THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

Exploring Agentic Retrieval-Augmented Generation Frameworks in Healthcare Recommendation Applications

by

Branden Lee

School of Information Technology and Electrical Engineering,

University of Queensland.

Submitted for Master of Data Science Capstone Project 1

1st May 2025

Introduction

This project proposes an application powered by Large Language Models (LLMs) that delivers accurate and reliable healthcare recommendations by developing an agentic Retrieval-Augmented Generation (RAG) framework. While LLMs present outstanding potential within the healthcare field, the serious nature of the field demands the utmost accuracy and reliability, areas where current LLM applications often fall short (Millett & Stekelenburg, 2024). Implementing RAG in LLM-powered applications can generate highly accurate and relevant responses, but they may still hallucinate and struggle to adapt in real-time situations because they rely on static training data (Merritt, 2025). This project tackles these critical limitations by exploring the implementation and evaluation of an agentic RAG system in LLM-powered applications in healthcare. The system will retrieve its information from relevant, substantial healthcare-related datasets including detailed medical research papers and up-to-date medical information from reliable online sources. With these sources processed, intelligent artificial intelligence (AI) agents capable of sophisticated query planning, multi-source retrieval, and information validation will also be deployed. This approach aims to significantly improve the accuracy and reliability of LLMs in healthcare. In this proposal, the critical problems, the proposed methods, and objectives will be discussed in detail further, though understanding the baseline is crucial. Therefore, this next section highlights the current landscape from which the identified problems emerge.

Background

Understanding the potential and limitations of RAG requires examining recent developments in AI tools for healthcare. AI is advancing quickly, generating significant excitement for its potential as a tool to be used in healthcare applications. Technologies like LLMs are becoming much more capable and are already behind many advanced virtual assistants and chat bots. LLMs are trained on vast amounts of data and allows them to excel at understanding, summarising, generating, and predicting complex forms of text (Clusmann et al., 2023). These factors make them promising tools for healthcare applications, such as assisting with diagnoses or informing treatment strategies. However, it is crucial to use robust, reliable methods for AI tools to access medical information and deliver precise responses, given the importance of this field and the need for accuracy (Millett & Stekelenburg, 2024). To overcome these challenges and improve LLM-powered applications, numerous AI frameworks and modern methods have emerged (Aman, 2025).

One such method that is gaining attention is Retrieval-Augmented Generation (RAG). It enhances accuracy and reliability by utilising powerful search algorithms to perform targeted searches within a

specific, trusted collection of external information based on the user's prompt (Merritt, 2025). This external information may be a database containing modern medical journals or lists of approved drug information. The next step of RAG involves augmenting the prompt by adding onto the prompt with the retrieved information. This augmentation will provide the LLMs a more comprehensive understanding and allows them to produce more factual and precise responses (Martineau, 2023). Although, what would happen if the information it finds is incorrect or not up to date?

Problem identification

While RAG offers a promising solution for improving LLM performance, there are two critical problems in many LLMs that need to be addressed: hallucinations and the lack of real-time adaptability. Firstly, hallucinations are outputs that are grammatically coherent but are factually incorrect, nonsensical, or disjoint from an input prompt that can occur due to limitations in training data and models such as inaccuracy or bias (IBM, 2023). Since they sound coherent, it makes it difficult for those who are not as knowledgeable in the topic at hand to detect misinformation. Secondly, there is the lack of real-time adaptability. As discussed previously, LLMs can use RAG to gather information and provide themselves with relevant and factual context. However, traditional RAG typically relies on static retrieval mechanisms that do not adapt to evolving tasks or user interactions (Nosta, 2024). For the healthcare industry, it is a clear priority to provide factually correct and up-to-date information. Thus, there is a need for a more advanced approach.

Agentic Retrieval-Augmented Generation (Agentic RAG) is an improved approach to traditional RAG. Rather than simply retrieving and generating information, agentic RAG utilises multiple AI agents to perform several vital tasks. It can handle complex requests by planning a step-by-step process, retrieve and assess information from multiple data sources, and adapt to changing contexts, which results in more accurate, comprehensive, and contextually relevant responses (Belcic & Stryker, 2025). To explore agentic RAG in more detail, the project is guided by several key objectives.

Objectives

In this project exploring the effectiveness of agentic RAG, there are several key objectives that can be broken down into two main parts: development and evaluation. For development, it involves the following steps:

- deploy open-source LLMs and shape them with prompt engineering
- implement a data ingestion pipeline to use real-time and static data
- develop traditional and agentic RAG frameworks within an LLM

After the development stage, the evaluation stage can commence and involves the following steps:

- investigate how multiple AI agents can improve retrieval efficiency and adaptability
- compare agentic RAG directly to traditional RAG
- identify key performance bottlenecks and propose optimisations

To achieve these objectives, the following methods are proposed in the next section.

Proposed methods

To begin, a significant objective in exploring RAG involves collecting factual data. Collecting data that is correct and reliable is crucial as it will directly affect the overall system's retrieval accuracy and reasoning capabilities (Davydova, 2025). The data will be collected from multiple trustworthy sources, including Healthdirect Australia and PubMed Central, through a data ingestion pipeline. Data from the Healthdirect website will act as the main source for the agentic RAG framework as it contains up-to-date and reliable information for a high number of health topics. Collecting this information requires implementing a specific strategy.

Retrieving data from a website involves legally extracting it through a process known as web scraping, which will be a significant process for the RAG frameworks. At first, the extracted data is unstructured text data with HTML tags. However, this data will be processed and transformed into structured data, allowing it to be analysed more easily. A random sample of the 639 web pages have been scraped and analysed; see Figure 1. Initial findings show that each page has a significant number of words, which is important as many words, or data, is necessary for the LLMs to contextualise the data (Davydova, 2025). Furthermore, a high word count is required to effectively evaluate RAG frameworks. To avoid legal and ethical issues, no private information was and will be collected during this process. Analysing this dataset is simple as data from each page has a clear topic and format, but the same cannot be said for the other dataset from PubMed.

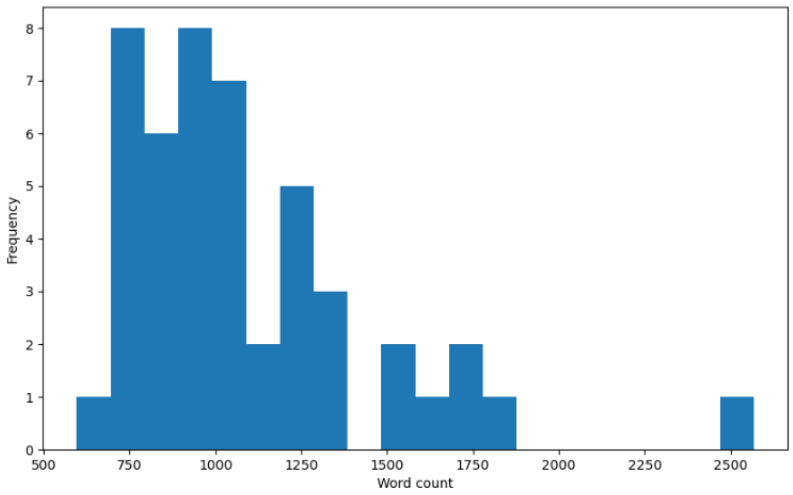


Figure 1: Word counts for webpages from Healthdirect.gov.au

PubMed consists of millions of medical research papers in the form of text documents that will be useful data for RAG frameworks for validation or to find contradictions. Although the size of the dataset is excellent, it can also be problematic. Due to the vast amount of unstructured text data, a significant effort is required for it to be used effectively. 3,027 research papers were sampled to explore the data. Like the Healthdirect data, there was a high number of words for each research paper, meaning enough data is available; see Figure 2. However, identifying the usefulness of the data is uncertain as the text documents unlabelled. An analysis performed on the samples shows a glimpse of the topics that can be found; see Appendix A. Some of the most common topics included DNA, cancer, and treatment. Although, expanding the view showed that there was a clear bias towards certain topics; see Appendix B. This finding suggests that a specialised tool is required to use this dataset effectively.

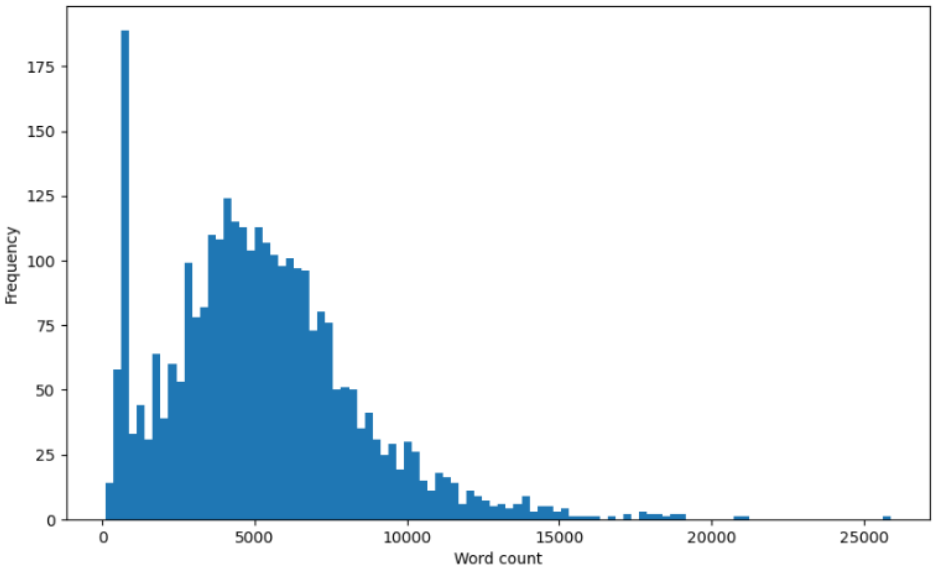


Figure 2: Word counts for research papers from PubMed

Vector databases is one of many specialised databases and will greatly assist RAG frameworks (Aquino, 2024). It is the final stage of the data ingestion pipeline where all collected data will be stored. A vector database helps an LLM contextualise text data by associating different words and sentences with each other. This process will significantly reduce processing time when it is required to search for relevant information (Holdsworth & Kosinski, 2024). For an application where it is important to provide a timely response, implementing a vector database is essential and thus, is a major objective for this project. Another major objective is using a method called prompt engineering.

One of the objectives describes the use of prompt engineering on the open source LLMs that will be used for the RAG frameworks (Jain, 2024). Typically, LLMs are designed to generate broad responses. However, they can be instructed to generate content that meets specific requirements through prompt engineering. For instance, it can instruct LLMs to ask certain questions to the user or to provide a response in a specific format (Jain, 2024). This method is simple and is particularly useful for the evaluation stage, which will involve small fine tunings to evaluate the outputs generated by the LLMs. Prompt engineering still has its own limitations such as lacking the ability to adapt to unforeseen scenarios so it cannot be used on its own. However, it can be very beneficial when used with other frameworks like RAG (Jain, 2024).

Both traditional and agentic RAG systems will be implemented to comparatively evaluate the two approaches. Initially, a traditional RAG system will be established and integrated into an LLM. Following this, the core of the project involves implementing a sophisticated agentic RAG system. This system will be built around multiple specialised AI agents which are essentially assistants that each have a designated task in the agentic RAG process. Some key AI agents include interpreting user intent and formulating precise search queries, dynamically selecting the most appropriate data and deploying search strategies, assessing retrieved information for relevance and quality, and rigorous validation by cross referencing information across different sources (Belcic & Stryker, 2025). Subsequently, a significant effort will be placed on optimising its efficiency and effectiveness. This can involve refining the retrieval workflows to ensure smooth data flow or enhancing agent collaboration protocols so that they work together seamlessly. With the traditional and agentic RAG systems operational, they can be directly compared and measured against each other. Performance metrics such as retrieval accuracy, answer coherence and relevance, and adaptability will be analysed to gain a deep understanding of each system's strengths and weaknesses. The following timeline demonstrates how long it will take to execute the proposed methods.

Project Timeline

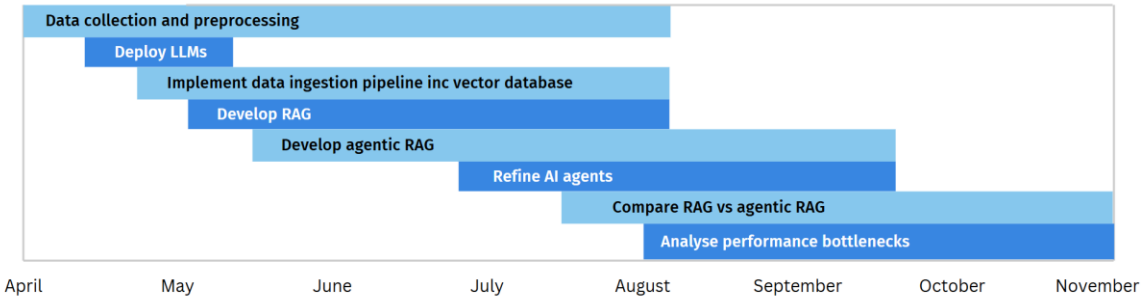


Figure 3: Planned project timeline

Figure 3 portrays a schedule that will allow the completion of the project in a timely manner while following the proposed methods. As previously mentioned in the Objectives section, the two stages of the project consist of implementation and evaluation. The project aims to complete the implementation stage as early as possible to place more emphasis on the evaluation stage. This will result in limiting the scope of the implementation, which will be discussed in greater detail in the next section.

Resources

This project will be implemented in a cloud environment due to the requirements for implementing an agentic RAG system. As such, the Google Cloud Platform (GCP) will be used to host most of the system, including the Open WebUI interface. GCP allows computers of various specifications to be remotely accessed and utilised while the Open WebUI interface will allow streamlined interactions with LLMs. Due to the heavy requirements of running one or more LLMs on a local computer, the Open WebUI interface will host Application Programming Interfaces (APIs). Using APIs will allow the interface to remotely connect to LLMs. Once implementation is complete, the cloud environment will conveniently allow the application to be accessed from anywhere with an internet connection. However, these resources do have costs.

While using cloud or remote-based resources do lower costs, some costs are still involved and can vary widely. For instance, the cost of using GCP will increase if a more powerful remote computer is used, where a computer with more power will result in faster performances. Another cost includes the APIs for the LLMs. The cost for APIs will increase if it is used more frequently, or if a more powerful model is selected. To lower costs, a less powerful LLM can be run locally, but this may negatively impact the evaluation of the RAG frameworks. Further testing is required to confirm how powerful an LLM is required to adequately complete this project.

Conclusion

To summarise, this project directly addresses the critical limitations of current LLM applications in healthcare: hallucinations and a lack of real-time adaptability. By proposing the development of an agentic Retrieval-Augmented Generation (RAG) framework which includes leveraging intelligent AI agents, implementing a sophisticated data ingestion pipeline, utilising relevant and reliable data sources, and robust evaluation metrics, this study aims to demonstrate the key advantages of agentic RAG. Once development is complete, the results will demonstrate enhanced accuracy, reliability, adaptability of LLM-powered healthcare recommendations. These results will be shown by the comparative analysis between traditional and agentic RAG frameworks. By the end, the project is expected to provide concrete evidence for the potential of agentic RAG and how it can make LLM-powered applications more reliable and effective within the demanding healthcare domain.

References

Alowais, S. A., Alghamdi, S. S., Alsuhebany, N., Alqahtani, T., Alshaya, A., Almohareb, S. N., Aldairem, A., Alrashed, M., Saleh, K. B., Badreldin, H. A., Yami, A., Harbi, S. A., & Albekairy, A. M. (2023). Revolutionizing healthcare: The role of artificial intelligence in clinical practice. *BMC Medical Education*, *23*(1). https://doi.org/10.1186/s12909-023-04698-z

Aman, Y. (2025, February 14). *LLM Model Optimization Techniques and Frameworks - Yugank .Aman - Medium*. Medium. https://medium.com/@yugank.aman/llm-model-optimization-techniques-and-frameworks-e21d57744ca1

Aquino, S. (2024, March 19). *What is RAG: Understanding Retrieval-Augmented Generation - Qdrant*. Qdrant. https://qdrant.tech/articles/what-is-rag-in-ai/

Belcic, I., & Stryker, C. (2025, March 3). *Agentic RAG*. IBM. https://www.ibm.com/think/topics/agentic-rag

Clusmann, J., Kolbinger, F. R., Muti, H. S., Carrero, Z. I., Eckardt, J.-N., Laleh, N. G., Löffler, C. M. L., Schwarzkopf, S.-C., Unger, M., Veldhuizen, G. P., Wagner, S. J., & Kather, J. N. (2023). The future landscape of large language models in medicine. *Communications Medicine*, *3*(1), 1–8. https://doi.org/10.1038/s43856-023-00370-1

Davydova, M. (2025, February 20). *Data Governance for Retrieval-Augmented Generation (RAG) - Enterprise Knowledge*. Enterprise Knowledge. https://enterprise-knowledge.com/data-governance-for-retrieval-augmented-generation-rag/

Holdsworth, J., & Kosinski, M. (2024, July 29). *Vector database*. IBM. https://www.ibm.com/think/topics/vector-database

IBM. (2023, September 1). *AI hallucinations*. IBM. https://www.ibm.com/think/topics/ai-hallucinations

Jain, P. (2024, May 6). *Use LLMs: Pre-training, Fine-tuning, RAG and Prompt Engineering*. Medium. https://medium.com/@jainpalak9509/use-llms-pre-training-fine-tuning-rag-and-prompt-engineering-564d5670f44d

Martineau, K. (2023, August 22). *What is retrieval-augmented generation?* IBM Research Blog; IBM. https://research.ibm.com/blog/retrieval-augmented-generation-RAG

Merritt, R. (2025, January 31). *What Is Retrieval-Augmented Generation?* NVIDIA Blog; NVIDIA. https://blogs.nvidia.com/blog/what-is-retrieval-augmented-generation/

Millett, D., & Stekelenburg, N. (2024, April 3). *Large language models in health: useful, not a miracle cure*. CSIRO. https://www.csiro.au/en/news/all/articles/2024/april/large-language-models

Nosta, J. (2024, September 12). *The Evolution of LLMs Through Real-Time Learning*. Psychology Today. https://www.psychologytoday.com/au/blog/the-digital-self/202409/the-evolution-of-llms-through-real-time-learning

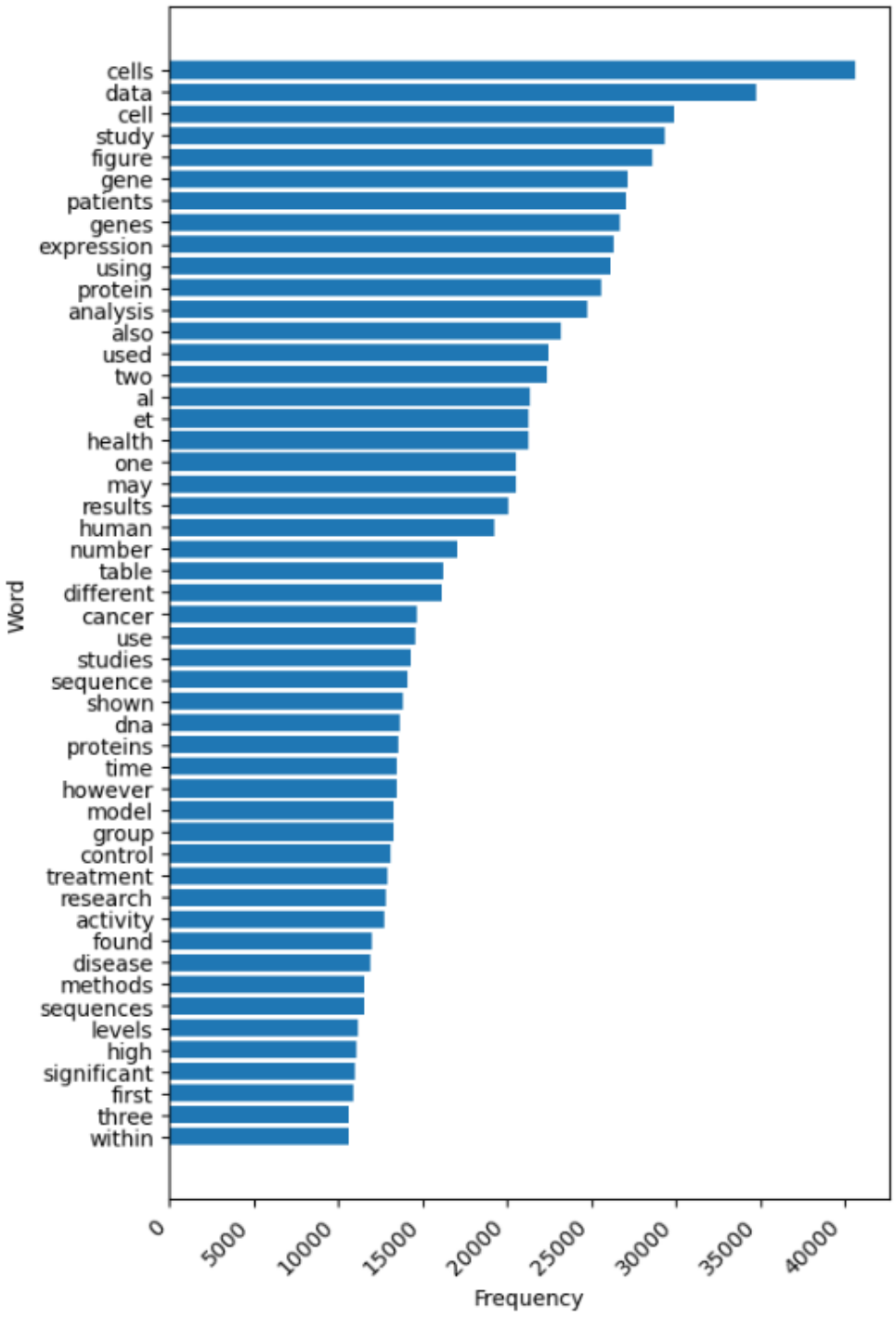
Datasets

healthdirect (2017, September 7). *Health topics* [Unpublished raw data]. healthdirect. https://www.healthdirect.gov.au/health-topics

PubMed Central (2017, September 7). *Health topics* [Dataset]. National Library of Medicine. https://ftp.ncbi.nlm.nih.gov/pub/pmc/oa_bulk/oa_comm/txt/

Appendix A

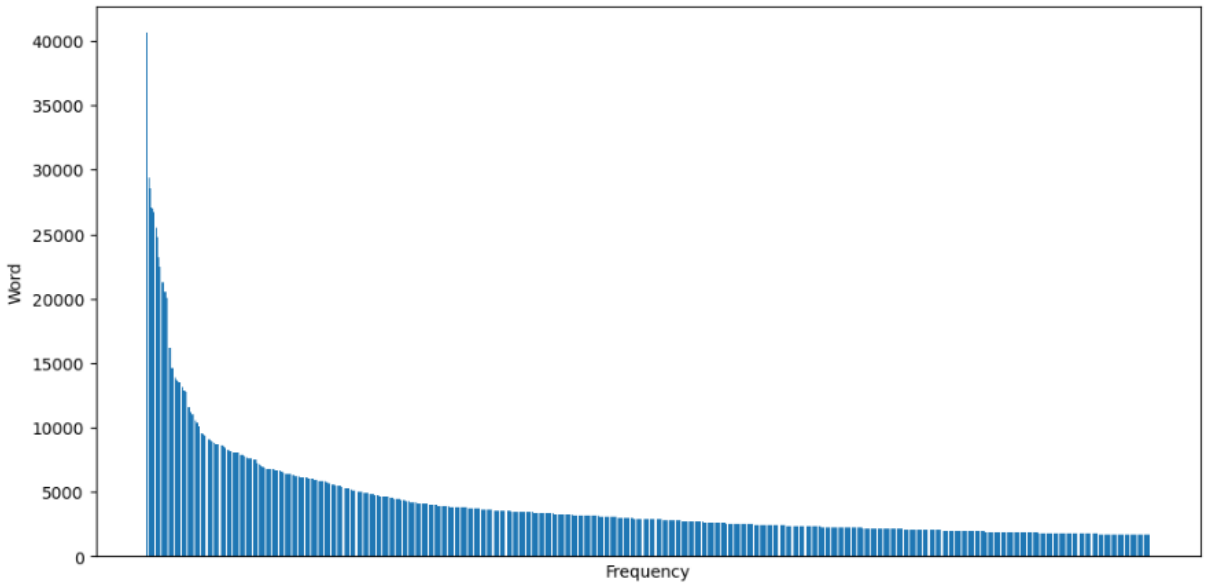
Top words with highest frequency



Note: The horizontal bar graph shows the top 50 words with the highest frequency from a random sample of about 3000 medical papers. The common words such as “the”, “is”, etc. have been filtered out.

Appendix B

Word frequencies for the top 1000 words



Note: This bar graph shows the top 1000 words with the highest frequencies from the same sample of medical papers. This graph shows a logarithmic-like shape where it suddenly drops and then tapers off. This indicates a bias towards a small number of medical topics.