrics [15]. State-of-the-art evaluation tools such as EXAM and RAGAS propose various quantifications for RAG retrieval and generation effectiveness, including context relevance and answer faithfulness [2, 25]. BARTScore and SelfCheckGPT focus on generation factuality and coherency. RAG evaluation also encompasses abilities indicative of its adaptability and efficiency: noise robustness, negative rejection, counterfactual robustness, and guideline adherence [6, 19].

Despite developments in evaluation metrics and tools, quantifying different aspects in RAG remains challenging due to uncertainties in inputs and outputs and limitations of existing benchmarks in capturing human preferences. The Large Model Systems Organization (LMSYS) group explores the feasibility and pros/cons of using various LLMs as automated judge for tasks in writing, math, and world knowledge [43]. Their results reveal that strong LLM judges like GPT-4 can match both controlled and crowdsourced human preferences well, achieving over 80% agreement, the same level of agreement between humans. The G-EVAL proposed by Microsoft Cognitive Services Research group with GPT-4 as the backbone model achieves a Spearman correlation of 0.514 with human on summarization task, along with other studies confirming GPT's ability to achieve state-of-the-art or competitive correlation with human judgments [20, 32]. Furthermore, several initiatives leverage LLM prompting to evaluate performance across diverse tasks such as translation, summarization, and dialogue [14]. These studies point out that LLMs offer a scalable and explainable alternative to human evaluation, which are otherwise very expensive to obtain [43].

Lastly, given that RAGs rely on a retrieval model to retrieve relevant documents, their performance is pegged to the efficacy of the semantic search within the retriever. As the quality of semantic search is dependent on document ingestion and chunking strategies employed, the retriever can be made more robust by a re-ranking

mechanism. This is where cross-encoder models have emerged as a prominent architecture in Natural Language Processing (NLP) for tasks requiring semantic similarity assessment and textual relationship understanding [10, 27]. These models, often leveraging transformer-based encoders like BERT, process text pairs and generate a joint embedding that encodes their semantic connection [11]. This functionality allows for various applications, including sentence retrieval, question answering, and paraphrase detection [24]. Cross-encoders offer advantages in efficiency compared to methods like Siamese networks, particularly for large datasets. Additionally, their ability to leverage pre-trained language models enables effective performance even with limited task-specific training data [24].

3 VERA METHOD

VERA first systematically assesses the integrity of document repositories using LLM-based metrics, such as Retrieval Precision, Recall, Faithfulness, and Answer Relevance. It then applies advanced techniques like rank-based aggregation and bootstrapping to enhance the usability, reliability, and reproducibility of these metrics. Finally, it conducts contrastive analysis to evaluate document repository topicality [35]. This approach not only evaluates the relevance and accuracy of document retrieval but also ensures the integrity and thematic consistency of the information retrieved. Section 3.1 covers how VERA uses LLMs to generate integrity related metrics. Section 3.2 discusses rank-based aggregation. Section 3.3 introduces the bootstrapping technique. Section 3.4 details how contrastive analysis is used to assess document topicality.

3.1 LLMs as Evaluators

Recent advances in LLMs' information retrieval, understanding of nuances, and reasoning abilities have made their applications in high-stakes tasks such as system evaluations practical and feasible [4, 8]. VERA uses Anthropic Claude V3 Haiku through Amazon Bedrock service as the default LLM for RAG evaluations, due to Haiku's balance between cost and effectiveness. Haiku achieves competitive performance on major reasoning dataset: 75.2% on MMLU [9], 89.2% on ARC-Challenge [40] and 85.9% on HellaSwag [1]. On each of the dataset, it has surpassed GPT-3.5 over all those three evaluation benchmark datasets. A different LLM can be chosen based on the model's merits, specific use cases, and costs.

Like existing LLM-based RAG evaluation system such as RAGAS or ARES [25], VERA has measured the following LLM-based evaluation metrics. The prompts to create the metrics are listed in Appendix 8.1.

Faithfulness: This metric evaluates whether answers are based solely on the provided contexts, without any fabrication. The prompt will instruct the language model to generate a binary "yes" or "no" label for each (q, a, c) pair, where q , a , and c represent the question, answer, and context, respectively. The faithfulness metric for a set of (q, a, c) pairs is calculated as the average of all the binary labels.

Retrieval Recall: This metric evaluates the system's effectiveness in fetching all relevant information related to a query from the given context, ensuring that no significant data is omitted. This metric is determined by assessing whether each piece of information in the answer is explicitly supported by the context. The retrieval

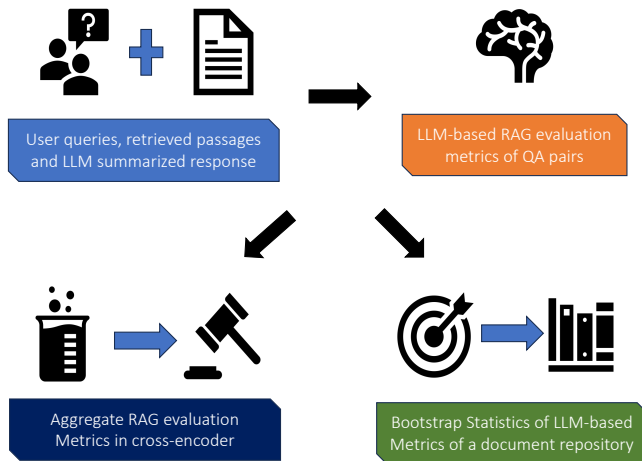


Figure 1: Overview of VERA: VERA begins with user queries, pairing them with retrieved and LLM summarized responses from a given RAG system. These elements form the basis for the LLM-based RAG evaluation of individual question-answer pairs, ensuring that the context relevance, answer faithfulness, and answer relevance metrics are meticulously assessed. These metrics are then consolidated using a cross-encoder to generate an aggregate score, enabling users to prioritize certain metrics over others and quickly make outcome-oriented decisions for development. The process then culminates with Bootstrap Statistics, which apply LLM-based metrics across the entire document repository to establish confidence bounds and gauge the overall performance of retrieval systems. This robust evaluation pipeline is essential for maintaining high standards of precision and trustworthiness in document retrieval, particularly critical in domains where the accuracy of information is paramount.

recall metric is calculated based on the proportion of sentences in the answer that are classified as "[Supported by Context]", which is used in the evaluation metric prompt in Appendix 8.1. This involves:

- Counting the total number of sentences in the answer.
- Counting the number of sentences classified as "[Supported by Context]".
- Calculating the ratio of "[Supported by Context]" sentences to the total number of sentences.

Retrieval Precision: This metrics assesses the system’s ability to focus on and retrieve the most relevant parts of the context in response to a query, minimizing the inclusion of irrelevant content. High precision ensures that the model considers only the information that is directly pertinent to the question. The retrieval precision metric is calculated by evaluating the relevance of sentences extracted from the context using LLM. This involves:

- Extracting sentences from the context that directly support the answer to the question.
- Ensuring that sentences are not altered when extracted.

- Returning "Insufficient Information" if no relevant sentences are found or if the context does not provide enough information to answer the question.
- Measuring the similarity between the extracted sentences and the context using embedding models. In this work, we used Amazon Titan embedding model [3].
- Calculating the precision based on the ratio of relevant sentences to the total number of sentences in the context.

Answer Relevance: This metrics evaluates whether the generated response directly addresses the given question, ensuring alignment with both the query and the retrieved context. This metric penalizes responses that are incomplete, redundant, or contain unnecessary information, providing a score that ranges from 0 to 1, with 1 being the highest level of relevance. The answer relevance metric is calculated through the following steps:

- For each generated answer, multiple questions are generated to assess the alignment of the answer with the query.
- The similarity between the generated questions and the original question is measured using embeddings. This involves:
 - (1) Embedding the original question and the generated questions using an embedding model. In this work, we used Amazon Titan embedding model [3].
 - (2) Calculating the cosine similarity between the original question embedding and each generated question embedding.
- The final score is computed as the mean cosine similarity across all generated questions for each answer, reflecting the degree of relevance.

3.2 Consolidation of Multi-Dimensional Evaluation Metrics

The concept of consolidating evaluation metrics into a single comprehensive score involves integrating the utilities of each metric, allowing users to make informed decisions despite the inherent fluctuations in these metrics. Appropriate consolidation eases the burden of users having to parse through multiple metrics to then make a decision based on the outcome - which would improve iteration speed in a development cycle. Furthermore, given that each of these multi-dimensional metrics has its nuances, the question of how to prioritize certain metrics over others arises (e.g. does a system with higher faithfulness and lower relevance outperform the system with lower faithfulness and higher relevance). This would assist users to swiftly take action during regression testing, to make decisions on whether to roll-back a deployment or not.

Traditional techniques like simple aggregation or rank fusion often suffer from compensatory effects and lack clarity, as they obscure the subtleties of individual metrics [5, 26].

To address these challenges, VERA utilizes cross-encoder models that leverage a cross-attention mechanism for a more precise evaluation of document relevance. Traditional cross-encoder models are effective at highlighting the most relevant text segments within large texts, based on capturing semantic relationships between words and phrases. It generates a relevance score for every question-answer pair, enabling an effective comparison and ranking of these pairs. Formally, for a user-input question q and an answer

a, the logit score σ is determined as:

$$\sigma(q, a) = CE([CLS] \ q \ [SEP] \ a \ [SEP]) \times W$$

where CE is the cross-encoder, CLS and SEP are special tokens to represent the classifier token and the separator token, and W is a learned matrix that represents the relationship between the query and answer representations [30].

Recently, effective multi-dimensional retrieval models are typically implemented by performing a first-stage retrieval (to efficiently identify a subset of relevant documents from a corpus); and a second-stage re-ranking on this subset (where additional dimensions of relevance may be considered) [30]. An example of this to conduct the first stage retrieval using the BM-25 algorithm (which is a ranking algorithm that determines a document’s relevance to a given query and ranks documents based on their relevance scores). After this, the second-stage re-ranker modifies the architecture of existing cross-encoders, whereby the BM-25 score obtained in the first-stage retrieval is fed as an input token to the cross-encoder. Mathematically, this is represented by:

$$\sigma(q, a) = CE([CLS] \ q \ [SEP] \ BM25 \ [SEP] \ a \ [SEP]) \times W$$

In this paper, we follow a similar process to incorporate additional dimensions of relevance into a cross-encoder to re-rank evaluation records against each other. However, instead of manipulating the input structure of the cross-encoder, we integrate additional "relevance statements" into each question-answer pair that is fed into the cross-encoder [30]. These relevance statements pertain to texts related to the utility of each of the multi-dimensional evaluation metrics, as well as their actual scores. As shown in [30], this exercise yields 4-5 percent improvement in Mean Average Precision, Normalized Discounted Cumulative Gain and Mean Reciprocal Rank metrics against baseline cross-encoder models.

The process involves two key steps: first, enhancing the input text with additional relevant information mentioned above, and second, providing the queries and "enhanced" answer as inputs to the pretrained cross-encoder to obtain the final aggregated score (which can be used to rank these records against each other) [30]. This structured approach ensures a thorough and nuanced assessment of document relevance.

Text Enhancement: The cross-encoder requires an input of an input and output text. Within VERA, the input text will be the user query input to a RAG system; and the output text will be an enriched answer encapsulating the original response from the RAG system, together with supplementary information about each evaluated metric’s utility as well as their scores. For instance, a question-answer pair (q,a) obtains a faithfulness score of 0.7 and its enriched answer \bar{a} is generated by appending the original response from the RAG system with the following text:

"Faithfulness measures the factual consistency of the generated answer against the given context. It is considered faithful if all the claims that are made in the answer can be inferred from the given context. It is measured between 0 and 1; where a lower score is given to answers consisting of claims that are not in the context; and a higher score indicates that the answer is using information from the contexts. For the given question, context and answer, the faithfulness score is 0.7."

Cross-Encoder Ranking: Once the text enhancement step is done, VERA feeds in the question and the enhanced answer into the ms-marco-MiniLM-L-12-v2 (top cross-encoder model on MTEB Leaderboard). Formally, for a user-input question q and enriched answer \bar{a} , the logit score σ is determined as:

$$\sigma(q, \bar{a}) = CE([CLS] \ q \ [SEP] \ \bar{a} \ [SEP]) \cdot W$$

As this cross-encoder model was trained to learn logit values, it can be normalized to a value between 0 and 1 by taking the expit. However, this paper will present the results as logit scores.

3.3 Bootstrap LLM-based RAG Evaluation Metrics

Evaluating RAG systems requires measuring retrieval precision, recall, faithfulness, and relevance. However, these metrics can vary due to LLMs’ stochasticity, reasoning limitations, and document repository topicality. To address this, we used bootstrapping on pre-computed metric values. This approach enhances result reliability and reproducibility by providing a robust statistical framework to analyze metric variability and distribution, while also supporting document repository topicality assessment for specific content types [14].

LLMs can produce varying outputs due to factors like random seed values, causing traditional evaluation to capture only a snapshot of this variability and potentially misleading performance conclusions. By applying bootstrap directly on the metric values, we can simulate multiple runs of model evaluations, capturing a broader spectrum of possible outcomes and thus providing a more comprehensive picture of system performance.

Bootstrapping metric values allows for repeated sampling from a set of observed metric computations, essentially creating numerous virtual evaluation scenarios. The bootstrapping metric values computation are identical for all metrics.

Given a known metric M : Firstly, compute its values for a document repository dataset $D = \{d_1, d_2, \dots, d_n\}$ as $M(d_i)$ for each document d_i . This results in a set of metric values $M = \{m_1, m_2, \dots, m_n\}$. Then, for each metric M , generate B bootstrap samples. Each sample s is created by randomly selecting metric values from M with replacement. Each bootstrap sample for metric M can be represented as $M_s = \{m_1^s, m_2^s, \dots, m_n^s\}$. For each bootstrap sample, compute the desired statistics, such as the sample mean and variance as below:

- **Estimates the Mean and Variability:** Provides a statistically robust way to estimate the mean and variance of performance metrics, incorporating the inherent randomness of LLM outputs. The mean \bar{M} of the bootstrap samples is estimated as:

$$\bar{M} = \frac{1}{B} \sum_{s=1}^B \bar{m}_s$$

And the variance $\sigma^2(M)$ is:

$$\sigma^2(M) = \frac{1}{B-1} \sum_{s=1}^B (\bar{m}_s - \bar{M})^2$$

- **Confidence intervals:** Can be derived from the percentiles of the bootstrap distribution, typically the 2.5th and 97.5th percentiles for a 95% confidence interval.

Bootstrapping Size B versus Sample Size n : There is no strict universal rule for the optimal bootstrapping size relative to the sample size. However, bootstrapping tends to work well when the sample size (n) is at least 30 and ideally 50 or more for accurate estimations of standard errors and confidence intervals [29, 37]. Bootstrapping size is recommended to be at least 1000 and 5000 or more for stable convergence and complex statistics. With larger sample size, a smaller bootstrapping size is possible with similar accuracy. It is recommended to monitor the convergence of standard errors or other statistics as B increases, to determine the optimal bootstrapping size for each use case.

Unbiased Estimator: The bootstrap estimator serves as an unbiased estimator for metrics based on LLMs, effectively estimating the expectation of the original estimator and its bootstrap distribution. Detailed assumptions and mathematical derivations supporting this conclusion are outlined in Appendix 8.2.

3.4 Evaluating Document Repository Topicality Using Contrastive Query Analysis

Document repositories often contain diverse content, leading to high entropy for domain-specific information retrieval and making it challenging to ascertain the repository’s thematic topicality, especially in specialized industry domains. To address this, we implement a contrastive analysis framework, differentiating responses to topic perfect relevant queries (positive instances) from responses to unrelated queries (negative controls). Within the framework, we have proposed a bootstrap estimation approach that provides a structured statistical analysis to evaluate the repository’s thematic consistency.

The approach involves several key steps with the idea ignited from contrastive learning [7]:

- **Query Generation:** Develop two distinct sets of queries. Positive queries set are relevant to a specific domain of interest, and negative queries are deliberately chosen to be unrelated to that domain.
- **Retrieval and Evaluation:** Utilize a large language model (LLM) or a similar retrieval system to fetch and evaluate responses for each query. Evaluation metrics such as Retrieval Precision, Recall, Faithfulness, and Answer Relevance are calculated to assess the quality and relevance of the responses.
- **Bootstrap Statistics:** Apply bootstrap sampling techniques to each evaluation metrics. This involves generating numerous subsamples from the collected metrics and computing statistical measures (e.g., mean, variance) for these samples to analyze the data robustly.
- **Comparative Analysis:** Compare the distributions of these bootstrap statistics between the positive and negative query sets. This step quantitatively assesses the repository’s content alignment with the domain of interest and identifies any significant disparities in content handling between relevant and irrelevant queries.

4 EXPERIMENTS

In this section, we present the models and data used by VERA. VERA uses both public and proprietary datasets to ensure a comprehensive analysis. We utilize the open-source MS MARCO in

TREC 2023 Deep Learning Track for a general knowledge [17]. Simultaneously, we incorporate proprietary datasets tailored to AWS sales and marketing domain-specific evaluations, reflecting the unique challenges and requirements of different industry sectors. This combination allows us to assess the general applicability and targeted performance of our RAG systems, facilitating a thorough understanding of their capabilities and areas for optimization in real-world scenarios.

4.1 Models

For domain-specific synthetic data generation, we employ Anthropic V3 Haiku to create high-quality synthetic queries and responses tailored for our experimental needs. This model’s advanced generative capabilities ensure that the synthetic datasets are both diverse and closely aligned with the task-specific requirements. For the evaluation of responses, we utilize Anthropic V3 Sonnet, which serves as our LLM judge. The examples of synthetic generation prompts, evaluation prompts and RAG summarization prompt using Llama 3 supported in POE Web UI [28] are in Appendix 8.1.

In our experiments, we compared the performance of multiple RAG systems by pairing different combinations of LLMs—specifically Anthropic Haiku and Llama3—with a selection of advanced retrievers. The retrievers we’ve chosen include e5-mistral-7b-instruct, titan-embedding-text-G1, and bge-large-en-v1.5, all of which are recognized as top models on the MTEB leaderboard, indicative of their superior performance and capability in handling complex retrieval tasks [3, 33, 34, 38]. This diverse combination of cutting-edge LLMs and retrievers allows us to thoroughly assess and contrast the strengths and limitations of different RAG configurations in producing relevant and accurate responses.

4.2 Datasets

The TREC 2023 Deep Learning Track emphasizes enhancing information retrieval with large-scale datasets suitable for deep learning, focusing on passage and document ranking tasks. It utilizes the MS MARCO dataset to analyze and develop effective retrieval and reranking systems in real-world scenarios. In this research, we’ll focus on using the smaller passage ranking data from the TREC 2023 Deep Learning Track for experimental purpose. For the purpose of our experiments, we have used all 887 unique perfectly relevant query-passage pairs (score=3) from "2021.qrels.docs.final.txt" and 500 randomly sampled irrelevant query-passage pairs (score=0).

Additionally, we have generated 400 passages related to cloud computation sales and marketing topic and 100 passages related to basketball topic. Then, we have created 200 queries about cloud computation sales and marketing, 200 queries about basketball and 200 random queries not related to both topics.

5 RESULTS & ANALYSIS

5.1 VERA LLM-Based RAG Evaluation Metrics

In this experiment, we evaluate the performance of several RAG systems by comparing perfectly relevant and irrelevant query-passages pairs on faithfulness, answer relevance, retrieval recall, retrieval precision, as well as the logit values returned by the cross-encoder when performing the aggregation step ("Agg" column in tables).

Findings are presented in Table 1 and Appendix Table 3. All the metric values in both tables are reported as mean.

We feed the top 5 retrieved passages for a given query into a LLM to generate the final summarized response. In this experiment, the dataset labels "PR" and "IR" stand for "Perfectly Relevant" and "Irrelevant," respectively. To make the results more deterministic and less affected by the randomness inherent in LLMs, we have implemented the following settings: *temperature* = 0; *top p* = 0.01. The metric values in both tables are mean of all queries' results.

The experimental results demonstrate that Llama3, a powerful open-source LLM, performs comparably to established models like Anthropic's Claude V3 Haiku. In Table 1, these models manage effective fact-checking and capture semantic relationships well, as indicated by high faithfulness and relevance scores. Additionally, the retrieval recall and precision are reasonably high, suggesting that the models retrieve most relevant information accurately. In the opposite way, the low precision score in Appendix Table 3 may suggest that the queries either fall outside the scope of the covered topics in the knowledge base, or that the topics within the knowledge base are too varied relative to the generality of the queries. The Agg-Logit scores in the comparison between different model configurations highlight the nuanced performance differences across various metrics.

Lastly, the performance of these powerful LLMs and embedding models have been compared to that of a weaker LLM (T-5 FLAN Base) and embedding model (all-MiniLM-L6-v2) as a "baseline", to highlight the differences in the evaluation metrics. As seen in both Table 1 and Table 2, the individual evaluation metrics as well as the Agg-Logit scores are consistently lower when using a weaker LLM + embedding model, regardless of the scenario of evaluating against a perfectly relevant or irrelevant dataset.

5.2 Bootstrap Metrics for Document Repository Topicality Analysis

As outlined in section 3.3, we utilized bootstrap statistics to analyze a synthetic dataset described in section 4.2 and the results are in Table 3. We used bootstrap sampling with replacement on the synthetic query sets and the overall passage set. In the synthetic query sets, we have 200 synthetic queries in each set and we labeled them in Table 3 as "Sales", "Basketball" and "Random" based on the topics. This approach enabled us to calculate critical statistical measures like the mean and variance, providing a robust foundation for assessing the thematic topicality of the data repository. We used sample size 50 and bootstrapping size 500 to ensure fairly stable convergence of the statistics for each metric and each query set. This comparative analysis helps in quantifying the document repository's content topicality to distinguish and accurately process content relevant to its designated domain.

In our study, the use of bootstrap statistics enabled us to compute the mean and confidence intervals for each performance metric across three different synthetic query sets on the same document repository. This comparison revealed notable differences in retrieval-related metrics among the query sets regarding different topics. The "Sales" query set results are with higher values in recall, precision, and relevance as the majority (80%) of the synthetic passage set is related cloud computation sales and marketing data.

As comparison, the "Basketball" query set results are much higher than the "Random" query set and fairly lower than the "Sales" query set, which is within expectation and validated the effectiveness of bootstrapping approach to evaluate the document repository topicality.

6 CONCLUSION

In this paper, we introduced VERA, a framework tailored for evaluating Retrieval-Augmented Generation (RAG) systems. By generating LLM-based RAG evaluation metrics such as faithfulness, answer relevance, retrieval precision and retrieval recall, VERA can help evaluate and validate if the response from RAG based AI assistant is accurate or not. This framework boosts the reliability and transparency of RAG systems and build the trust in AI applications for users.

Our findings demonstrate VERA's capacity to enhance decision-making processes effectively. The framework has been applied across several use cases, illustrating its ability to adapt to dynamic environments and maintain the integrity of data repositories. This adaptability makes VERA an important tool in the landscape of modern AI technologies, where the accuracy and relevance of information are paramount.

Looking forward, we aim to further refine the metrics within VERA and expand its applicability to a broader range of domains and languages. Continuous advancements in VERA's methodologies will allow it to keep pace with rapid technological developments in AI. This evolution will ensure that emerging AI technologies are leveraged responsibly, maximizing their potential benefits for society.

7 LIMITATIONS

This paper presents several limitations that could potentially impact the comprehensiveness and applicability of its conclusions. Firstly, the analysis omits scenarios involving fine-tuned LLMs. Potential enhancements or specific use-case efficiencies brought by fine-tuned models might not be fully captured. This omission could lead to an incomplete understanding of the capabilities and limitations of the models under different experimental conditions. And the exclusion of some top proprietary LLMs in our experiments, such as OpenAI models, limits the evaluation's scope and understanding of our selected models' performance against the best available options.

Secondly, our study does not address multilingual capabilities. The focus solely on English-language tasks may limit the generalizability of our conclusions to multilingual applications. This oversight could restrict the utility of our findings for developers and researchers working on systems intended for diverse linguistic environments, potentially overlooking significant performance variations across languages.

Thirdly, although our bootstrap estimators offer a more convincing assessment of the content complexity within a document repository, they are computationally intensive. We aim to develop a theoretically grounded, cost-effective measurement approach by constructing a pseudo-bootstrap strategy. This strategy will utilize pre-calculated evaluation metrics instead of relying on bootstrap sampling from queries.

Table 1: VERA LLM-Based RAG Evaluation Metrics on 500 Perfectly Relevant MS MARCO TREC 2023 Query-Passage Pairs

Models	Dataset	Faithfulness	Relevance	Recall	Precision	Aggregate Logit Score
Llama3 + e5-mistral-7b-instruct	PR	0.94	0.87	0.76	0.68	8.72
Llama3 + titan-embedding-text-G1	PR	0.93	0.81	0.74	0.63	8.69
Llama3 + bge-large-en-v1.5	PR	0.94	0.83	0.77	0.64	8.70
Haiku + e5-mistral-7b-instruct	PR	0.95	0.85	0.75	0.65	8.48
Haiku + titan-embedding-text-G1	PR	0.93	0.81	0.73	0.64	8.59
Haiku + bge-large-en-v1.5	PR	0.94	0.82	0.74	0.63	8.46
T-5 Flan + all-MiniLM-L6-v2	PR	0.82	0.53	0.61	0.50	6.11

Table 2: VERA LLM-Based RAG Evaluation Metrics on 500 Irrelevant MS MARCO TREC 2023 Query-Passage Pairs

Models	Dataset	Faithfulness	Relevance	Recall	Precision	Agg
Llama3 + e5-mistral-7b-instruct	IR	0.94	0.20	0.10	0.12	6.45
Llama3 + titan-embedding-text-G1	IR	0.94	0.13	0.09	0.11	6.23
Llama3 + bge-large-en-v1.5	IR	0.95	0.12	0.11	0.14	6.39
Haiku + e5-mistral-7b-instruct	IR	1.0	0.08	0.03	0.12	6.38
Haiku + titan-embedding-text-G1	IR	1.0	0.29	0.22	0.12	6.26
Haiku + bge-large-en-v1.5	IR	1.0	0.21	0.10	0.13	6.55
T-5 Flan + all-MiniLM-L6-v2	IR	0.87	0.12	0.02	0.05	3.81

Table 3: VERA Bootstrap Statistics on Three Comparative QuerySet

Models	QuerySet	Faithfulness	Relevance	Recall	Precision
Llama3 + e5-mistral-7b-instruct	Sales	0.93±0.03	0.71±0.04	0.61±0.07	0.54±0.09
Llama3 + titan-embedding-text-G1	Sales	0.94±0.03	0.70±0.04	0.62±0.08	0.55±0.08
Llama3 + bge-large-en-v1.5	Sales	0.93±0.03	0.70±0.05	0.60±0.07	0.53±0.10
Haiku + e5-mistral-7b-instruct	Sales	0.93±0.02	0.72±0.05	0.62±0.07	0.55±0.06
Haiku + titan-embedding-text-G1	Sales	0.94±0.03	0.71±0.04	0.63±0.07	0.56±0.07
Haiku + bge-large-en-v1.5	Sales	0.93±0.02	0.70±0.05	0.61±0.08	0.56±0.09
Llama3 + e5-mistral-7b-instruct	Basketball	0.94±0.03	0.67±0.05	0.56±0.07	0.43±0.09
Llama3 + titan-embedding-text-G1	Basketball	0.93±0.03	0.66±0.06	0.53±0.07	0.42±0.08
Llama3 + bge-large-en-v1.5	Basketball	0.94±0.03	0.66±0.05	0.54±0.08	0.45±0.10
Haiku + e5-mistral-7b-instruct	Basketball	0.95±0.02	0.66±0.06	0.53±0.09	0.43±0.09
Haiku + titan-embedding-text-G1	Basketball	0.94±0.03	0.65±0.06	0.52±0.08	0.45±0.08
Haiku + bge-large-en-v1.5	Basketball	0.93±0.02	0.66±0.05	0.54±0.08	0.44±0.08
Llama3 + e5-mistral-7b-instruct	Random	0.93±0.02	0.23±0.04	0.13±0.05	0.09±0.05
Llama3 + titan-embedding-text-G1	Random	0.93±0.03	0.21±0.05	0.15±0.04	0.10±0.04
Llama3 + bge-large-en-v1.5	Random	0.94±0.03	0.16±0.05	0.11±0.04	0.08±0.05
Haiku + e5-mistral-7b-instruct	Random	0.94±0.02	0.24±0.06	0.12±0.04	0.10±0.04
Haiku + titan-embedding-text-G1	Random	0.92±0.03	0.22±0.06	0.14±0.05	0.09±0.05
Haiku + bge-large-en-v1.5	Random	0.93±0.03	0.17±0.05	0.14±0.04	0.08±0.05

Lastly, our study did not analyze all popular publicly available benchmarks such as the Knowledge Intensive Language Tasks (KILT) benchmark, which could have provided additional insights into the models’ capabilities in retrieving, reasoning, and synthesizing information from knowledge bases in real-world scenarios [16, 23, 28, 39].

REFERENCES

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
- [2] John Allard and Colin Jarvis. 2023. A survey of techniques for maximizing llm performance. (2023).

- [3] Amazon Web Services. 2024. Amazon Titan Embeddings G1 - Text. <https://docs.aws.amazon.com/bedrock/latest/userguide/model-parameters-titan-embed-text.html>. Accessed on: April 19, 2024.
- [4] Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku. <https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/ModelCardClaude3.pdf>. Accessed: 2024.
- [5] KS Arun, VK Govindan, and SD Madhu Kumar. 2017. On integrating re-ranking and rank list fusion techniques for image retrieval. *International Journal of Data Science and Analytics* 4 (2017), 53–81.
- [6] Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 17754–17762.
- [7] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. 2020. Debaised contrastive learning. *Advances in neural information processing systems* 33 (2020), 8765–8775.
- [8] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research* 25, 70 (2024), 1–53.
- [9] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457* (2018).
- [10] Dave, Neil. 2024. Understanding Cross-Encoders: Revolutionizing Textual Relationships in NLP. <https://theneildave.medium.com/understanding-cross-encoders-revolutionizing-textual-relationships-in-nlp-86b69254e08>. Accessed on: May 23, 2024.
- [11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [12] Zhangyin Feng, Xiaocheng Feng, Dezhi Zhao, Maojin Yang, and Bing Qin. 2024. Retrieval-generation synergy augmented large language models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 11661–11665.
- [13] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*. PMLR, 3929–3938.
- [14] Tim Hesterberg. 2011. Bootstrap. *Wiley Interdisciplinary Reviews: Computational Statistics* 3, 6 (2011), 497–526.
- [15] Yasuto Hoshi, Daisuke Miyashita, Youyang Ng, Kento Tatsuno, Yasuhiro Morioka, Osamu Torii, and Jun Deguchi. 2023. Ralle: A framework for developing and evaluating retrieval-augmented large language models. *arXiv preprint arXiv:2308.10633* (2023).
- [16] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics* 7 (2019), 453–466.
- [17] Dawn Lawrie, Sean MacAvaney, James Mayfield, Paul McNamee, Douglas W Oard, Luca Soldaini, and Eugene Yang. 2024. Overview of the TREC 2023 NeuCLIR Track. *arXiv preprint arXiv:2404.08071* (2024).
- [18] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.
- [19] Yi Liu, Lianzhe Huang, Shicheng Li, Shishuo Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023. Recall: A benchmark for llms robustness against external counterfactual knowledge. *arXiv preprint arXiv:2311.08147* (2023).
- [20] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634* (2023).
- [21] Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting for retrieval-augmented large language models. *arXiv preprint arXiv:2305.14283* (2023).
- [22] David Nadeau, Mike Kroutikov, Karen McNeil, and Simon Baribeau. 2024. Benchmarking Llama2, Mistral, Gemma and GPT for Factuality, Toxicity, Bias and Propensity for Hallucinations. *arXiv preprint arXiv:2404.09785* (2024).
- [23] Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. 2020. KILT: a benchmark for knowledge intensive language tasks. *arXiv preprint arXiv:2009.02252* (2020).
- [24] Reimers, Nils. 2024. Sentence Transformers. <https://www.sbert.net/docs/>. Accessed on: May 23, 2024.
- [25] Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2023. Ares: An automated evaluation framework for retrieval-augmented generation systems. *arXiv preprint arXiv:2311.09476* (2023).
- [26] Krystian Safjan. 2023. Rank Fusion Algorithms - From Simple to Advanced. (2023). <https://safjan.com/Rank-fusion-algorithms-from-simple-to-advanced/>.
- [27] Soman, Aneesh B. 2024. Exploring Diverse Techniques for Sentence Similarity. <https://medium.com/@aneeshb161994/exploring-diverse-techniques-for-sentence-similarity-bc62058c7972>. Accessed on: May 23, 2024.
- [28] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. *arXiv preprint arXiv:1803.05355* (2018).
- [29] Robert J Tibshirani and Bradley Efron. 1993. An introduction to the bootstrap. *Monographs on statistics and applied probability* 57, 1 (1993), 1–436.
- [30] Rishabh Upadhyay, Arian Askari, Gabriella Pasi, and Marco Viviani. 2023. Enhancing Documents with Multidimensional Relevance Statements in Cross-encoder Re-ranking. *arXiv preprint arXiv:2306.10979* (2023).
- [31] Boxin Wang, Wei Ping, Lawrence McAfee, Peng Xu, Bo Li, Mohammad Shoneybi, and Bryan Catanzaro. 2023. Instructretro: Instruction tuning post retrieval-augmented pretraining. *arXiv preprint arXiv:2310.07713* (2023).
- [32] Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048* (2023).
- [33] Liang Wang, Nan Yang, Xiaolong Huang, Bingxin Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533* (2022).
- [34] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368* (2023).
- [35] Peiling Wang and Dagobert Soergel. 1998. A cognitive model of document use during a research project. Study I. Document selection. *Journal of the American Society for Information Science* 49, 2 (1998), 115–133.
- [36] Shi Weijia, Min Sewon, Yasunaga Michihiro, Seo Minjoon, James Rich, Lewis Mike, et al. 2023. REPLUG: Retrieval-augmented black-box language models. *ArXiv: 2301.12652* (2023).
- [37] Chien-Fu Jeff Wu. 1986. Jackknife, bootstrap and other resampling methods in regression analysis. *the Annals of Statistics* 14, 4 (1986), 1261–1295.
- [38] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighof. 2023. C-pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597* (2023).
- [39] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600* (2018).
- [40] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830* (2019).
- [41] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).
- [42] Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Stefan Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. *arXiv preprint arXiv:1909.02622* (2019).
- [43] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems* 36 (2024).
- [44] Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022. Dialoglm: Pre-trained model for long dialogue understanding and summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 11765–11773.

8 APPENDIX

8.1 Prompt Examples

Prompt of Faithfulness Metric

Consider the given context and following statements, then determine whether they are supported by the information present in the context. Provide a brief explanation for each statement before arriving at the verdict (Yes/No). Provide a final verdict for each statement in order at the end in the given format. Do not deviate from the specified format.

Context:

Emma is a graduate student specializing in marine biology at Coastal University. She has a keen interest in coral reefs and is conducting her thesis on coral bleaching. Emma attends several seminars

related to marine ecosystems and is actively involved in field research in the nearby coral reefs. She often collaborates with other researchers to publish her findings.

Statements:

- (1) Emma is studying mechanical engineering.
- (2) Emma is working on a project related to coral reefs.
- (3) Emma often attends computer science workshops.
- (4) Emma collaborates with other researchers.
- (5) Emma's research focuses on marine ecosystems.

Answer:

1. Emma is studying mechanical engineering.

Explanation: The context specifies that Emma is specializing in marine biology, not mechanical engineering. There is no information suggesting she is studying mechanical engineering.

Verdict: No.

2. Emma is working on a project related to coral reefs.

Explanation: It is mentioned that Emma is conducting her thesis on coral bleaching, which directly relates to coral reefs.

Verdict: Yes.

3. Emma often attends computer science workshops.

Explanation: The context only mentions that Emma attends seminars related to marine ecosystems. There is no mention of her attending computer science workshops.

Verdict: No.

4. Emma collaborates with other researchers.

Explanation: The context explicitly states that Emma often collaborates with other researchers to publish her findings.

Verdict: Yes.

5. Emma's research focuses on marine ecosystems.

Explanation: Emma's interest in coral reefs and her participation in seminars related to marine ecosystems suggest that her research focuses on marine ecosystems.

Verdict: Yes.

Final verdict for each statement in order: No. Yes. No. Yes. Yes.

Context: {context}

Statements: {statements}

Answer:

Prompt of Retrieval Recall Metrics

Task: Given a context and an answer, analyze each sentence in the answer and classify whether the sentence is supported by the given context or not. Think in steps and reason before coming to a conclusion.

Guidelines:

- (1) Read each sentence in the answer carefully.
- (2) Compare the sentence with the context to see if the information is explicitly mentioned.

- (3) Classify each sentence as either [Supported by Context] or [Not Supported by Context].

- (4) Provide a brief reasoning for your classification.

Example 1:

Context:

Isaac Newton (25 December 1642 – 20 March 1726/27) was an English mathematician, physicist, astronomer, alchemist, and author. He is widely recognized as one of the most influential scientists of all time and a key figure in the scientific revolution. His book "Philosophiæ Naturalis Principia Mathematica," first published in 1687, laid the foundations of classical mechanics. Newton made seminal contributions to optics and shares credit with Gottfried Wilhelm Leibniz for developing calculus.

Answer:

Isaac Newton was an English mathematician, physicist, and astronomer. He is known for writing "Philosophiæ Naturalis Principia Mathematica." Newton invented calculus independently of Leibniz.

Classification:

- (1) Isaac Newton was an English mathematician, physicist, and astronomer. This information is in the context. So [Supported by Context]
- (2) He is known for writing "Philosophiæ Naturalis Principia Mathematica." This is explicitly mentioned in the context. So [Supported by Context]
- (3) Newton invented calculus independently of Leibniz. The context mentions Newton shares credit with Leibniz for developing calculus but does not state he did it independently. So [Not Supported by Context]

Example 2:

Context:

Marie Curie (7 November 1867 – 4 July 1934) was a Polish and naturalized-French physicist and chemist who conducted pioneering research on radioactivity. She was the first woman to win a Nobel Prize, the only woman to win the Nobel prize twice, and the only person to win the Nobel Prize in two different scientific fields. Her achievements include the development of the theory of radioactivity, techniques for isolating radioactive isotopes, and the discovery of two elements, polonium and radium.

Answer:

Marie Curie was a Polish physicist who won the Nobel Prize twice. She discovered the elements polonium and radium. Curie was the first person to win Nobel Prizes in two different fields.

Classification:

- (1) Marie Curie was a Polish physicist who won the Nobel Prize twice. This is explicitly mentioned in the context. So [Supported by Context]
- (2) She discovered the elements polonium and radium. This is explicitly mentioned in the context. So [Supported by Context]

(3) Curie was the first person to win Nobel Prizes in two different fields. This is explicitly mentioned in the context. So [Supported by Context]

Context: {context}

Answer: {ground_truth}

Classification:

Prompt of Retrieval Precision Metric

Task: Evaluate whether the provided context can answer the given question by extracting relevant sentences. Follow these guidelines:

- (1) **Extract Sentences:** Identify and extract sentences from the context that directly support an answer to the question.
- (2) **No Modifications:** Do not alter the sentences when extracting them.
- (3) **Insufficient Information:** If no relevant sentences are found or if the context does not provide enough information to answer the question, return "Insufficient Information".

Examples:

Example 1:

Question: What causes the tides to rise and fall?

Context: The gravitational pull of the moon and the sun causes the tides to rise and fall. The moon’s gravity has a greater effect because it is closer to the Earth, creating high and low tides. The sun also plays a role, but to a lesser extent.

Candidate Sentences:

- The gravitational pull of the moon and the sun causes the tides to rise and fall.
- The moon’s gravity has a greater effect because it is closer to the Earth, creating high and low tides.

Example 2:

Question: Who discovered penicillin?

Context: Penicillin was discovered by Alexander Fleming in 1928. He noticed that a mold called *Penicillium notatum* had killed a staphylococcus bacterium in a petri dish. This discovery led to the development of antibiotics, which have saved countless lives.

Candidate Sentences:

- Penicillin was discovered by Alexander Fleming in 1928.
- He noticed that a mold called *Penicillium notatum* had killed a staphylococcus bacterium in a petri dish.

Example 3:

Question: What is the capital of Atlantis?

Context: Many myths surround the lost city of Atlantis, but no concrete evidence has ever been found to confirm its existence. Some legends suggest it was a powerful civilization located in the Atlantic Ocean, but its exact location and details remain unknown.

Candidate Sentences:

Insufficient Information

Question: {question}

Context:

{context}

Candidate Sentences:

Question Generation Prompt for Answer Relevance Metric

Task: Generate a question based on the given answer. The question should be specific, clear, and directly related to the information provided in the answer.

Guidelines:

- (1) **Identify Key Information:** Carefully read the given answer to identify the key pieces of information. These may include dates, times, locations, events, people, etc.
- (2) **Formulate the Question:** Create a question that specifically asks about the key information you identified in the answer. The question should be comprehensive and direct, ensuring it covers all the important details provided in the answer.
- (3) **Ensure Clarity and Specificity:** The question should be clear and specific, leaving no ambiguity about what information it seeks. It should be framed in a way that the answer provided directly responds to it.
- (4) **Maintain Formality and Precision:** Use formal language and precise wording to ensure the question is professional and easy to understand.

Examples:

Example 1

Answer:

The PSLV-C56 mission is scheduled to be launched on Sunday, 30 July 2023 at 06:30 IST / 01:00 UTC. It will be launched from the Satish Dhawan Space Centre, Sriharikota, Andhra Pradesh, India.

Question:

When is the scheduled launch date and time for the PSLV-C56 mission, and where will it be launched from?

Example 2

Answer:

The Great Wall of China, built over several dynasties, stretches approximately 13,170 miles. It was primarily constructed to protect Chinese states and empires from invasions and raids.

Question:

What is the length of the Great Wall of China, and why was it primarily constructed?

Example 3

Answer:

Marie Curie was awarded the Nobel Prize in Physics in 1903 and the Nobel Prize in Chemistry in 1911 for her work on radioactivity and the discovery of radium and polonium.

Question:

In which years did Marie Curie receive her Nobel Prizes, and for what contributions were they awarded?

Example 4**Answer:**

The Amazon Rainforest, often referred to as the "lungs of the Earth," covers over 5.5 million square kilometers and spans across nine countries in South America.

Question:

What is the total area of the Amazon Rainforest, and across how many countries does it span?

Now, generate a question based on the given answer below:

Answer:

{answer}

Question:

Enhanced Document Context for Cross-Encoder

Question: At about what age do adults normally begin to lose bone mass?

Enhanced Text: The actual answer to the following question is: Based on the given context, adults typically begin to lose bone mass around the age of 40. The key points are: - Bone mass reaches its peak during young adulthood, and then there is a slow but steady loss of bone beginning about age 40. - After about age 30, people can start to lose bone faster than their body makes it, which can weaken the bones and increase the risk of breakage. - The reduction of bone mass begins between ages 30 and 40, and continues to decline. So the summarized response is that adults normally begin to lose bone mass around the age of 40.

The context (which refers to text that was used to answer this question) is: '[Age. There's no way around it: loss of bone mass comes with age, laying the groundwork for low bone density and the potential of osteoporosis. We typically lose bone mass starting at age 40 and one in two women and one in four men over the age of 50 will fracture a bone at some point.', 'After about age 30, you can start to lose bone faster than your body makes it, which can weaken the bones and increase the risk of breakage. Some bone loss is natural as men and women age, but women are at higher risk of significant bone loss.', 'Bone mass reaches its peak during young adulthood. Then, after a period of stability, there is a slow but steady loss of bone beginning about age 40. In women, normal aging and menopause significantly increase susceptibility to osteoporosis.', 'In adults, this can take ten years. Until our mid-20s, bone density is still increasing. But at 35 bone loss begins as part of the natural ageing process. This becomes more rapid in post-menopausal women and can cause the bone-thinning condition osteoporosis.', 'The reduction of bone mass begins between ages 30 and 40, and continues to decline. Women lose about 8% of skeletal mass every decade, while men lose about 3%. Epiphyses, vertebrae, and the jaws lose more mass than other sites, resulting in fragile limbs, reduction in height, and loss of teeth.'].'

Answer Relevancy assesses how pertinent the actual answer is to the given context. It is measured between 0 and 1; where a lower score is given to answers that are incomplete or contain redundant

information; and a higher score indicates better relevancy. For the given question, context and answer, the answer relevancy score is: 0.9531866263993314.

Context Precision assesses how relevant is every context towards answering the question. Ideally all of the text in all of the contexts should be relevant to the question. It is measured between 0 and 1; where a lower score is given to lower precision contexts; and a higher score indicates more precision. For the given question and contexts, the context precision score is: 0.06666666666666667.

Context recall measures the extent to which the context aligns with the ground truth. It is computed based on attributing text in the ground truth to the context, and is measured between 0 and 1; where a lower score is given to lower recall contexts; and a higher score indicates better performance. For the given ground truth and contexts, the context recall score is: 0.2727272727272727

Faithfulness measures the factual consistency of the generated answer against the given context. It is considered faithful if all the claims that are made in the answer can be inferred from the given context. It is measured between 0 and 1; where a lower score is given to answers consisting of claims that are not in the context; and a higher score indicates that the answer is using information from the contexts. For the given question, context and answer, the faithfulness score is: 1.0.

Synthetic Data Prompt Generation

Prompt 1: Cloud Computation Sales and Marketing**Task:**

Generate a passage related to cloud computation sales and marketing and a corresponding question based on the passage.

Example:**Passage:**

Cloud computation has revolutionized sales and marketing by enabling businesses to analyze large datasets in real-time. This allows for more precise targeting of potential customers and more effective allocation of marketing resources. Companies can now leverage cloud-based tools to track consumer behavior, predict trends, and personalize marketing campaigns.

Question:

How has cloud computation changed the way businesses approach sales and marketing?

Prompt 2: Basketball**Task:**

Generate a passage related to basketball and a corresponding question based on the passage.

Example:**Passage:**

Basketball is a fast-paced sport that requires a combination of physical skill, strategic planning, and teamwork. Players must constantly

communicate and adapt to the changing dynamics on the court. Successful teams often have a mix of strong defense, effective offense, and the ability to capitalize on opponents’ mistakes.

Question:
What are the key components of a successful basketball team?

Prompt 3: Random Topics (Unrelated to Cloud Computation Sales and Marketing and Basketball)

Task:
Generate a passage on a random topic unrelated to cloud computation sales and marketing or basketball and a corresponding question based on the passage.

Example:
Passage:
The history of the automobile is marked by continuous innovation and technological advancements. From the invention of the internal combustion engine to the development of electric cars, the automotive industry has always been at the forefront of engineering and design. Modern cars are equipped with advanced safety features, autonomous driving capabilities, and environmentally friendly technologies.

Question:
How has the automotive industry evolved over the years?

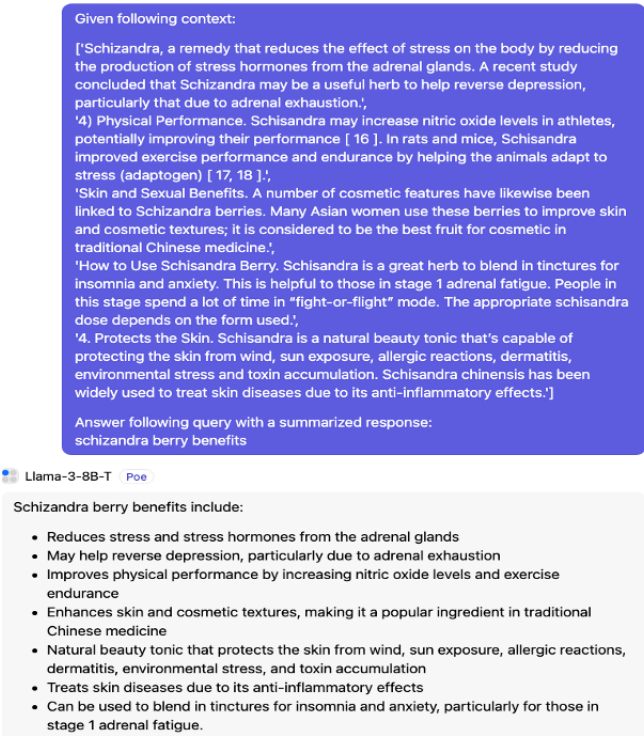


Figure 2: Example of Prompt of RAG Summarization with Retrieved Chunks

8.2 Unbiased Estimator

Let $f : X \rightarrow Y$ be a black box function mapping from the input space X (queries) to the output space Y (LLM-based metrics space). For each query x , the output is a vector $y = (y_1, y_2, y_3, y_4)$ where y_1, y_2, y_3 and y_4 represent Retrieval Precision, Retrieval Recall, Faithfulness, and Answer Relevance, respectively.

Assume a dataset $D = \{x_i\}_{i=1}^n$ where each x_i is a query. Associated with each query is a vector of metrics $m_i = (m_{i1}, m_{i2}, m_{i3}, m_{i4})$. For each metric k (where $k = 1, 2, 3, 4$ corresponding to the four metrics), the estimator $\hat{\theta}_k$ based on n observations is given by the sample mean:

$$\hat{\theta}_k = \frac{1}{n} \sum_{i=1}^n m_{ik}$$

Based on the bootstrap procedure described in 3.2, for each bootstrap sample D_b^* and for each metric k , compute the bootstrap replicate:

$$\hat{\theta}_k^* = \frac{1}{n} \sum_{i=1}^n m_{bik}^*$$

The bootstrap expectation for each metric k over all bootstrap samples is:

$$E^*[\hat{\theta}_k^*] = \frac{1}{B} \sum_{b=1}^B \hat{\theta}_k^*$$

We need to know that:

$$E^*[\hat{\theta}_k^*] = \hat{\theta}_k$$

to prove the bootstrap statistics $\hat{\theta}_k^*$ are unbiased estimators of the empirical means $\hat{\theta}_k$ for each metric.

Proof: Given that each $\hat{\theta}_k^*$ is an average of n independently and identically distributed (i.i.d.) bootstrap samples drawn with replacement from m_k , we apply the law of large numbers in the bootstrap world, stating:

$$E^*[\hat{\theta}_k^*] \approx \frac{1}{n} \sum_{i=1}^n E^*[m_{ik}^*]$$

Since the bootstrap samples are drawn from the empirical distribution of m_k , $E^*[m_{ik}^*] = \hat{\theta}_k$. Therefore:

$$E^*[\hat{\theta}_k^*] \approx \hat{\theta}_k$$

Thus, $\hat{\theta}_k^*$ is an unbiased estimator of $\hat{\theta}_k$ under the bootstrap distribution, assuming that the original sample is representative of the population and the metrics are i.i.d.

Received 10 June 2024; accepted 20 July 2024