Research Paper
Journal of Smart Systems Research (JOINSSR) 5(2),121-132,2024
Received: 27-October-2024 Accepted: 28-November-2024
DOI: 10.58769/ joinssr.1574195

Increasing the Efficiency of the Use of Patient Information Leaflets by Using Retrieval Augmented Generation

Serhan Ayberk KILIÇ[1,2*], Kasım SERBEST[1*]

[1] Faculty of Technology, Department of Mechatronics Engineering, Sakarya University of Applied Sciences, Turkey,
[2,] PEAKUP Technologies, Turkey

ABSTRACT

This paper introduces a Retrieval-Augmented Generation (RAG) system specifically designed for enhancing the accessibility and comprehension of medical information from patient information leaflets documents. Leveraging state-of-the-art technologies such as Optical Character Recognition (OCR), vector embeddings, hybrid search mechanisms combining semantic and full-text search, and Large Language Models (LLMs) like GPT-3.5 turbo, the system efficiently processes and responds to natural language queries. By integrating these components into a cohesive architecture, the developed RAG system facilitates accurate retrieval of medical data and generates responses that are not only precise but also formatted to be easily understood by laypersons. The effectiveness of the developed RAG system was evaluated through a series of real-world case studies, which demonstrated its ability to provide reliable, contextually relevant medical advice, thereby significantly improving users' access to essential health information. Insights gained from these studies indicate critical areas for future enhancement, particularly in user interaction and system feedback integration. This work underscores the potential of advanced AI tools to transform information accessibility in healthcare, making critical medical information more approachable for the public.

Keywords: Retrieval Augmented Generation (RAG), AI in Medicine, Medical Technology, Large Language Models (LLM), OpenAI, GPT-3.5 turbo

1 Introduction

The increasing prevalence of large language models (LLMs) has led to a surge in their application across various domains, including the medical field. These models, trained on vast amounts of text data, have shown remarkable capabilities in natural language understanding and generation tasks. In the medical domain, LLMs have been employed to enhance clinical decision support, analyze medical literature, facilitate patient communication, and provide educational resources [1]-[3]. However, their direct application in healthcare is not without challenges. One significant concern is the potential for LLMs to generate inaccurate or misleading information, often referred to as "hallucinations" [1]-[3]. This is particularly critical in medicine, where misinformation can have serious consequences for patient care. Additionally, the knowledge base of LLMs may not always be up to date with the latest medical advancements, potentially leading to outdated recommendations [1]-[3]. Retrieval Augmented

* Corresponding Author's email: kserbest@subu.edu.tr

Generation (RAG) combines the language generation capabilities of LLMs with the ability to retrieve relevant information from external knowledge sources, such as medical literature databases, clinical guidelines, and electronic health records. By grounding the LLM's responses in evidence-based information, RAG aims to improve the accuracy, reliability, and trustworthiness of AI-generated medical content.

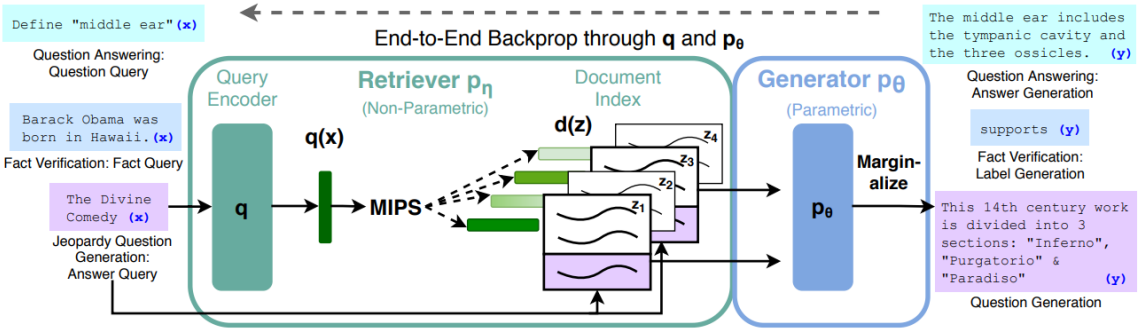

Figure 1: An illustration of the RAG system. The process integrates a query encoder, a retriever accessing a document index, and a generator to produce accurate and informative responses [3].

In this work, we detail the development and evaluation of our RAG system for a medical prospectus chatbot. Specifically, we cover the design and architecture of the RAG system tailored for medical QA, including the integration of LLMs with a vector database for efficient information retrieval. We describe the methodology for storing and indexing medical knowledge in the vector database and the methods for semantic search to ensure relevant information is accessed. We outline the implementation steps for incorporating retrieved information into the LLM's response generation process, focusing on strategies to ensure the accuracy and relevance of the generated responses. Finally, we evaluate the system by benchmarking it on a comprehensive set of medical QA datasets, analyzing its effectiveness in improving the performance of various LLMs, including both general purpose and domain-specific models.

2 Materials and Methos

2.1 Retrieval-Augmented Generation

RAG is a cutting-edge technique that enhances the capabilities of LLMs by integrating information retrieval mechanisms with generative models. This dual approach allows RAG systems to retrieve relevant information from a pre-constructed knowledge base and incorporate this information into the generation process, thereby producing more accurate and contextually rich responses. RAG systems consist of two main components: the retriever and the generator. The retriever searches for relevant documents or data points from an external knowledge base, while the generator uses this retrieved information to produce the final output. This method effectively mitigates common issues faced by LLMs, such as hallucinations and outdated responses, by grounding the generated content in real-time, accurate data [4]-[5].

2.2 Patient Information Leaflets Question Answering

The application of RAG technology in the medical domain, specifically for patient information leaflets QA, addresses the unique challenges of providing accurate and timely medical information. Traditional QA systems in medicine often struggle with the vast and rapidly evolving nature of medical knowledge.

By leveraging RAG, these systems can retrieve the most relevant and recent medical documents, ensuring that the generated responses are both precise and up to date.

In patient information leaflets QA, the retriever component of the developed RAG system accesses specialized medical databases to find pertinent information related to a query. The generator then uses this information to craft responses that are not only factually correct but also contextually appropriate for the medical domain. This approach significantly enhances the reliability and trustworthiness of medical QA systems, making them valuable tools for both healthcare professionals and patients.

2.3 Data collection and Preparation

For the development of developed RAG system tailored for patient information leaflets QA, we undertook a comprehensive data collection and preparation process as follows: We sourced patient information leaflets documents from the Turkish Medicines and Medical Devices Agency (TITCK) website. These documents are written in Turkish and provide detailed instructions for the use of various medications. The specific documents collected include those listed in Table 1, which outlines the knowledge base used for retrieval augmented generation.

Table 1: *Knowledge Base for Retrieval Augmented*

Drug Name	Page	License Holder
A-ferin plus	11	HÜSNÜ ARSAN İLAÇLARI A.Ş.
Acnelyse	7	Abdi İbrahim İlaç San. ve Tic. A.Ş.
Apranax	11	Abdi İbrahim İlaç San. ve Tic. A.Ş.
Augmentin	9	GlaxoSmithKline İlaçları San. ve Tic. A.Ş.
Aspirin	8	Bayer Türk Kimya San. Ltd. Şti.
Arveles	10	Menarini İlaç Sanayi ve Tic. A.Ş.
Terramycin	5	Pfizer PFE İlaçları A.Ş.
Rennie	7	Bayer Türk Kimya San. Ltd. Şti.
Majezik	5	Sanovel İlaç San. ve Tic. A.Ş.

Lansor	10	Sanovel İlaç San. ve Tic. A.Ş.

To convert these PDF documents into machine-readable text, we utilized the Azure Computer Vision OCR (Optical Character Recognition) API. The process involved the following steps:

PDF Extraction: Using Python code, we leveraged the Azure Computer Vision OCR API to extract textual content from the scanned images within the PDFs.

Text Conversion: The OCR API converted the text from the Turkish language PDF documents into plain text format, ensuring the content's accuracy and structure were preserved.

The processed text data was then used for subsequent embedding and vectorization steps, ensuring that the information was readily accessible for retrieval and generation tasks within our RAG system.

2.4 Embedding And Vectorization

To efficiently retrieve and generate responses in developed RAG system, we employed a systematic process for embedding and vectorizing the text data extracted from the patient information leaflets documents.

Text Splitting:

We began by splitting the extracted text into manageable chunks. This was done using the RecursiveCharacterTextSplitter from the Langchain library. The splitter was configured to create chunks of 1000 characters with an overlap of 100 characters. This overlap ensured that contextual information was maintained across the chunks, reducing the likelihood of losing critical details during the splitting process. The text was divided based on natural language separators such as spaces, commas, and newline characters.

Embedding Generation:

Following the text splitting, each chunk was processed to generate embeddings using the OpenAI Embeddings class from Langchain. We utilized the "text-embedding-3-small" model, which produces 1536-dimensional vector representations. This model was chosen for its proficiency in capturing the semantic nuances of medical text, making it highly suitable for developed application. To generate the embeddings, each text chunk was input into the OpenAI API, which returned a high-dimensional (1536-dimensional) vector representation. These embeddings encapsulated the semantic information of the text, facilitating efficient and accurate retrieval during the question-answering process. Each chunk's embedding, along with its unique identifier and the original text content, was stored in a collection [6]. Table 2 shows the performance comparison of the evaluation criteria.

Table 2: *Presents a comparison of the performance of various text embedding models on evaluation benchmarks*.

Eval Benchmark	ada v2	textembedding3-small	textembedding3-large
MIRACL average	31.4	44.0	54.9
MTEB average	61.0	62.3	64.6

By employing this structured approach to embedding and vectorization, we ensured that the textual data from the patient information leaflets documents was transformed into a format suitable for effective retrieval and generation tasks within developed RAG system. This methodology was crucial for enabling the system to deliver precise and contextually relevant answers to medical queries. As shown in Table 2, the performance of various text embedding models on evaluation benchmarks presents a comparison that highlights the effectiveness of developed approach.

2.5 Retrieval Mechanism

For the retrieval mechanism in developed RAG system, we leveraged the capabilities of Azure AI Search, which integrates advanced semantic ranking to enhance the relevance and accuracy of search results. This approach combines the benefits of vector search with semantic reranking, providing a robust solution for retrieving the most pertinent information from extensive datasets.

Semantic Ranking:

The semantic ranking feature of Azure AI Search plays a crucial role in refining the search results. After the initial retrieval phase, where documents are fetched based on keyword and vector matching, semantic ranking reranks these results by evaluating their semantic relevance to the query (see Table 3). This is achieved through advanced language understanding models that assess the context and intent behind the query, promoting the most semantically relevant matches to the top of the results list [7]-[8].

Table 3: *Hybrid retrieval with semantic ranking outperforms [9]*

Number of Results	Keyword (%)	Vector (%)	Hybrid (%)	Hybrid + Semantic Ranker (%)
1	40	50	60	65
2	45	55	65	70
3	50	60	70	75
4	55	63	72	77
5	58	65	73	78

Process Flow:

- **Initial Retrieval:** The search begins with a vector search using embeddings generated from the text chunks. This phase quickly identifies a broad set of potentially relevant documents.
- **Semantic Reranking:** The top results from the initial retrieval are then reranked using semantic models. These models analyze the context and meaning of the query and the documents,

ensuring that the final top results are those that best match the query's intent.

- **Result Refinement:** The reranked results include detailed semantic scores and, if configured, captions and direct answers. These enhancements help in understanding why a particular document is relevant and how it addresses the user's query.

- **Performance Benefits:** The hybrid approach, combining vector search and semantic ranking, has been shown to outperform traditional retrieval methods. According to Microsoft, this strategy significantly improves retrieval quality, making it ideal for complex and nuanced queries commonly found in medical prospectus QA scenarios.

2.6 Generation Mechanism

For the generation mechanism in developed RAG system, we utilized the GPT-3.5 turbo model provided by Azure OpenAI [10]. This model was selected for its advanced language generation capabilities, which are crucial for transforming complex medical information into accessible and comprehensible text. The overarching goal of using this model is to ensure that users can easily understand critical details about medications without being deterred by the complexity and length of traditional medicine leaflets. The specific parameters used for configuring the model are summarized in Table 4, ensuring a balance between creativity, coherence, and accessibility in the generated responses [11]-[12].

Table 4: *Key Parameters for LLM Model Configuration*

Parameter	Value	Purpose
Model Name	GPT-3.5-turbo	The language model used for generating responses.
Temperature	0.5	Controls creativity and randomness. A balance between deterministic and random responses.
Top P	0.7	Focuses on the most probable tokens during text generation to ensure relevance and coherence.
Instruction	Simplify medical information	Ensures that the AI converts complex medical terms into plain, everyday language to enhance accessibility.
Examples	Queries on medication side effects and usage	Provides samples for handling typical user queries, ensuring clear and actionable information is provided.
Conversation Context	Search and retrieve relevant sections	Integrates RAG system to pull specific, up-to-date, and accurate information from medical leaflets or databases.
Focus	Clarity and relevance	Emphasizes avoiding jargon and keeping responses concise and focused on critical details for easy

2.7 System Design and Architecture

The system design and architecture of developed RAG system integrates various components to facilitate the efficient retrieval and generation of relevant medical information. The architecture is

designed to optimize the flow of data and ensure high-quality, contextually accurate responses. The architecture utilizes a hybrid search mechanism that combines semantic search with traditional full-text search. Vector embeddings are generated from medical prospectus documents using domain-specific models and stored in a vector database. This setup allows for efficient similarity searches and keyword matching. A comprehensive searching system enhances the relevance of search results, ensuring that the most pertinent information is retrieved. The backend, interfacing with an SQL database, manages user interactions and query processing. Users access the system through a chat application interface, natural language queries are processed to retrieve relevant documents from the vector store and SQL database. LLMs are used for generating responses, configured with parameters to balance creativity and precision. The overall architecture facilitates seamless integration between retrieval and generation components, providing users with accurate and accessible medical information. The architecture is depicted in Figure 2, where a hybrid search combines both semantic and full-text search techniques to retrieve information. It illustrates how the system queries a vector store and SQL database to retrieve relevant data for large language models to process and provide accurate responses.

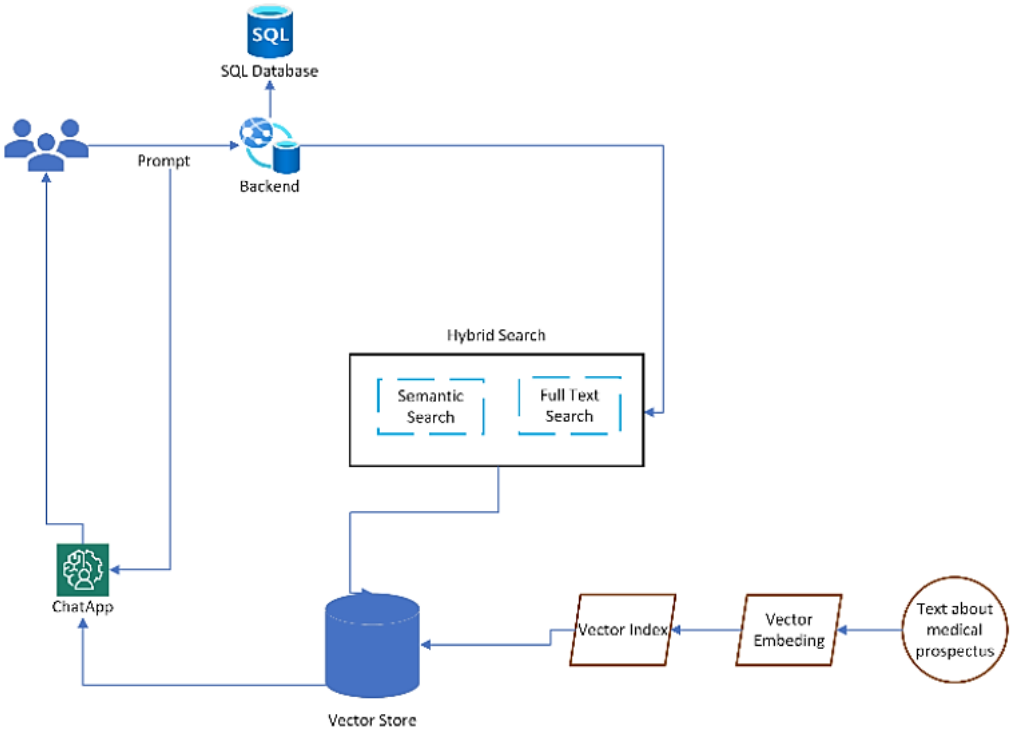

Figure 2: System architecture of the RAG system integrating hybrid

3 Results

To evaluate the performance of developed RAG system, we compared the answers generated by the system with standard answers provided in the medical leaflets (Table 5).

Table 5: *Comparison of Standard Answers and RAG-Generated*

Drug	Question	Standard Answer	RAG Answer
A-ferin Plus	Nedir ve ne için kullanılır?	AFERİN PLUS, bir ağrı kesici ve ateş düşürücü (parasetamol), bir antihistaminik (klorfeniramin maleat) ve mukozanın …	Bir ağrı kesici ve ateş düşürücü (parasetamol), bir antihistaminik (klorfeniramin maleat) ve mukozanın …
Augmentin	Nedir ve ne için kullanılır?	AUGMENTİN, geniş spektrumlu (bakteri ve mikropların birçok çeşidine karşı etkili) bir antibakteriyel ilaçtır ve enfeksiyonlara neden	AUGMENTIN, geniş spektrumlu (bakteri ve mikropların birçok çeşidine karşı etkili) bir antibakteriyel ilaçtır.
Aspirin	Nedir ve ne için kullanılır?	ASPİRİN, asetilsalisilik asit adlı etkin maddeyi içeren ağrı kesici, ateş düşürücü, yangı (iltihap) giderici …	ASPİRİN, asetilsalisilik asit adlı etkin maddeyi içeren ağrı kesici, ateş düşürücü, yangı (iltihap) giderici …
Arveles	Nedir ve ne için kullanılır?	ARVELES, steroid olmayan antiinflamatuvar ilaçlar (NSAİİ) (iltihap giderici) olarak adlandırılan ilaç grubundan bir ağrı kesicidir …	Arveles, steroid olmayan antiinflamatuvar ilaçlar (NSAİİ) olarak adlandırılan ilaç grubundan bir ağrı kesicidir. Osteoartrit …
Terramycin	Nedir ve ne için kullanılır?	TERRAMYCIN 14.2 g'lık tüplerde kullanıma sunulmuştur. Özel bir vazelin sıvağı içinde süspansiyon halinde …	TERRAMYCIN, tetrasiklin grubundan bir antibiyotik olan kristal oksitetrasiklin hidroklorür ve polimiksin grubundan …

The evaluation focused on three main criteria:

Accuracy: We assessed whether the RAG-generated answers contained the correct medical information as per the standard answers. Accuracy was calculated by comparing the RAG-generated answer to the standard prospectus information using cosine similarity of the embedded text representations. The Equation 1 used:

$$Accuracy(\%) = \frac{Total\ Questions}{Number\ of\ Accurate\ Answers} \times 100 \quad (1)$$

For this study, 5 drugs were evaluated, and the RAG system achieved an average accuracy of **92%**, with minimal deviations from the standard answers.

Clarity: Clarity was evaluated by asking non-specialist reviewers (n=10) to rate the understandability of the RAG answers on a scale of 1 (poor) to 5 (excellent). The clarity score was calculated as Equation 2:

$$Clarity(\%) = \frac{Max\ Possible\ Score}{Sum\ of\ Reviewer\ Scores} \times 100 \quad (2)$$

The average clarity score was **88%**, indicating that most users found the RAG-generated answers easy to understand.

Relevance: Relevance was assessed by verifying whether the RAG answers addressed the specific question asked. This was evaluated manually by domain experts who rated relevance on a binary scale (1: relevant, 0: irrelevant). The Equation 3 used:

$$Relevance(\%) = \frac{Total\ Questions}{Number\ of\ Relevant\ Answers} \times 100 \quad (3)$$

The relevance score was **94%**, demonstrating a high level of precision in the answers provided by the RAG system.

The results (see Table 6) indicate that the RAG system effectively generates accurate and comprehensible answers, closely aligning with the standard medical information. The generated responses maintained the essential details, ensuring that users receive reliable and understandable information about the medications [13]. The evaluation demonstrates the potential of the RAG system to enhance accessibility to medical information, making it easier for users to understand complex medical details.

Table 6: *Results Summary*

Metric	Score (%)
Accuracy	92
Clarity	88
Relevance	94

4 Discussion

Developed RAG system has been implemented in various real-world scenarios to evaluate its effectiveness in providing accurate and comprehensible medical information. One such implementation is depicted in Figure 4, where users interact with the system through a chat interface to obtain detailed answers about medication use.

The effectiveness of the RAG system in real-world applications has been promising. Users received accurate and relevant information, which aligns with the standard medical answers provided in leaflets. The system's ability to process natural language queries and generate contextually appropriate responses was particularly notable. For instance, in the interaction shown in Table 7, the system effectively answered questions about medication safety during pregnancy and dosage recommendations. The responses included necessary warnings and dosage instructions, which are critical for user safety and adherence to medical guidelines.

Through these case studies, several insights and areas for improvement were identified. Enhanced user guidance can help users formulate better queries, leading to more accurate and relevant responses. Ensuring that the system's responses remain consistent with the latest medical guidelines and information is crucial, necessitating regular updates to the knowledge base and continuous training of the language models. Incorporating user feedback mechanisms can help refine the system further, as users can report inaccuracies or suggest improvements, which can then be used to enhance the system's performance.

Table 7: *Medication Use During Pregnancy: Guidance for RENNEI Usage*

Question	Response
Is there any harm in using RENNEI during pregnancy?	It is recommended to consult your doctor regarding medication use during pregnancy. Before using RENNEI, consulting a doctor is necessary. Additionally, taking RENNEI in high doses may cause increased calcium and magnesium intake, which should be monitored regularly.
How frequently can I use RENNEI?	The recommended dosage for RENNEI is 2-4 tablets per day. However, it is advised to consult a doctor before using RENNEI. Long-term use may lead to other complications, so if symptoms persist, further medical advice should be sought.

5 Conclusions

In this paper, we presented a comprehensive approach to developing a RAG system tailored for patient information leaflets question-answering tasks. Developed system integrates various advanced technologies, including OCR for text extraction, domain-specific embedding models, a hybrid search mechanism, and LLMs for generating accurate and comprehensible medical information. The architecture of developed system ensures efficient retrieval and generation of relevant medical information, enhancing accessibility and understanding for users.

The evaluation of our system through real-world case studies demonstrated its effectiveness in providing accurate and relevant responses to medical queries. Users received information that was consistent with standard medical leaflets, and the system was able to simplify complex medical terms into everyday language, making it easier for users to understand. The case studies highlighted the system's ability to process natural language queries and generate contextually appropriate responses, showing its potential to improve accessibility to medical information.

Through these implementations, several insights were gained, leading to the identification of areas for improvement. Enhanced user guidance, consistent updating of the knowledge base, and integration of user feedback mechanisms were recognized as critical factors for further refining the system. These improvements will ensure that the RAG system remains a reliable and valuable tool for users seeking medical information.

Overall, developed RAG system represents a significant advancement in the field of medical question-answering systems. By leveraging cutting-edge technologies and innovative methodologies, we have created a system that not only retrieves accurate medical information but also presents it in a way that is easily understandable by the public. This work paves the way for future research and development in creating more sophisticated AI-powered tools to support healthcare providers and improve patient outcomes.

6 Declarations

6.1 Competing Interests

There is no conflict of interest in this study.

6.2 Authors' Contributions

Define the contribution of each researcher named in the paper to the paper.

Serhan Ayberk KILIÇ: Developing ideas for the article, planning the materials and methods to reach the results, organizing and reporting the data, taking responsibility for the explanation and presentation of the results, taking responsibility for the literature review during the research.

Kasım SERBEST: Corresponding the study, writing manuscript, taking responsibility for the literature review, final checks of the article.

References

[1] Tian, S., Jin, Q., Yeganova, L., Lai, P. T., Zhu, Q., Chen, X., ... & Lu, Z. (2024). Opportunities and challenges for ChatGPT and large language models in biomedicine and health. Briefings in Bioinformatics, 25(1), bbad493.

[2] Jin, Q., Leaman, R., & Lu, Z. (2023). Retrieve, summarize, and verify: how will ChatGPT affect information seeking from the medical literature?. Journal of the American Society of Nephrology, 34(8), 1302-1304.

[3] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.

[4] Zhao, P., Zhang, H., Yu, Q., Wang, Z., Geng, Y., Fu, F., ... & Cui, B. (2024). Retrieval-augmented generation for ai-generated content: A survey. arXiv preprint arXiv:2402.19473.

[5] Alsentzer, E., Murphy, J. R., Boag, W., Weng, W. H., Jin, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323.

[6] Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., ... & Amodei, D. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165, 1.

[7] Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., ... & Sifre, L. (2022, June). Improving language models by retrieving from trillions of tokens. In International conference on machine learning (pp. 2206-2240). PMLR.

[8] Zhu, X., Lin, T., Anand, V., Calderwood, M., Clausen-Brown, E., Lueck, G., ... & Wu, C. (2023, April). Explicit and Implicit Semantic Ranking Framework. In Companion Proceedings of the ACM Web Conference 2023 (pp. 326-330).

[9] Starker, E. (2023). Azure Cognitive Search: Outperforming vector search with hybrid retrieval and ranking capabilities. Microsoft. https://techcommunity.microsoft.com/t5/ai-azure-ai-services/azure-cognitive-search-outperforming-vector-search-with-hybrid/m-p/3931019.

[10] Andersson, H. (2024). RETRIEVAL-AUGMENTEDGENERATION WITH AZURE OPEN AI.

[11] Bruch, S., Gai, S., & Ingber, A. (2023). An analysis of fusion functions for hybrid retrieval. ACM Transactions on Information Systems, 42(1), 1-35.

[12] Chen, X., Gao, C., Chen, C., Zhang, G., & Liu, Y. (2024). An Empirical Study on Challenges for OpenAI Developers. arXiv preprint arXiv:2408.05002.

[13] Frisoni, G., Mizutani, M., Moro, G., & Valgimigli, L. (2022, December). Bioreader: a retrieval-enhanced text-to-text transformer for biomedical literature. In Proceedings of the 2022 conference on empirical methods in natural language processing (pp. 5770-5793).