

Domain-specific Question Answering with Hybrid Search

Dewang Sultania *, Zhaoyu Lu *, Twisha Naik *, Franck Dernoncourt, *
David Seunghyun Yoon *, Sanat Sharma, Trung Bui, Ashok Gupta,
Tushar Vatsa, Suhas Suresha, Ishita Verma, Vibha Belavadi, Cheng Chen, Michael Friedrich

Adobe Inc.
345 Park Avenue
San Jose, CA, 95110

Abstract

Domain-specific question answering is an evolving field that requires specialized solutions to address unique challenges. In this paper, we show that a hybrid approach—combining a fine-tuned dense retriever with keyword-based sparse search methods—significantly enhances performance. Our system leverages a linear combination of relevance signals, including cosine similarity from dense retrieval, BM25 scores, and URL host matching, each with tunable boost parameters. Experimental results indicate that this hybrid method outperforms our single-retriever system, achieving improved accuracy while maintaining robust contextual grounding. These findings suggest that integrating multiple retrieval methodologies with weighted scoring effectively addresses the complexities of domain-specific question answering in enterprise settings.

Introduction

With the increasing adoption of Large Language Models (LLMs) in enterprise settings, ensuring accurate and reliable question-answering systems remains a critical challenge. Building upon our previous work on domain-specific question answering about Adobe products (Sharma et al. 2024), which established a retrieval-aware framework with self-supervised training, we now present a production-ready, generalizable architecture alongside a comprehensive evaluation methodology. Our core contribution is a flexible, scalable framework built on Elasticsearch that can be adapted for any LLM-based question-answering system. This framework seamlessly integrates hybrid retrieval mechanisms, combining dense and sparse search with boost matching, while maintaining production-grade performance requirements. While we demonstrate its effectiveness using our organization’s domain-specific data, the architecture is designed to be domain-agnostic and can be readily deployed for other enterprise applications. We evaluate our system through a rigorous methodology that assesses performance across multiple dimensions: context relevance through normalized Discounted Cumulative Gain (nDCG), response

groundedness, answer accuracy, and answer null rate. To ensure robust testing, we compile a diverse evaluation dataset including:

1. Human-annotated responses for common feature queries
2. A carefully curated “negative” dataset containing jail-break attempts, NSFW content, and irrelevant queries to test system boundaries
3. LLM-based comparative analysis between system outputs and human-annotated ground truth

This comprehensive evaluation approach allows us to assess not only the system’s ability to provide accurate information but also its robustness against inappropriate queries and its capability to acknowledge when information is not available. Our hybrid search mechanism demonstrates significant improvements across these metrics compared to single-method approaches. The key contributions of this work include:

- A production-ready, generalizable framework for LLM-based QA systems built on Elasticsearch
- A flexible hybrid retrieval mechanism combining dense and sparse search methods
- A comprehensive evaluation framework for assessing QA system performance
- Empirical analysis demonstrating the effectiveness of our approach across various metrics

Through this work, we provide not only theoretical insights but also a practical, deployable solution for building reliable domain-specific question-answering systems that can be adapted to various enterprise needs.

Related Work

Our research builds upon recent advancements in retrieval-augmented language models, domain-specific question answering (QA), and hybrid retrieval methods. We review relevant literature across these key areas.

Retrieval-Augmented Language Models

Retrieval-Augmented Generation (RAG) models integrate information retrieval with language models to enhance per-

*These authors contributed equally.

formance on knowledge-intensive tasks such as question answering. These models employ a retriever to identify pertinent documents, conditioning the language model on the retrieved context during response generation. Lewis et al. (2020) introduced the RAG framework, demonstrating its effectiveness in open-domain QA by combining dense passage retrieval with sequence-to-sequence generation. Further, Lazaridou, Cancedda, and Baroni (2022) explored the application of RAG in an enterprise customer support setting, highlighting its potential for domain-specific QA.

Domain-Specific Question Answering

Developing QA systems tailored to specialized domains necessitates techniques that effectively capture domain-specific knowledge and terminology. Sharma et al. (2024) proposed a framework for compiling domain-specific QA datasets and introduced retrieval-aware fine-tuning of language models, which reduces hallucinations and enhances contextual grounding. Eppalapally et al. (2024) introduced a model for query rewriting to take the domain of interest into account. Additionally, hybrid retrieval methods that combine dense and sparse techniques have shown promise in domain-specific QA. For instance, Zhu et al. (2023) presented a hybrid text generation-based query expansion method, improving retrieval accuracy and QA system performance.

Hybrid Retrieval Methods

Integrating multiple retrieval signals can improve ranking performance compared to single-method approaches. Arivazhagan et al. (2023) explored hybrid hierarchical retrieval, combining sparse and dense methods in a two-stage document and passage retrieval setup, demonstrating improved in-domain and zero-shot generalization. Similarly, Li et al. (2021) proposed a dual reader-parser architecture that leverages both textual and tabular evidence, enhancing open-domain QA performance.

While prior work has made significant strides, developing reliable and robust domain-specific QA systems for enterprise settings remains an open challenge. Our work addresses this by proposing a flexible, generalizable framework that integrates hybrid retrieval with weighted relevance scoring. We extend prior hybrid approaches with a multi-phase scoring algorithm and introduce a comprehensive evaluation methodology to assess system performance across multiple dimensions.

Methodology

To optimize document retrieval accuracy, we employed a multi-phase scoring algorithm that evolved through iterations, starting from a baseline BM25 model, progressing to a chunking-based approach, and finally integrating host-based score adjustments.

Scoring Algorithm

Given a user query, we calculate relevance scores for each document as follows:

$$\begin{aligned} \text{score} = & \max(\text{matched chunks cosine score}) \\ & + \text{bm25_boost} \times \text{BM25 score} \\ & + \text{host_boost} \times \text{host_score} \end{aligned} \quad (1)$$

Here, the final score combines the highest cosine similarity score from matched content chunks within each document, boosted BM25 scores, and a host-based weighting. Each document’s chunked cosine score ensures finer-grained relevance, while BM25 and host boosts adjust for term frequency relevance and content source authority, respectively. This combination provides a more nuanced ranking by integrating multiple retrieval signals.

Parameter Selection

The values for `bm25_boost` and `host_boost` were empirically determined by optimizing for the highest average similarity score between generated outputs and a predefined golden set. This iterative tuning process ensured that the scoring weights effectively balanced relevance and content source reliability. Table 2 and Table- 3 summarize how we select these parameters.

Document Ranking

After calculating scores for all documents, they were sorted in descending order, and the top 3 documents were returned as the final output, prioritizing the documents most likely to yield relevant information.

Experiments and Evaluation

Datasets

We evaluate our system using two carefully curated datasets: a golden dataset for measuring accuracy and performance, and a negative dataset for testing system robustness.

Golden Dataset Our golden dataset comprises question-answer pairs from three primary documentation sources for our organization, with the following distribution:

- creativecloud.adobe.com (Express Learn): 62 pairs
- www.adobe.com: 74 pairs
- helpx.adobe.com: 51 pairs

Each entry in the dataset contains:

- A help query, initially LLM-generated and human-revised
- Relevant document URLs (separated by delimiters)
- Step-by-step answers, LLM-generated and human-revised
- Source documentation URL

Negative Dataset To evaluate system robustness, we compiled a negative dataset consisting of:

- 12 jailbreak attempts
- 6 NSFW queries
- 12 irrelevant queries

All queries were collected from internet sources and revised by human annotators, with an expected "content not found" response.

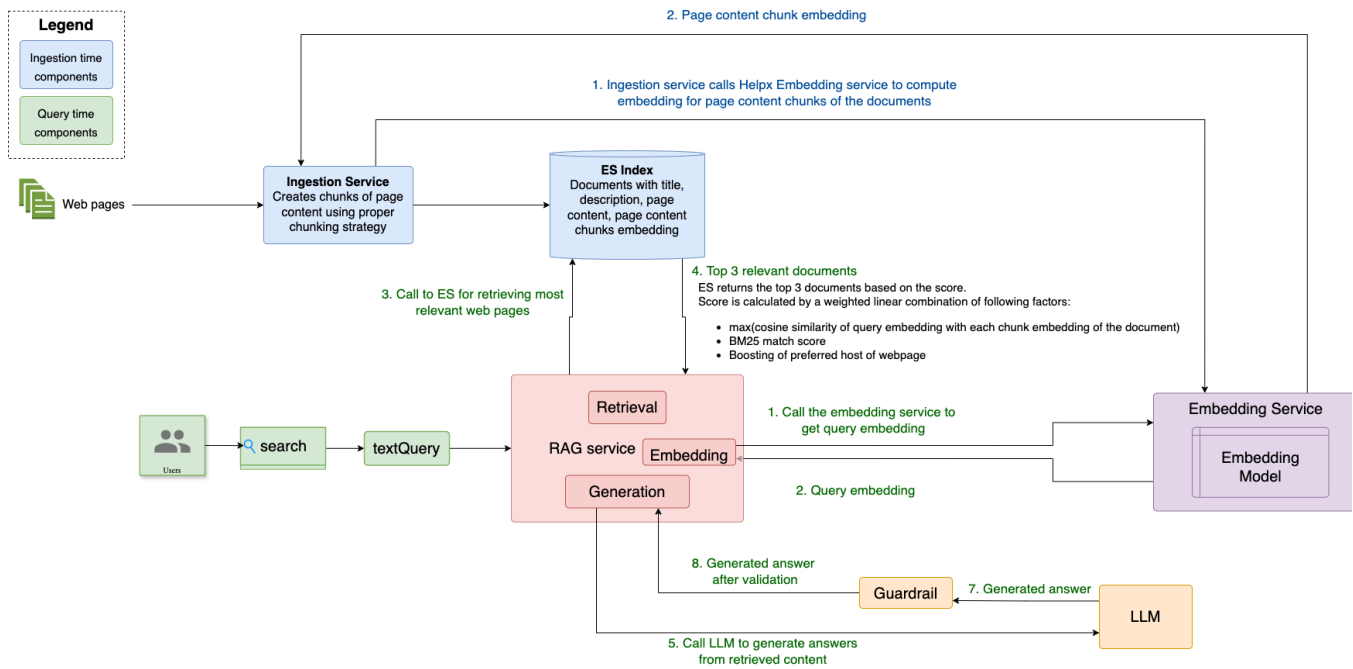


Figure 1: Production ready RAG deployment

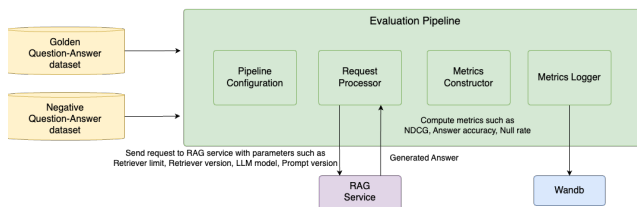


Figure 2: Evaluation Setup

Evaluation Setup

Our evaluation framework uses the following common parameters across all experiments:

- Generator Model: GPT-4o
- Retrieved URLs Limit: 3

Results and Analysis

Effect of Chunk Size on Retrieval Performance To determine the optimal chunk size for document segmentation, we experimented with various chunk sizes and overlap values. Table 1 summarizes the NDCG scores obtained for different configurations.

Chunk size	Chunk Overlap	Context nDCG
1000	100	0.828
2000	500	0.802
5000	1000	0.795

Table 1: Effect of Approximate Chunk Size on Retrieval Performance

The results indicate that an approximate chunk size of 1000 characters with a split ensuring it ends on a sentence delimiter, with an overlap of 100 characters provides a favorable balance between retrieval granularity and computational efficiency.

Boost Parameter Tuning We performed empirical tuning of the boost parameters to optimize the scoring algorithm for document retrieval.

BM25 Boost Tuning Using 60% of the golden dataset as a validation set, we adjusted the BM25 content match boost. Table 2 shows the Top-3 NDCG scores for different boost values.

bm25-boost	Context nDCG
0.1	0.863
0.3	0.868
0.6	0.840
1	0.831

Table 2: Choosing optimal parameter for BM25

The optimal BM25 content boost was found to be 0.3, achieving the highest NDCG score.

Host Boost Tuning We also tuned the host boost parameter to enhance retrieval from authoritative sources. Table 3 presents the NDCG scores for varying host boost values.

host-boost	Context nDCG
0.1	0.853
0.3	0.842
0.6	0.830
1	0.781

Table 3: Choosing optimal parameter for BM25

A host boost value of 0.1 provided the best trade-off between relevance and authority.

Discussion of Boost Tuning Results The tuning of boost parameters significantly impacted retrieval performance. Higher boost values for BM25 content match improved NDCG up to a point, after which performance declined due to overemphasis on term frequency. Similarly, moderate host boosting enhanced the retrieval of authoritative documents without overshadowing relevance.

Retrieval Strategy We evaluated five increasingly sophisticated retrieval strategies to assess their effectiveness in our domain-specific question-answering system. The strategies are:

- **Keyword Search (BM25):** This baseline approach utilizes the BM25 algorithm for sparse keyword-based retrieval. It serves as a foundational benchmark for comparison.
- **OpenAI text-embedding-3-large:** This strategy employs OpenAI’s pre-trained text-embedding-3-large model for dense retrieval. It leverages semantic embeddings to capture the meaning of queries and documents without domain-specific fine-tuning.
- **Fine-tuned Retriever:** This approach uses a dense retriever that has been fine-tuned on our domain-specific dataset, as discussed briefly in the Methods section. The fine-tuning enhances the model’s ability to understand domain-specific terminology and nuances.
- **Fine-tuned Retriever + Keyword Search:** This hybrid method combines our fine-tuned dense retriever with the BM25 keyword search. The integration aims to leverage both semantic understanding and exact keyword matching to improve retrieval performance.
- **Fine-tuned Retriever + Keyword Search + Host Boost** Building upon the hybrid approach, this strategy incorporates URL host boosting. By assigning additional weight to documents from preferred URL hosts, we aim to enhance the relevance of the retrieved results in the enterprise context.

Strategy	Context NDCG
Keyword Search (BM25)	0.640
text-embedding-3-large (openai)	0.760
Fine tuned retriever	0.828
Fine tuned retriever + Keyword Search	0.845
Fine tuned retriever + Keyword Search + Host boosting	0.847

Table 4: Detailed Performance Comparison of Retrieval Strategies

Guardrails around generated answer In order to protect any jailbreak attempts, we have implemented a guardrail mechanism on top of answer generation. Before returning the generated answer to the user, we compare the similarity of the generated answer with the system prompt, excluding the user query. If this similarity is very high, we do not return any content to the user. We have added jailbreak attempt queries in the evaluation dataset as well, which proves that we have made the system robust with this additional check.

Generated Answer Analysis To validate our hypothesis that a better retriever leads to better final responses, we conducted an analysis comparing the generated answers from the LLM to the golden answers using an LLM as an evaluator.

Methodology Using GPT-4 as the evaluation model, we assessed the quality of the generated answers across different retrieval strategies. We focused on two key metrics:

Answer Similarity: Measures the degree of correspondence between the generated answer and the golden answer. **Groundedness:** Evaluates how well the generated answer is supported by the retrieved documents, ensuring factual correctness and context relevance. For each query in the golden dataset, we generated answers using the LLM with contexts retrieved by each strategy. We then prompted GPT-4 to rate the similarity and groundedness of each generated answer compared to the golden answer on a scale from 0 to 1.

Results Table 5 presents the mean and std for answer similarity and groundedness across the different retrieval strategies.

The results indicate that strategies incorporating both the fine-tuned retriever and keyword search, especially with host boosting, achieved the highest scores in both answer similarity and groundedness.

Analysis The incremental improvements in both answer similarity and groundedness scores align with our hypothesis that enhanced retrieval strategies lead to better final responses. Specifically:

Strategy		Answer Similarity mean	Answer Similarity Std	Groundedness mean	Groundedness Std
Keyword Search (BM25)		0.717	0.27	0.919	0.27
text-embedding-3-large (openai)		0.748	0.23	0.979	0.13
Fine tuned retriever		0.755	0.22	0.974	0.15
Fine tuned retriever + Keyword Search		0.767	0.20	0.983	0.12
Fine tuned retriever + Keyword Search + Host boosting		0.780	0.19	0.983	0.12

Table 5: Performance Comparison of Final answer with different retrieval strategies

- **Fine-tuned Retriever:** By capturing domain-specific semantics, it provides more relevant context to the LLM, improving answer quality.
- **Hybrid Approach:** Combining dense retrieval with keyword search leverages both semantic understanding and exact term matching, further enhancing the context.
- **Host Boosting:** Prioritizing authoritative sources increases the reliability of the information provided, which is reflected in higher groundedness scores.

These findings demonstrate that optimizing retrieval not only improves retrieval metrics but also has a significant positive impact on the quality of the final answers generated by the LLM.

Discussion

Our comprehensive evaluation of the hybrid search framework reveals several key findings and implications for domain-specific question answering systems:

Performance Advantages of Hybrid Retrieval The experimental results demonstrate that our hybrid approach, combining fine-tuned dense retrieval with sparse search and host boosting, consistently outperforms single-method approaches across multiple metrics. This superior performance can be attributed to several factors:

- **Complementary Retrieval Signals:** The integration of dense and sparse retrieval methods allows the system to leverage both semantic understanding and exact keyword matching. While the fine-tuned dense retriever captures domain-specific context and terminology, the BM25-based keyword search ensures critical term matches aren't overlooked.
- **Host Boosting Effect:** The addition of host-based boosting further refined results by prioritizing authoritative sources, leading to a modest but consistent improvement in both retrieval accuracy (NDCG: 0.84701) and answer quality metrics (Answer Similarity: 0.78, Groundedness: 0.983).

Impact on Answer Generation The improvement in retrieval quality shows a direct correlation with enhanced answer generation:

- **Answer Similarity:** The progression from basic keyword search (0.717) to the full hybrid approach (0.78) demonstrates that better context retrieval leads to more accurate answers.
- **Groundedness:** High groundedness scores across all methods indicate that the system maintains strong factual alignment, with the hybrid approach achieving the highest score (0.983).

System Robustness Our evaluation framework's inclusion of negative test cases (jailbreak attempts, NSFW queries, and irrelevant queries) demonstrates the system's resilience to potential misuse. The implemented guardrail mechanism, which checks answer similarity against system prompts, provides an effective defense against jailbreak attempts while maintaining system usability.

Practical Implications The findings have several important implications for enterprise deployments:

- **Scalability:** The framework's architecture, built on Elasticsearch, provides a production-ready solution that can handle enterprise-scale deployments while maintaining performance.
- **Adaptability:** The tunable boost parameters (BM25_boost and host_boost) offer flexibility in adjusting the system for different domain-specific applications and content sources.
- **Cost-Effectiveness:** The chunking optimization (optimal size: 1000 tokens with 100 token overlap) provides an effective balance between retrieval accuracy and computational efficiency.

These findings provide a foundation for future research in domain-specific question answering systems while offering practical insights for enterprise implementations.

References

- Arivazhagan, M. G.; Liu, L.; Qi, P.; Chen, X.; Wang, W. Y.; and Huang, Z. 2023. Hybrid Hierarchical Retrieval for Open-Domain Question Answering. In *Findings of the Association for Computational Linguistics: ACL 2023*, 10680–10689.
- Eppalapally, S.; Dangi, D.; Bhat, C.; Gupta, A.; Zhang, R.; Agarwal, S.; Bagga, K.; Yoon, S.; Lipka, N.; Rossi, R.;

and Dernoncourt, F. 2024. KaPQA: Knowledge-Augmented Product Question-Answering. In Yu, W.; Shi, W.; Yasunaga, M.; Jiang, M.; Zhu, C.; Hajishirzi, H.; Zettlemoyer, L.; and Zhang, Z., eds., *Proceedings of the 3rd Workshop on Knowledge Augmented Methods for NLP*, 15–29. Bangkok, Thailand: Association for Computational Linguistics.

Lazaridou, A.; Cancedda, N.; and Baroni, M. 2022. Internet-Augmented Language Models through Few-Shot Prompting for Open-Domain Question Answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 437–456.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474.

Li, A. H.; Ng, P.; Xu, P.; Zhu, H.; Wang, Z.; and Xiang, B. 2021. Dual Reader-Parser on Hybrid Textual and Tabular Evidence for Open Domain Question Answering. arXiv:2108.02866.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Reimers, N. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Sharma, S.; Yoon, D. S.; Dernoncourt, F.; Sultania, D.; Bagga, K.; Zhang, M.; Bui, T.; and Kotte, V. 2024. Retrieval Augmented Generation for Domain-specific Question Answering. arXiv:2404.14760.

Truera. 2023. Trulens. <https://github.com/truera/trulens>.

Zhu, W.; Zhang, X.; Zhai, Q.; and Liu, C. 2023. A Hybrid Text Generation-Based Query Expansion Method for Open-Domain Question Answering. *Future Internet*, 15(5): 180.

Model Fine-tuning Details

Retriever Model Configuration

We build a text encoder that computes a vector representation of a sentence. The model is trained to map user queries and corresponding documents in similar latent spaces so that it can be used as a semantic retriever (Reimers 2019).

Training Process

We use user behavior data to train the retriever. The behavior data consists of a “user query” and a list of “documents” that were shown to the user by the in-house system. Among the list of “documents,” we have a signal that a user clicked on one of the provided documents. Since each document consists of a “title” and “body text,” we formulate the training loss as follows:

$$L_{\text{total}} := L_1(f_{\theta}(q), f_{\theta}(\text{title})) + L_1(f_{\theta}(q), f_{\theta}(\text{body text})),$$

where q is the user query, $f_{\theta}()$ is the sentence encoder parameterized by θ , and L_1 represents the InfoNCE loss function that operates between images and text (Oord, Li, and Vinyals 2018).

Limitations

While our hybrid search approach demonstrates strong performance for domain-specific question answering, several important limitations should be noted:

Context Awareness Limitations

The current system lacks crucial contextual awareness in several dimensions:

1. **User Location Context:** The system cannot differentiate between users on different pages or sections of the product (e.g., homepage versus editor), leading to potentially misaligned responses. This becomes particularly evident in answers that reference generic locations like “open homepage” without considering the user’s current position in the product journey.
2. **Platform and Device Context:** The system does not receive or account for information about:
 - Operating system platform (Windows, iOS, etc.)
 - Device type (mobile phone, desktop computer, tablet)
 - Screen size or resolution

This limitation can result in instructions that may not be applicable to the user’s specific platform or device configuration.

Language Support

The current implementation is restricted to English-language queries and content. This represents a significant limitation for global enterprise deployment, particularly for multinational organizations requiring multilingual support. The system would need substantial modifications to:

- Handle queries in multiple languages
- Process multilingual documentation
- Generate responses in the user’s preferred language
- Account for cultural and regional differences in terminology

Multimodal Content Handling

A notable limitation exists in the system’s ability to process and leverage visual content:

1. **Visual Context Loss:** Many retrieved documents contain valuable screenshots, diagrams, or step-by-step visual guides that demonstrate product functionality. However, these visual elements are currently not:
 - Included in the context provided to the LLM
 - Used to enhance response generation
 - Incorporated into the scoring mechanism
2. **Response Limitations:** The system cannot:
 - Reference specific visual elements in responses

- Generate or modify visual content
- Provide visual feedback or confirmation

This limitation is particularly significant in software documentation, where visual guides often provide crucial information that text alone cannot effectively convey.

Future Work

Addressing these limitations suggests several promising directions for future research:

1. **Enhanced Context Integration:** Developing mechanisms to capture and utilize user context, including:
 - Real-time user location tracking within the product
 - Platform and device detection
 - Context-aware response generation
2. **Multilingual Extension:** Expanding the system to support multiple languages through:
 - Multilingual dense retriever training
 - Cross-lingual document retrieval
 - Language-specific response generation
3. **Multimodal Enhancement:** Incorporating visual content through:
 - Multimodal embedding models
 - Visual content understanding and extraction
 - Integration of visual elements in response generation
 - Screenshot and image-aware context processing

These improvements would significantly enhance the system's utility and applicability across diverse enterprise environments.

Evaluation Metrics Details

Normalized Discounted Cumulative Gain (nDCG)

nDCG is a standard metric in information retrieval that measures ranking quality. It is calculated as:

$$\text{nDCG@k} = \frac{\text{DCG@k}}{\text{IDCG@k}} \quad (2)$$

where DCG@k is defined as:

$$\text{DCG@k} = \sum_{i=1}^k \frac{2^{\text{rel}_i} - 1}{\log_2(i + 1)} \quad (3)$$

Here:

- rel_i is the relevance score of the result at position i
- IDCG@k is the DCG@k value for the ideal ranking
- The score ranges from 0 to 1, with 1 being perfect ranking

Interpretation of nDCG Scores

This metric is particularly suitable for our evaluation because:

- It accounts for both relevance and position in the ranking
- It is normalized, allowing comparison across queries

LLM-based Evaluation Metrics

In addition to nDCG, we employ GPT as an automated judge to evaluate two key aspects of the generated responses: groundedness and accuracy.

Groundedness Score We evaluate how well the generated responses are grounded in the retrieved context using the following prompt structure (Truera 2023):

System: You are an impartial groundedness judge. You will be given a context and a response. Your task is to determine how grounded the response is in the given context. A response is considered grounded if it is supported by and does not contradict the given context.
Rate the groundedness on a scale from 0 to 1 (0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0), where 0 is completely ungrounded and 1 is perfectly grounded.
Context: \$context
Response: \$response
Groundedness Score:

Accuracy Score We evaluate the accuracy of generated responses by comparing them with ground truth answers using the following prompt (Truera 2023):

System: You will be given one Model_answer and a Groundtruth_answer for a question about \$product. Your task is to rate the similarity of the two answers on one metric.
Evaluation Criteria:
Similarity (0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0)
- Similarity of overall text and similarity of each step or multiple steps also need to be considered.
Question: \$question
Groundtruth_answer: \$ground_truth
Model_answer: \$model_answer
Evaluation Steps: 1. Read the question carefully 2. Do not modify the Groundtruth_answer and Model_answer 3. Read the Groundtruth_answer and identify the steps and order 4. Read the Model_answer and assess similarity to Groundtruth 5. Assign similarity score (0.0 to 1.0) 6. Reply with the Similarity score only

Both metrics output scores in the range [0,1], where:

- Groundedness Score measures how well the response is supported by the retrieved context
- Accuracy Score measures the semantic and structural similarity between the generated response and the ground truth

This combination of retrieval-based (nDCG) and generation-based (Groundedness, Accuracy) metrics provides a comprehensive evaluation framework for our system's performance.