Proceedings of MSS 2025 Symposium: Automation and Smart Technologies in Petroleum Engineering
DOI: 10.54254/2755-2721/142/2025.KL22312

Retrieval-Augmented Generation: Methods, Applications and Challenges

Yicheng Liu

Faculty of Science, National University of Singapore, Singapore
e0952418@u.nus.edu

Abstract: The Retrieval-Augmented Generation (RAG) has been proven to have a promising approach. It can address the limitations of purely generative models in knowledge-intensive tasks caused by their reliance on static, pre-trained knowledge. RAG addresses these challenges by integrating a retrieval mechanism with a generative model, enabling dynamic access to external knowledge sources during the generation process. This paper presents a comprehensive study of the RAG framework, focusing on its architecture, training strategies, and applications. The framework combines a dense passage retriever (DPR) with a sequence-to-sequence generator (GPT-3.5-turbo), jointly optimized in an end-to-end manner to retrieve and utilize relevant knowledge effectively. This paper evaluates RAG on MS MARCO, demonstrating its superiority over state-of-the-art purely generative models and traditional retrieval-based systems. Experimental results show that RAG achieves significant improvements in factual accuracy, relevance, and interpretability, as measured by metrics such as term frequency–inverse document frequency, bidirectional encoder representation from transformer Score, and Q-Bilingual Evaluation Understudy-1.

Keywords: Retrieval-Augmented Generation, Dense Passage Retrieval, Large-scale language Model, Term Frequency-Inverse Document Frequency, Bidirectional Encoder Representations from Transformers

1. Introduction

Pre-trained large-scale language models (LLMs) such as ChatGPT, Gemini, and DeepSeek have been proven to have great abilities to acquire extensive and profound knowledge from the dataset. They can function as a parameterized implicit knowledge base without the need for external memory, relying on the vast amount of pre-trained knowledge stored within their parameters [1]. Despite these advantages, LLMs may produce a phenomenon called “hallucinations”, especially for time-sensitive queries due to their reliance on static knowledge bases. Furthermore, the LLMs are a black box, making it harder to verify the reliability of the outputs [2].

To address these problems, retrieval-augmented generation (RAG) has emerged as a promising paradigm that combines the advantages of retrieval-based methods and generative models. RAG integrates a retrieval mechanism, dynamically extracting relevant information from external knowledge sources. By leveraging external knowledge, RAG not only enhances the factual accuracy and timeliness of generated text but also improves interpretability by providing explicit references to the retrieved documents.

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/142/2025.KL22312

Despite its potential, the development and deployment of RAG systems face several challenges. First, the retrieval process must be both efficient and accurate, ensuring that the most relevant documents are selected from potentially massive knowledge corpora. Second, the generative model must effectively integrate retrieved information into its responses, balancing the need for factual correctness with the ability to generate fluent and contextually appropriate text. Finally, the end-to-end training of RAG systems requires careful optimization to ensure that both the retriever and generator components work collaboratively.

This paper tests a RAG framework that dynamically integrates external knowledge retrieval with text generation. The approach leverages a dense passage retriever and a sequence-to-sequence generator, jointly optimized in an end-to-end manner to enhance the accuracy and relevance of generated responses. Through extensive experiments on benchmarks such as MS MARCO, this paper demonstrates the effectiveness of RAG in improving factual correctness and interpretability for knowledge-intensive tasks.

2. Related Work

The development of RAG builds upon advancements in two key areas: (1) generative models that produce fluent and contextually relevant text, and (2) retrieval-based systems that extract precise information from external knowledge sources. Below, this paper reviews the relevant literature in these domains and highlights the evolution of hybrid approaches that combine retrieval and generation.

2.1. Generative Models

Generative models have achieved remarkable success in natural language processing (NLP) tasks, such as text generation, summarization, and dialogue systems. However, they face limitations in handling knowledge-intensive tasks due to their reliance on static, pre-trained knowledge and their tendency to generate factually incorrect or outdated information, a phenomenon known as "hallucination."

To address these limitations, sequence-to-sequence models like T5 and BART have been proposed. While they exhibit improved flexibility and performance, they still struggle with dynamically incorporating external knowledge, especially for queries requiring up-to-date or domain-specific information.

2.2. Retrieval-Based Systems

Retrieval-based systems aim to provide accurate and interpretable answers by extracting relevant information from external knowledge sources. Traditional approaches, such as BM25 and TF-IDF, rely on term-based matching to retrieve documents or passages. While effective for simple queries, these methods often fail to capture semantic relationships between the query and the retrieved content.

More recent approaches, such as Dense Passage Retrieval (DPR), leverage dense vector representations to improve retrieval accuracy. This approach has demonstrated significant improvements in tasks like open-domain question answering (QA). However, retrieval-based systems are inherently limited by their inability to generate novel or synthesized responses, as they can only return existing content from the knowledge corpus.

2.3. Hybrid Approaches: Combining Retrieval and Generation

The limitations of purely generative and retrieval-based systems have motivated the development of hybrid approaches that combine the strengths of both paradigms. Early attempts, such as REALM

DOI: 10.54254/2755-2721/142/2025.KL22312

and ORQA, introduced retrieval-augmented pre-training to jointly optimize retrieval and generation tasks [3-5]. These models demonstrated the potential of integrating external knowledge into generative models but were limited by their two-stage training pipelines and computational inefficiency.

This paper's work builds on these advancements by proposing RAG, a framework that seamlessly integrates retrieval and generation in an end-to-end manner. Unlike previous approaches, RAG jointly optimizes the retriever and generator components, allowing the model to dynamically retrieve and utilize relevant knowledge while maintaining the flexibility and fluency of generative models.

3. Methods

In this section, this paper presents the architecture and implementation details of RAG framework, which integrates a retrieval mechanism with a generative model to dynamically access external knowledge and generate contextually appropriate responses. The evaluation focuses on three key aspects:

This paper first assesses the performance of RAG in comparison to purely generative models and traditional retrieval-based systems, aiming to highlight the advantages of combining retrieval and generation capabilities. Then this paper tests the effectiveness of RAG in handling knowledge-intensive tasks, such as open-domain question answering and dialogue generation, to demonstrate its ability to leverage external knowledge for improved accuracy and relevance. Finally, this paper analyzes the key factors that contribute to RAG's performance, including the interaction between the retriever and generator components, to provide insights into the framework's strengths and potential areas for improvement. In this paper, the RAG-Sequence as the RAG method.

3.1. Models

The framework consists of two main components, a retriever that retrieves the most similar paragraph from an external knowledge base and a generator that combines the retrieved paragraph with the question query and generates a fluent response.

For each question query, the model uses the corresponding paragraph to generate the complete sequence. For a given input query q, the retriever retrieves a set of relevant passages $P = \{p_1, p_2, \cdots, p_k\}$ from a large-scale knowledge corpus. Then the generator takes the query q and the retrieved passage P as input, and generates an output response r.

The retriever and generator are jointly optimized in an end-to-end manner, allowing the model to learn how to effectively retrieve and utilize external knowledge.

3.2. Retriever: Dense Passage Retrieval (DPR)

The retriever identifies and retrieves the most relevant document from a knowledge corpus. This paper employs a DPR based on a dual-encoder architecture [6]:

For query encoder, this paper encodes the input query q into a dense vector q, and uses BERT as the transformer-based encoder. For passage encoder, this paper encodes each passage p_i into a dense vector p_i using the same separate transformer-based encoder as the query encoder. For similarity scoring, this paper computes the similarity of a passage p_i between the query q as the dot product of their embeddings:

$$\text{score}(q, p_i) = q^T p_i \tag{1}$$

DOI: 10.54254/2755-2721/142/2025.KL22312

3.3. Generator: GPT-3.5-Turbo

The generator is responsible for synthesizing the retrieved documents into a coherent response. This paper applies GPT-3.5-turbo, a sequence-to-sequence (Seq2Seq) architecture based on a pre-trained language model [7]. This paper concatenates the question query q and retrieved passage, together. GPT-3.5-turbo was pre-trained with a denoising objective and a variety of noise functions. It has achieved state-of-the-art results across a range of generation tasks and outperforms models of similar size.

4. Experiments

To evaluate the performance of RAG, this paper experiments with a wide range of knowledge-intensive questions. For the experiments, this paper uses the related websites from the dataset. Each website is split into disjoint 100-word chunks. During training, this paper retrieves the most similar chunk for each query. This paper now discusses experimental details for each task.

4.1. Open-domain Question Answering

Open-domain question answering (OpenQA) is a key benchmark for evaluating the ability of models to retrieve and generate accurate answers from large-scale knowledge sources. In this study, this paper evaluates the RAG framework on the MS MARCO dataset, a widely used benchmark for OpenQA tasks. Below, this paper describes the dataset, experimental setup, baseline models, and results.

MS MARCO (Microsoft Machine Reading Comprehension) is a large-scale dataset designed for question answering and passage-ranking tasks [8]. It consists of real-world user queries sourced from Bing search logs, paired with human-generated answers and relevant passages from web documents.

For the experiments, this paper focuses on the question-answering task, where the goal is to generate concise and accurate answers using the provided passages.

4.2. Evaluation

To comprehensively evaluate the performance of the RAG framework, this paper employs a combination of traditional and task-specific metrics. this paper introduces TF-IDF cosine similarity, BERT-based cosine similarity, and Q-BLEU-1 to assess the quality of the generated responses. Below, details will be described:

TF-IDF (Term Frequency-Inverse Document Frequency) is a classic information retrieval technique that measures the importance of a word in a document relative to a corpus [9]. It is calculated as:

$$\mathrm{TF-IDF}(t, d) = \mathrm{TF}(t, d) \times \mathrm{IDF}(t) \tag{2}$$

This paper uses TF-IDF to compute the cosine similarity between the generated response and the reference answer, providing a measure of their lexical overlap and relevance. Both the generated response and the reference answer are represented as TF-IDF vectors, where each dimension corresponds to a unique word in the corpus, weighted by its TF-IDF score. The similarity between the two vectors is computed as the cosine of the angle between them:

$$\mathrm{similarity}_{\mathrm{TF-IDF}} = \frac{v_{gen} \cdot v_{ref}}{\|v_{gen}\| \|v_{ref}\|} \tag{3}$$

Where v_{gen} is the TF-IDF vector for generated response and v_{ref} is the TF-IDF vector for reference answer.

DOI: 10.54254/2755-2721/142/2025.KL22312

BERT-based cosine similarity leverages contextualized embeddings to measure the semantic similarity between the generated response and the reference answer [10]. The generated response and reference answer are encoded into dense vector representations using a BERT model. The similarity between the two embeddings is computed as:

$$\text{similarity}_{\text{BERT}} = \frac{e_{gen} \cdot e_{ref}}{\|e_{gen}\| \|e_{ref}\|} \tag{4}$$

Where e_{gen} is the BERT embedding for generated response and e_{ref} is the BERT embedding for reference answer.

BLEU-1 is a variant of the traditional BLEU (Bilingual Evaluation Understudy) metric, adapted for question-answering and knowledge-intensive tasks [11]. Unlike standard BLEU, Q-BLEU-1 focuses on n-gram overlap between generated and reference texts. Q-BLEU-n incorporates a question-aware weighting mechanism to better align with the semantic relevance of the generated content. Specifically: Unigram Precision: Q-BLEU-1 calculates the precision of unigrams (single words) in the generated response that match the reference answer, weighted by their relevance to the input question. Question-Answer Alignment: The metric assigns higher weights to unigrams that are semantically aligned with the question, ensuring that the evaluation focuses on the most contextually important aspects of the response.

The Q-BLEU-1 score is computed as:

$$\text{Q} - \text{BLEU} - 1 = \frac{\Sigma_{\omega \in \text{response}} \text{weight}(\omega) \cdot \text{match}(\omega, \text{reference})}{\Sigma_{\omega \in \text{response}} \text{weight}(\omega)} \tag{5}$$

where $\text{weight}(\omega)$ is the question-aware weight of word ω, and $\text{match}(\omega, \text{reference})$ is a binary indicator of whether ω appears in the reference answer.

The combination of TF-IDF, BERT Score, and Q-BLEU-1 provides a multi-faceted evaluation framework:

TF-IDF ensures that the retriever provides high-quality input to the generator. BERT Score evaluates the semantic quality and fluency of the generated responses. R-BLEU-1 provides a lightweight measure of lexical overlap, ensuring that the generated text aligns with the reference at a surface level.

The differences among the three metrics are shown in Table 1.

Table 1: Role in Evaluation and Complementary Value of Metrics

Metric	Role in Evaluation	Complementary Value
TF-IDF	Assess retrieval relevance	Ensures high-quality input for generation
BERT	Evaluate semantic quality of generation	Captures deep semantic understanding
Q-BLEU-1	Measure lexical overlap	Provides efficient surface-level evaluation

Together, these metrics address the key aspects of RAG's performance, including retrieval relevance, semantic accuracy, and lexical alignment, while mitigating the limitations of individual metrics.

5. Results

In this section, this paper presents the results of the experiments comparing the performance of the plain (non-RAG) model and the RAG model across three similarity metrics: BERT-based semantic similarity, BLEU score, and TF-IDF-based lexical importance matching. The analysis focuses on the

DOI: 10.54254/2755-2721/142/2025.KL22312

ability of both models to generate answers that are semantically and lexically aligned with the ground truth.

5.1. Overall Performance

In this part, this paper applies three metrics mentioned above to evaluate the performance of RAG, comparing with the plain method. For each metric, this paper use a figure to compare the scores between plain and RAG method. In the figure, the coordinate of the point is (plain score, RAG score). Then y=x is drawn in the figure to compare the scores.

5.1.1. TF-IDF Similarity Scores

The TFIDF similarity scores exhibit more variation compared to BERT and BLEU. As seen in Figure 1, there are some differences in similarity between the "plain" and "RAG" methods. While both methods show relatively low similarity values, there are a few instances where the "RAG" method yields higher scores, such as in sample 4, where the "plain" score is 0.000, but the "RAG" score rises to 0.0427. These results suggest that the TF-IDF method, which is based on term frequency and inverse document frequency, might benefit slightly from the RAG process, although the overall effect is still limited.

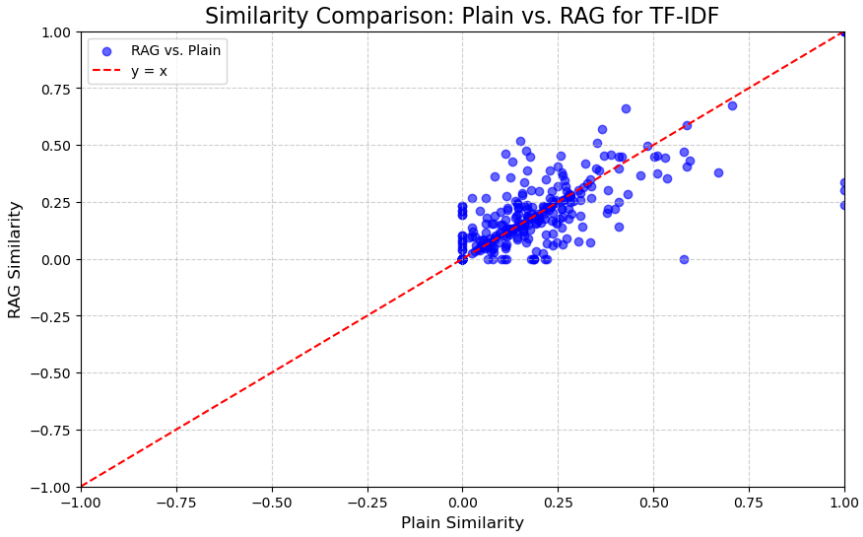

Figure 1: Similarity Comparison for TF-IDF (Photo/Picture credit: Original).

5.1.2. BERT Similarity Scores

The BERT-based similarity scores, which measure the semantic alignment between generated answers and ground truth, range from -1 (completely dissimilar) to 1 (completely similar). The majority of scores for both models fall within the range of 0.7 to 1, indicating that most generated answers are semantically close to the ground truth. However, there are instances of negative values (e.g., plain: -0.55374306, RAG: -0.31060708), suggesting that both models occasionally generate answers that are completely unrelated to the correct answers (Figure 2).

DOI: 10.54254/2755-2721/142/2025.KL22312

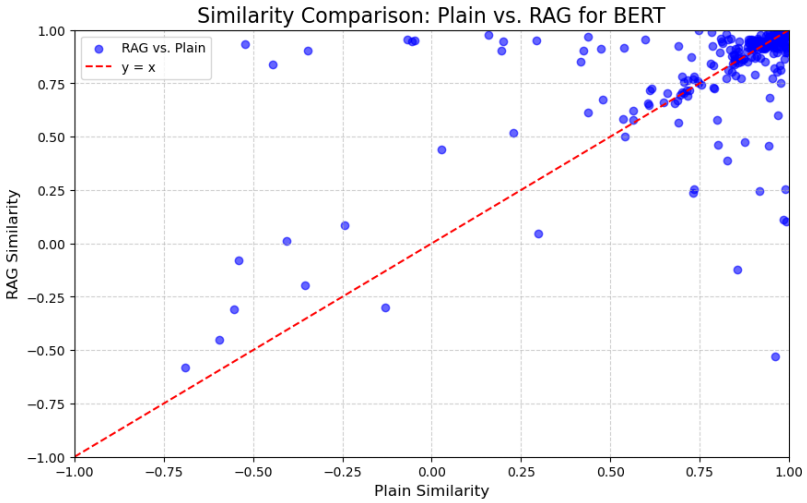

Figure 2: Similarity Comparison for BERT (Photo/Picture credit: Original).

5.1.3. Q-BLEU-1 Similarity Scores

The BLEU scores, however, show a different pattern. BLEU's similarity scores are mostly close to zero for both the "plain" and "RAG" methods (in Figure 3), indicating that this metric does not capture significant similarity between the compared texts in the majority of the samples. But for non-zero points, in most cases (about 87% cases), RAG achieves higher BLEU scores than the plain model.

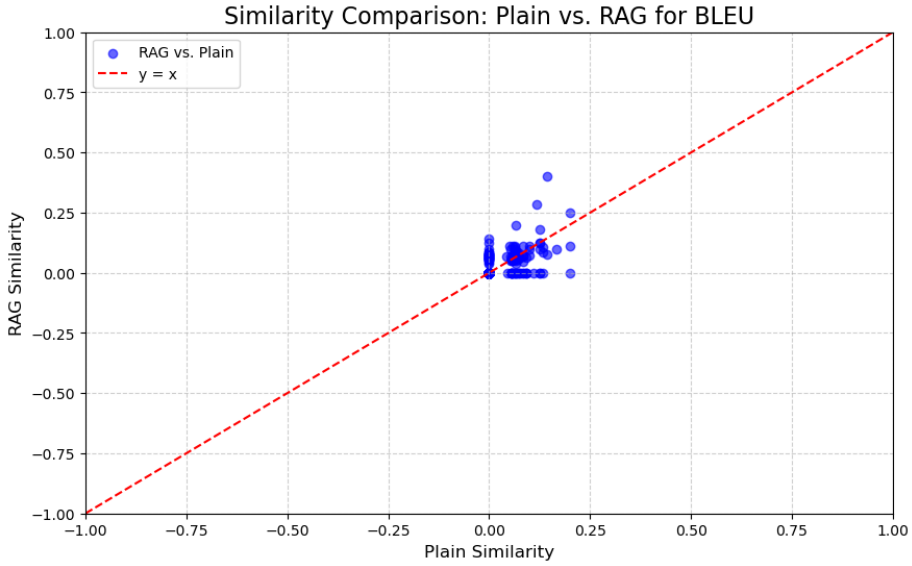

Figure 3: Similarity Comparison for Q-BLEU-1 (Photo/Picture credit: Original).

5.2. Statistical Analysis

To quantify the performance differences, this paper computed the mean, and standard deviation of the similarity scores and the rate of cases RAG outperforms the plain method for the RAG model. To compare the performance of RAG method under different metrics, this paper shows the result in Table 2.

DOI: 10.54254/2755-2721/142/2025.KL22312

Table 2: Mean, Std, and Rate of cases RAG Outperforms Plain Method

	Mean for RAG	Std Dev for RAG	Rate of cases RAG outperforms plain method
TF-IDF	0.1788	0.1654	63%
BERT	0.8381	0.2608	54%
Q-BLEU-1	0.0254	0.0484	86%

6. Challenges & Future Work

6.1. Current Challenges

6.1.1. Retrieval Quality and Efficiency

The performance of RAG heavily depends on the quality of the retrieved documents. Inaccurate or irrelevant retrievals can lead to suboptimal or incorrect generation results. Additionally, retrieving from large-scale knowledge corpora (e.g., Wikipedia) can be computationally expensive, especially for real-time applications [12].

Future work could explore more advanced retrieval techniques, such as hierarchical retrieval or active retrieval strategies, to improve both accuracy and efficiency. Additionally, integrating multi-modal retrieval (e.g., combining text, images, and structured data) [13,14] could enhance the relevance of retrieved content.

6.1.2. Multi-Hop Reasoning

Many knowledge-intensive tasks, such as complex question answering, require multi-hop reasoning—combining information from multiple documents to derive the correct answer. Current RAG systems often struggle with such tasks due to limitations in their ability to reason across multiple retrieved documents [15].

Developing models that explicitly support multi-hop reasoning, such as graph-based retrieval or iterative retrieval-generation pipelines, could address this challenge. Additionally, incorporating reinforcement learning to optimize retrieval and generation for multi-hop tasks may improve performance.

6.1.3. Knowledge Coverage and Timeliness

RAG relies on external knowledge corpora, which may lack coverage for niche domains or fail to include up-to-date information. This limits the model's ability to handle queries requiring specialized or real-time knowledge.

Expanding the knowledge corpus to include domain-specific sources (e.g., medical journals, legal databases) and integrating dynamic updates (e.g., live data feeds) could improve coverage and timeliness. Additionally, leveraging techniques like continual learning could help RAG systems adapt to evolving knowledge.

6.2. Future Directions

Current RAG systems typically retrieve documents once at the beginning of the generation process. However, dynamically retrieving documents during generation (e.g., based on intermediate outputs) could improve relevance and accuracy. Active retrieval strategies, where the model decides when and what to retrieve, could further enhance performance. Extending RAG to handle multi-modal inputs (e.g., text, images, videos) could enable more applications, such as generating responses based on

DOI: 10.54254/2755-2721/142/2025.KL22312

visual or audio context [13]. This would require advancements in multi-modal retrieval and generation techniques.

Deploying RAG in resource-constrained environments (e.g., edge devices) requires reducing its computational and memory footprint. Techniques like model distillation, quantization, and efficient retrieval algorithms could make RAG more accessible for real-world applications.

Incorporating human feedback into the retrieval and generation process could improve the quality and reliability of RAG systems. For example, users could validate retrieved documents or refine generated responses, enabling the model to learn from human expertise.

As RAG systems are deployed in sensitive domains (e.g., healthcare, and education), addressing ethical concerns such as bias, fairness, and privacy becomes critical. Future research should focus on developing frameworks to ensure responsible use of RAG technology.

7. Conclusion

This paper presented a study of RAG, a framework that combines the advantages of retrieval-based methods and generative models to address the limitations of purely generative models in knowledge-intensive tasks. This paper's work demonstrates that RAG effectively leverages external knowledge to enhance the accuracy, relevance, and interpretability of generated text, while maintaining the flexibility and fluency of state-of-the-art language models.

Through extensive experiments on MS MARCO, this paper showed that RAG outperforms both purely generative models (GPT-3.5-turbo) by evaluating three metrics, including TF-IDF, BERT Score and Q-BLEU-1. The ablation studies further highlighted the importance of end-to-end training and the synergistic interaction between the retriever and generator components.

Despite its promising results, RAG faces several challenges, including retrieval efficiency, multi-hop reasoning, and knowledge coverage. Addressing these limitations will be critical for advancing the capabilities of retrieval-augmented systems. Future work should explore dynamic retrieval strategies, multi-modal integration, and lightweight deployment techniques to make RAG more scalable and adaptable to real-world applications.

In conclusion, RAG represents a significant step forward in bridging the gap between retrieval and generation, offering a powerful framework for knowledge-intensive NLP tasks. This paper believes that these findings will inspire further research in this direction, ultimately leading to more robust, interpretable, and trustworthy AI systems.

References

[1] *Roberts, A., Raffel, C., & Shazeer, N. (2020). How much knowledge can you pack into the parameters of a language model?. arXiv preprint arXiv:2002.08910.*

[2] *Marcus, G. (2020). The next decade in AI: four steps towards robust artificial intelligence. arXiv preprint arXiv:2002.06177.*

[3] *Glendinning, I. (2013). Comparison of policies for academic integrity in higher education across the European Union.Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. REALM: Retrieval-augmented language model pre-training. ArXiv, abs/2002.08909, 2020.*

[4] *Lee, K., Chang, M. W., & Toutanova, K. (2019). Latent retrieval for weakly supervised open domain question answering. arXiv preprint arXiv:1906.00300.*

[5] *Karpukhin, V., Oguz, B., Min, S., Lewis, P. S., Wu, L., Edunov, S., ... & Yih, W. T. (2020, November). Dense Passage Retrieval for Open-Domain Question Answering. In EMNLP (1) (pp. 6769-6781).*

[6] *Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in neural information processing systems, 35, 27730-27744.*

[7] *Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., & Deng, L. (2016). Ms marco: A human-generated machine reading comprehension dataset.*

DOI: 10.54254/2755-2721/142/2025.KL22312

[8] Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., & Weston, J. (2018). Wizard of wikipedia: Knowledge-powered conversational agents. arXiv preprint arXiv:1811.01241.

[9] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers) (pp. 4171-4186).

[10] Nema, P., & Khapra, M. M. (2018). Towards a better metric for evaluating question generation systems. arXiv preprint arXiv:1808.10192.

[11] Liu, P. J., Saleh, M., Pot, E., Goodrich, B., Sepassi, R., Kaiser, L., & Shazeer, N. (2018). Generating wikipedia by summarizing long sequences. arXiv preprint arXiv:1801.10198.

[12] Faghri, F., Fleet, D. J., Kiros, J. R., & Fidler, S. (2017). Vse++: Improving visual-semantic embeddings with hard negatives. arXiv preprint arXiv:1707.05612.

[13] Wang, L., Li, Y., & Lazebnik, S. (2016). Learning deep structure-preserving image-text embeddings. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5005-5013).

[14] Lin, X. V., Socher, R., & Xiong, C. (2018). Multi-hop knowledge graph reasoning with reward shaping. arXiv preprint arXiv:1808.10568.

[15] Weston, J., Chopra, S., & Bordes, A. (2014). Memory networks. arXiv preprint arXiv:1410.3916.