Feedback-Driven Continuous Learning in Retrieval-Augmented Generation Systems

Shreekara Mahapatra[1], Rishi Raj[2], Divyansh Singhal[1], and Sarang Agrawal[3]

[1] School of Computer Engineering, Kalinga Institute of Industrial Technology (KIIT)
Deemed to be University, Bhubaneswar-751024, Odisha, India.
22053548@kiit.ac.in,
[2] Information Systems, Indian Institute of Management Visakhapatnam, India.
Rishi.raj@iimv.ac.in
[3] Computer Science and Data Science, Netaji Subhas University of Technology,
Delhi, India.
sarang.agrawal.ug23@nsut.ac.in

Abstract. The fusion of Retrieval-Augmented Generation(RAG) in Conversational AI has profoundly improved its capacity to deliver accurate and context-aware responses by combining the retrieval of relevant information from the knowledge base with human-like language generation. However, Conventional RAG models are often restricted to static knowledge bases which become outdated over time and hinder the capacity of these systems to adapt to proprietary domains in healthcare, education, finance and investments etc. This research proposes a new continuous learning and dynamic knowledge update framework of RAG to address these limitations. The proposed framework is designed to not only enhance the response accuracy and context relevance but also to improve user satisfaction by incorporating real-time data update and user feedback loops. Specifically, our method uses an implicit and explicit feedback-driven system to gather user feedback after each interaction to guide continuous learning, allowing the model to refine the knowledge base and improve its response. This not only reduces the problem of information staleness, but also enhances customer satisfaction. The proposed model has shown significant gains in performance parameters such as, 16% improvement in response accuracy and 24% improvement in user satisfaction over the conventional RAG system . The proposed framework is also a contribution to the growing field of intelligent conversational agents as it provides insights on building an adaptive RAG model to address the challenges of proprietary and dynamic information environment.

Keywords: Continuous Learning, Retrieval-Augmented Generation, Feedback Loop, Dynamic Knowledge Update, Conversational AI

1 Introduction

The swift development of natural language processing (NLP)[2] and artificial intelligence (AI)[3] technologies has opened the door for creative systems that can produce logical and contextually appropriate answers to user inquiries. One such paradigm is Retrieval-Augmented Generation (RAG), which blends generating models and retrieval methods in a synergistic way. This method improves the systems' capacity to generate precise and contextually relevant replies by allowing them to anchor their responses in a dynamic corpus of information. However, traditional RAG models are frequently constrained by static knowledge bases, which restrict their capacity to adjust to changing information and lessen their usefulness in practical applications that need for current knowledge.

It has become more and more clear in recent years that AI systems must be able to learn from encounters and dynamically update their knowledge in addition to retrieving pertinent information. Developing AI systems that can function well in dynamic contexts requires models to have self-learning capabilities, which allow them to enhance performance in response to user feedback and newly accessible data. One prospective solution to the problems with static knowledge bases is the integration of continuous learning techniques into RAG systems.

Key Contributions: Specifically, our key contributions are:

- Introduction of the feedback loop and dynamic document ingestion into the system for user interaction.
- Evaluate user interaction after each interaction and identify areas of improvement
- Trigger reprocessing and updating of the documents in the vector-store if a knowledge gap is found.
- Fine-tune the retrieval and generation mechanism as per the emerging requirements.

The remainder of this paper is organized as follows: section 2 briefly reviews the related works. In section 3, we present the methodology of our work. Section 4 describes the evaluation and results, followed by applications in section 5. Section 6 provides the limitations of the proposed RAG system. Finally section 7 includes the conclusion and references.

2 Related Work

Early conversational models mostly depended on rule-based systems, in which responses were produced using hard-coded rules or predetermined templates. The effectiveness of early retrieval-based models, like TF-IDF and BM25[10], in addressing complicated or varied queries were restricted since they were able to return documents based on keyword similarity but were unable to capture semantic meaning (Robertson & Zaragoza, 2009) . Dense Passage Retrieval (DPR) (Karpukhin et al., 2020) was developed as advanced conversational systems, using dense embeddings produced by transformers to more effectively match questions with pertinent passages [9]. Even though DPR showed an increase in retrieval accuracy, it was still unable to provide feedback-driven or continuous updates.

The original RAG architecture, proposed by Lewis et al. (2020)[1], represented a significant advancement by combining retrieval with generative language models . RAG retrieves relevant documents, which are then used as context for a transformer-based language model to generate a coherent response. This architecture effectively improved response contextualization by grounding generation in retrieved knowledge. However, a key limitation of the original RAG framework is that it is trained on a fixed knowledge base. Post-deployment, the model cannot dynamically update its knowledge or adapt to evolving information, making it insufficient for applications that require real-time accuracy, such as customer support or healthcare information systems.

Recent work on feedback loops in reinforcement learning (RL)[7] and retrieval mechanisms suggests that incorporating explicit user feedback can significantly enhance a system's relevance and personalization capabilities (Ziegler et al., 2019). Additionally, most existing RAG models are limited by a static vector store, which restricts their ability to dynamically ingest and re-process new documents based on emerging needs or identified knowledge gaps. Table 1 shows a clear comparison between the related works and the proposed RAG and conveys why a feedback-driven, self-learning RAG is a better choice over conventional RAG system.

	Models				
	Rule-Based Systems	**Early Retrieval-Based Models (TF-IDF, BM25)**	**Dense Passage Retrieval (DPR)**	**Original RAG (Lewis et al., 2020)**	**Self-Learning and Dynamic Knowledge Update RAG (Proposed)**
Knowledge Base Type	Fixed, rule-based	Static corpus	Static corpus	Static knowledge base	Dynamic knowledge base, capable of real-time updates
Retrieval Method	Hard-coded rules	Keyword similarity	Dense embeddings	Dense retrieval integrated with generation	Real-time dense retrieval with knowledge updates
Semantic Understanding	Limited	Limited	Improved, transformer-based	High, with contextual document grounding	High, with improved contextual grounding
Feedback Integration	None	None	None	Limited	Self-learning feedback loop, uses explicit user feedback
Knowledge Updating Mechanism	Manual updates	Manual updates	None	None, requires retraining	Continuous, automated updates through incremental indexing
Adaptability to New Information	Very low	Low	Low	Low	High, dynamically adapts to new data
Response Generation	Template-based responses	Returns documents only	Returns documents only	Generates responses based on retrieved context	Generates responses with real-time, updated context
Deployment Scalability	Limited by rule complexity	Moderate	High	High	High, with scalability optimized for continuous updates
Applications	Limited to simple, static scenarios	Static information retrieval	Static knowledge scenarios	Static and slow-evolving knowledge domains	Dynamic, real-time applications (e.g., customer support, healthcare)
Memory Management	Not applicable	Not applicable	Not applicable	Requires manual KB management	Automatic memory pruning and relevance-based retention
Example Use Cases	FAQ, basic decision trees	Document retrieval systems	Document retrieval in Q&A settings	Knowledge-grounded dialogue	Knowledge-rich, dynamic applications requiring updates

Table 1. Comparison of Related Works in RAG and the Proposed RAG System

3 Methodology

3.1 Dataset Preparation

We used information from 3 publicly accessible datasets covering a broad variety of issues in order to make it easier to construct a self-learning and dynamic knowledge updating RAG system. The following are the main datasets used in this study:

- NaturalQuestions (NQ)[4]- Approximately 300,000 examples (7GB). This dataset contains real-world questions posed by users, paired with relevant passages from Wikipedia. It serves as a rich source of open-domain question-answering data.
- MS MARCO[5] - Over 1 million query-passage pairs in the training set (9GB). This dataset provides questions along with multiple potential answers sourced from web documents, enabling the model to learn from a diverse set of contexts.
- HotpotQA[6] - Around 113,000 question-answer pairs (3.5GB). A dataset created especially for addressing multi-hop questions, which necessitates combining data from several publications. This component is essential for evaluating the model's capacity to integrate data from various settings.

3.2 Data Pre-processing

The system first loads documents from specified sources, where each document undergoes pre-processing to ensure data integrity before being stored in a vector-based storage system for future retrieval. Pre-processing of data involves, tokenization, which is segmenting the loaded texts into manageable chunks using a text splitter configured with specific separators or token limits. To enable semantic understanding, an embedding model, from Hugging Face, is initialized to convert these text chunks into dense vector embeddings, facilitating context-rich document representation. These embeddings are then stored in a vector store, optimized for efficient retrieval in response to queries. A detailed pre-processing architecture is shown in Figure 1

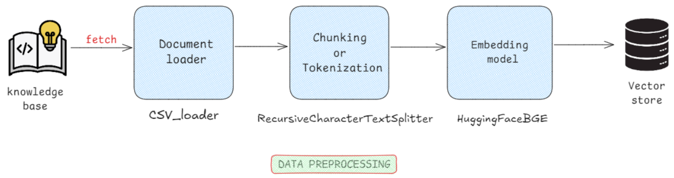

Fig. 1. Data pre-processing architecture

3.3 Retrieval, Generation Feedback Mechanism

Retrieval - The retrieval mechanism is responsible for identifying and retrieving relevant documents from a vector store, which houses embeddings created from pre-processed text chunks. To achieve this, the system employs a dense vector representation model, specifically the advanced embedding models from Hugging Face. Each query is transformed into an embedding vector, which is then matched against the vectors in the vector store using similarity metrics, cosine similarity in our case. By relying on semantic embeddings, the retrieval mechanism overcomes limitations of traditional retrieval systems, such as TF-IDF, which are typically limited to surface-level, keyword-based matching. The proposed system retrieves the top-k relevant documents, dynamically adjusting retrieval parameters based on user feedback and learned patterns to refine and optimize retrieval accuracy over time.

Generation - In order to generate a logical and contextually aware answer, we have used GPT(Generative Pre-Trained Tranformers). The generative model is grounded on the retrieved material, which directs it to produce answers that not only answer the query but also represent the most recent and pertinent data from the knowledge base. The output of the generative model is adjusted by conditioning on user feedback and domain requirements to improve response specificity and coherence

Feedback - The feedback mechanism, which records both explicit and implicit user feedback to continuously improve and refine the retrieval and generation processes, is essential to the system's capacity for self-improvement. User ratings, comments, and corrections are examples of explicit feedback. These are recorded and examined to pinpoint areas where answer relevance or accuracy may be lacking. The retrieval and creation components are then modified based on these feedback inputs. For example, if a specific query type repeatedly produces unsatisfactory results, the feedback loop identifies these interactions and re-assesses the retrieval parameters or relevant document embeddings.

4 Evaluation and Results

4.1 Experimental Setup

All the experiments were done on system having the following specifications: Intel i7-1165G7 CPU, Intel Iris Xe GPU, 16.0 GB of RAM, 64-bit operating system.

4.2 Evaluation Metrics

We have compared the proposed model with the conventional model on the basis of response accuracy, relavence and user satisfaction of the models.

Response Accuracy: Response accuracy is measured when the produced response is compared to a collection of ground-truth in order to determine the

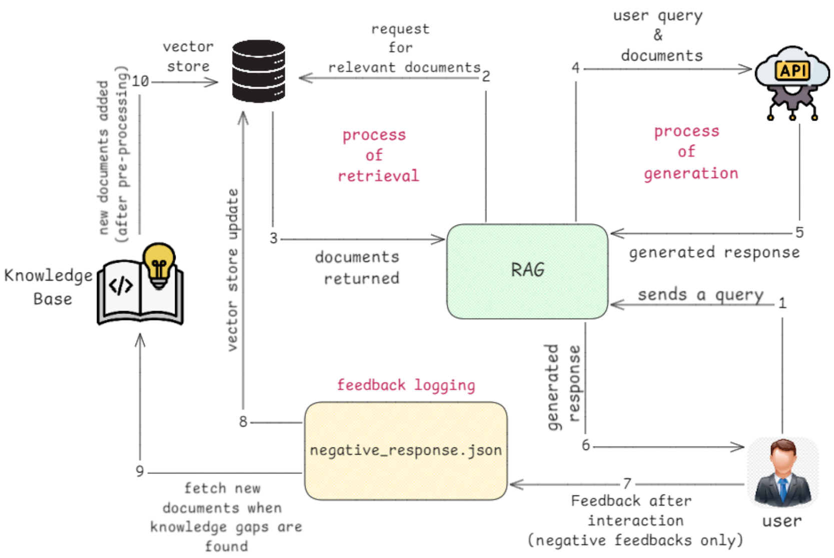

Fig. 2. The Retrieval, Generation & Feedback Mechanism

response accuracy. Metrics that measure how well the model's output and reference answers match, like BLEU[8] in our system, are frequently used to quantify this.

$$RA = \frac{1}{n}\sum_{i=1}^{n}(A_i \times S_i) \tag{1}$$

Let: A_i be the accuracy score for the i-th generated response, where $A_i = 1$ if it matches the expected answer within a tolerance threshold and $A_i = 0$ otherwise. S_i represent a similarity score derived from automated metrics, that is BLEU, indicating similarity to the reference answer. n be the total number of evaluated responses.

Response Relevance (RR): Response Relevance evaluates how relevant the retrieved documents are to the query, which directly impacts the quality of generated answers. Relevance is measured based on cosine similarity metrics ranging from 0-1. 1 if the response is very similar and 0, if the response if completely different.

$$RR = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{k}\sum_{d\in D_i}\mathrm{sim}(d, q_i) \tag{2}$$

Let:D_i be the set of retrieved documents for the i-th query. q_i be the i-th query. $\mathrm{sim}(d, q_i)$ be the similarity between document $d \in D_i$ and query q_i, calculated using cosine similarity metric. k be the number of top-relevant documents considered for each query.

User Satisfaction: User satisfaction is frequently assessed using both explicit and implicit feedback systems, including click-through rates, user ratings, and the frequency of query rephrasing. Rating scales ranging from 1-5 is used for explicit feedback(1 for poor and 5 for excellent), whereas behavioral indications that indicate possible gaps in the original response, such as the frequency of follow-up inquiries or the pace at which queries are rephrased, are used in the form of implicit feedback.

$$US = \frac{1}{n}\sum_{i=1}^{n}(F_i \times T \times R_i) \tag{3}$$

F_i be the feedback score provided by the user for the i-th response, ranging from 1 (unsatisfactory) to 5 (highly satisfactory). n be the total number of user feedback scores collected. T be a task completion score, where $T = 1$ if the user's task was completed successfully and $T = 0$ otherwise. R_i be the relevance score of the response for the i-th query, measured by cosine similarity or a semantic matching metric, with $R_i \in [0, 1]$.

4.3 Results

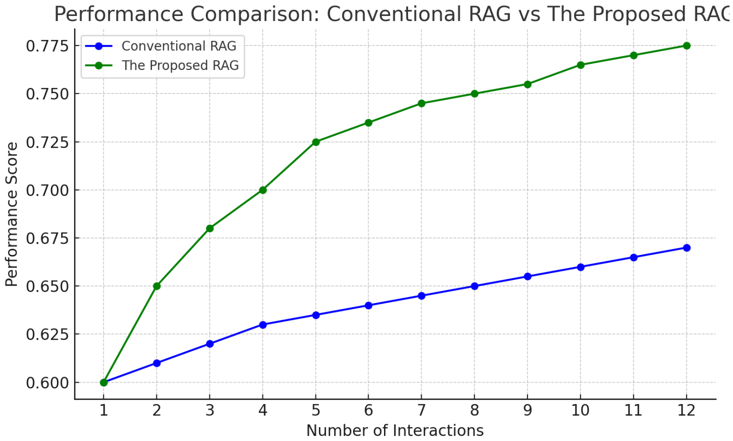

Fig. 3. Graph comparing the response accuracy of conventional RAGs and the proposed RAG

The graph in Figure 3 illustrates the response accuracy of both conventional RAG system and proposed RAG system. Both systems begin with similar baseline performance, with the Performance Score hovering around 0.6(60%).This indicates that at the onset, there is little to no significant difference between the conventional RAG system and the proposed RAG system, as both systems are likely initialized with the same knowledge base and models.The blue line representing the conventional RAG shows minimal improvement, going upto a maximum level of 0.64(64%) in 12 interactions. On the other hand, the red line representing the proposed RAG system shows a growth upto 0.80(80%) in 12 interactions.

Iteration	Conventional RAG	Proposed RAG
1	**User Query:** What is the recommended dosage for XyloMed? **Response:** ”I couldn’t find specific information on XyloMed in the current knowledge base. Please consult a healthcare provider for accurate details.” **Feedback:** N/A **Cosine Similarity Score:** 0.45	**User Query:** What is the recommended dosage for XyloMed? **Response:** ”I couldn’t find specific information on XyloMed in the current knowledge base. If you have more details or references, I can update my knowledge to assist better in the future.” **Feedback:** ”The recommended dosage for XyloMed is 50 mg once daily for adults. This information is available in the FDA’s latest release.” **Cosine Similarity Score:** 0.88
2	**User Query:** Can XyloMed be taken by pregnant women? **Response:** ”I couldn’t find specific information on XyloMed. Please consult a healthcare provider for guidance.” **Feedback:** N/A **Cosine Similarity Score:** 0.50	**User Query:** Can XyloMed be taken by pregnant women? **Response:** ”I couldn’t find specific information on the use of XyloMed during pregnancy. If you can provide references, I can assist better in the future.” **Feedback:** ”XyloMed is contraindicated during pregnancy due to potential risks, as per the FDA guidelines.” **Cosine Similarity Score:** 0.85
3	**User Query:** Can XyloMed be taken by pregnant women? (Follow-Up) **Response:** N/A (No updated information from Conventional RAG) **Feedback:** N/A **Cosine Similarity Score:** N/A	**User Query:** Can XyloMed be taken by pregnant women? (Follow-Up) **Response:** ”XyloMed is contraindicated during pregnancy due to potential risks, as per the FDA guidelines. Please consult a healthcare provider for further clarification.” **Feedback:** N/A **Cosine Similarity Score:** 0.90

Table 2. Comparison of Response Relevance Between Conventional RAG and The Proposed RAG Across 3 Iterations

The graph in Figure 4 illustrates the level of user satisfaction of the conventional and the proposed RAG system. We can infer from the graph that both the proposed RAG (blue line) and the Conventional RAG (red line) start with similar user satisfaction scores, around 60%. This suggests that both systems provide comparable initial performance, likely because both models share the

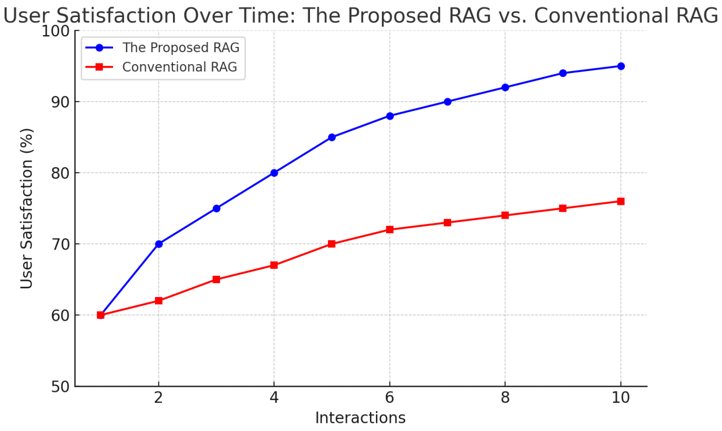

Fig. 4. Graph comparing the user satisfaction of conventional RAGs and the proposed RAG

same knowledge base and retrieval setup when first deployed. However, after 10 feedback interactions user satisfaction for the Conventional RAG shows only a slight improvement, stabilizing around 71%. The gradual, minor improvement indicates that this RAG system lacks the feedback-driven adaptation capabilities. The user satisfaction for proposed RAG shows a consistent upward trend across all user interactions, reaching around 95% satisfaction by the final interaction. This increase reflects the RAG's continuous learning mechanism, which incorporates feedback, ingests new information dynamically, and adjusts based on user needs.

5 Applications

The feedback-driven continuous learning RAG system has multiple applications:

- **Customer Service**: The proposed RAG system can be deployed as an intelligent customer support chatbot that leverages continuous learning algorithms. By analyzing user interactions and feedback, the chatbot dynamically updates its knowledge base regarding product details, company policies, and frequently asked questions (FAQs). This involves using natural language processing (NLP) techniques to extract insights from conversations, which are then utilized to refine the chatbot's responses and improve its accuracy over time.
- **Healthcare Assistant**: In the healthcare sector, a virtual assistant powered by a self-learning RAG system can provide support to patients, doctors, and nurses by retrieving relevant medical knowledge. Utilizing medical databases and guidelines, the system offers recommendations regarding medications and treatment plans. It continuously improves its recommendations based on

feedback from healthcare professionals, allowing it to adjust to new protocols or changes in treatment practices.

- **Financial Market Analysis**: In financial analysis, the proposed RAG system can assist analysts by providing real-time updates on market data and trends. By integrating feedback mechanisms, these systems learn from the accuracy and relevance of previous predictions and analyses. This involves utilizing financial datasets and employing advanced algorithms to refine predictive models continuously, ensuring that analysts receive timely and pertinent information for decision-making.
- **Education**: In the educational domain, the proposed self-learning RAG system can be implemented to enhance the learning experience by adapting educational content based on student feedback. The system analyzes responses to questions, quizzes, and other assessments to modify explanations and learning materials dynamically. This approach ensures that content remains relevant and tailored to individual learning needs, thus facilitating personalized education.

6 Limitations

The proposed framework, while advantageous for continuous learning and dynamic knowledge update, it has several limitations also. The major problem of the proposed RAG system is feedback that is biased or inaccurate. Personal prejudices, misinterpretations, or subjective judgments may affect user-provided input, which might skew the system's learning process and result in inaccurate knowledge base entries. When accuracy and dependability are essential in high-stakes situations, this might be very harmful. Reliance on the quantity and quality of feedback is still another major obstacle. The system's capacity to progress depends mostly on regular, high-quality input; therefore, any lack of or decrease in feedback volume might impede the learning process and make it more difficult for the system to adapt over time. Feedback noise and ambiguity might make it more difficult for the system to handle user input in a useful way. Not all input may be specifically instructive since users may provide ambiguous remarks or even contradicting information, which makes it challenging for the system to precisely identify insights that may be put to use. Security and privacy concerns are especially urgent in delicate fields where input could include private information, such as healthcare or the judicial system. Maintaining safe data handling procedures and adhering to privacy laws are crucial for reducing the risks of data breaches and abuse.

7 Conclusion

To improve conversational AI, the proposed Retrieval-Augmented Generation (RAG) system is a revolutionary solution. Incorporating feedback is a new method for dynamic knowledge updating and continual learning. In contrast to the conventional models, like TF-IDF or BM25, which match queries based on keyword similarity but lack semantic understanding, or even Dense Passage Retrieval (DPR), which addressed this shortcoming but is unable to adjust dynamically, our model actively learns from user feedback to improve its knowledge base, which over time increases its contextual relevance and responsiveness. This can effectively solve the issue of knowledge gaps as well. Together, the retrieval and production procedures enable the system to produce contextually relevant replies, and the feedback mechanism offers an organized method for gradually improving and growing its knowledge base. This learning loop enhances the system's relevance in knowledge for specific domains such as customer service, healthcare, legal research, and more, in addition to improving the system's accuracy in answering inquiries. To enable the wider deployment of self-learning RAG models in dynamic and high-stakes knowledge domains, future research should concentrate on improving the infrastructure's efficiency, resolving security issues, and refining the feedback mechanism to minimize possible biases.

References

1. Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks." Advances in Neural Information Processing Systems 33 (2020): 9459-9474.
2. Nadkarni, Prakash M., Lucila Ohno-Machado, and Wendy W. Chapman. "Natural language processing: an introduction." Journal of the American Medical Informatics Association 18, no. 5 (2011): 544-551.
3. Ertel, Wolfgang. Introduction to artificial intelligence. Springer Nature, 2024.
4. Kwiatkowski, Tom, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein et al. "Natural questions: a benchmark for question answering research." Transactions of the Association for Computational Linguistics 7 (2019): 453-466.
5. Nguyen, Tri, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. "Ms marco: A human-generated machine reading comprehension dataset." (2016).
6. Yang, Zhilin, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. "HotpotQA: A dataset for diverse, explainable multi-hop question answering." arXiv preprint arXiv:1809.09600 (2018).
7. Arulkumaran, Kai, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. "Deep reinforcement learning: A brief survey." IEEE Signal Processing Magazine 34, no. 6 (2017): 26-38.
8. Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. "Bleu: a method for automatic evaluation of machine translation." In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pp. 311-318. 2002.
9. Karpukhin, Vladimir, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. "Dense passage retrieval for open-domain question answering." arXiv preprint arXiv:2004.04906 (2020).
10. Robertson, Stephen, and Hugo Zaragoza. "The probabilistic relevance framework: BM25 and beyond." Foundations and Trends® in Information Retrieval 3, no. 4 (2009): 333-389.