Efficacy of Large Language Models in facilitating exam preparation for Medical students

by

Humayera Tabassum Tunan
22299539
Md Muntasir Mahmud Amit
21301088
Syed Tasrif Hasan
23141030
Md. Anwar Hossain
22101710

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
June 2025

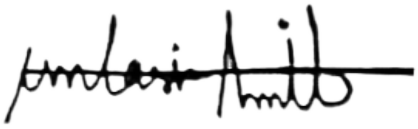
Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Humayera Tabassum Tunan
22299539



Md Muntasir Mahmud Amit
21301088

Syed Tasrif Hasan
23141030

Md. Anwar Hossain
22101710

Approval

The thesis/project titled "Efficacy of Large Language Models in facilitating exam preparation for Medical students" submitted by

1. Humayera Tabassum Tunan(22299539)
2. Md Muntasir Mahmud Amit(21301088)
3. Syed Tasrif Hasan(23141030)
4. Md. Anwar Hossain(22101710)

Of Spring 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on June 20, 2025.

Examining Committee:

Supervisor:
(Member)

Dr. Farig Yousuf Sadeque
Associate Professor
Department of Computer Science and Engineering
Brac University

Co-Supervisor:
(Member)

Nirjhor Datta
Lecturer
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

The complex medical curriculum compels the students to study a significant quantity of textbooks authored by various writers in each subject. These books contain a considerable amount of topics, and as a consequence, students frequently become overwhelmed by the heavy load of academic pressure. Moreover, in the four phases of medical studies, students perform six cards, three terms, and one professional examination in each phase to advance to the next phase. In this particular circumstance, students need adequate guidance to clarify their understanding and to address any questions that may arise during their preparation. To resolve these issues, an expert LLM developed with RAG, especially for medical students, is required. We intend to develop a retriever model powered by the proficiency of LLMs that will have access to the comprehensive resources of the medical curriculum and is able to solve any study-related confusion a medical student may have. We believe that our algorithm will help medical students overcome their fear of examinations, as well as improve their overall academic performance. LLM has great potential to give the perfect guide a medical student requires. LLMs integrated with the RAG system has the ability to give a medical student the perfect tutoring required to test their exam preparation. Our driving force motivation behind implementing this algorithm is to lessen the mental pressure of medical students and enhance their study experience by providing curriculum-aligned assessment tools.

Keywords: Large language Model, Retrieval Augmented Generation, Medical textbooks, Medical Studies, Medical education, Medical Examination Preparation, textbook retrieval, Artificial Intelligence, Generative AI, Natural Language processing, LLM, AI, RAG, NLP.

Table of Contents

Chapter 1

Introduction

1.1 History and Background

The increasing trajectory of technology is mostly the contribution of Artificial Intelligence, Large Language Models, Machine learning, Natural Language Processing, etc. LLMs like ChatGPT, Gemini, Meta Llama, and Claude ai have drastically changed the image of information technology in many fields including medicine, health care, and medical education. LLMs have the potential to assist medical students in their studies as they perform significantly well in medical education-related queries (Abd-alrazaq et al. [11]). Among all the other existing state-of-the-art LLMs, ChatGPT outperformed with an adequate high score (Safranek et al. [16]). It performs at a level expected from a third-year medical student (Gilson et al. [13]). Although it sometimes may face hallucinations and show misleading information as it is not solely dedicated to medical education. Some LLMs which are specialized in medical question answering such as JMLR, and LLM-AMT performed significantly better than ChatGPT and showed lesser hallucinations and incorrect information (Wang et al. [17]). It is evident that LLMs that are specialized in a particular field perform significantly better than those generalized ones. Therefore, we intend to implement a Large Language Model developed with medical textbooks, specifically based on the complex medical curriculum which will be solely dedicated to helping medical students in their learning process and exam preparation resulting in fewer mental health issues.

1.2 Research Problem Statement

Medical school provides a comprehensive curriculum featuring frequent distinguished assessments. Throughout 5 or 6 years of medical school, students have to perform 4 professional examinations and before every professional examination, they have to sit for 6 cards and 3 terms followed by an oral exam called item. The professional examinations consist of one written exam, one viva, one Objective Structured Practical Examination and one practical. This helps the students with the opportunity to reinforce their understanding of the materials with a clear idea of each topic and enhance their expertise through continued practice.

However, students often get overwhelmed by the pressure of significantly large lessons and back-to-back assessments within a very short period. Moreover, they have to study an adequate amount of books for every subject and distinguish writers for better understanding and clear knowledge. The overloaded syllabus, a considerable amount of books, and frequent assessments all together compel a student to study continuously without a break. It is challenging to maintain concentration for a long period. As a result, they suffer from tension, stress, depression, and demotivation which sometimes leads to degradation in academic performance as well as loss of interest in medical studies. Furthermore, students often struggle to find reliable and quick answers to their study-related confusion.

1.3 Research Objectives

In order to lessen the academic pressure and enhance the learning experience of medical students, we are introducing a retriever model which will utilize large language models and have access to widely used medical textbooks in most of the medical curriculum, which will help clarify any study-related confusions or queries of a student that may arise while taking exam preparation. The reason behind using medical textbooks is that most medical professionals prioritize medical textbooks as the most authentic source of medical information over Wikipedia (Wang et al. [25]). We utilized a number of widely used medical textbooks from various authors to train our model using Retrieval Augmented Generation so that students can blindly trust the information provided by our LLM.

We developed a Retriever Augmented generation (RAG) system, which includes LLMS (Large Language Models) and is developed with medical textbooks specialized in medical education, which will be solely dedicated to lessening the academic pressure and enhancing the overall learning experience of medical students.

Our research objective is to create and evaluate a Retrieval-Augmented Generation (RAG) system that uses the power of Large-Language-Models (LLM) and generates diverse, unique, subject-specific, board exam standard multiple-choice questions (MCQs) for first-year medical students in Anatomy, Physiology, and Biochemistry, enhancing medical education by providing curriculum-aligned assessment tools.

Chapter 2

Detailed Literature Review

2.1 Survey Methodology

We have reviewed plenty of research papers related to our chosen topic titled "Efficacy of Large Language Models in facilitating exam preparation for medical students". For choosing papers we took into consideration the number of citations, reputation of the publisher, latest research articles because LLMs are booming in recent years especially generative AI's, most notably those articles which are related to our thesis work. After selecting we thoroughly read each of the papers to find out what are the problems they have solved, which datasets they have used, also which technology they have used lastly, what was the ultimate performance and findings. All the paper summaries are below.

2.2 Related Works

Nagar et al. [24] proposed a unified framework called "uMedSum" for advancing medical abstractive summarization. It improves medical summarization by balancing faithfulness and informativeness compared to other summarization techniques. Previous summarization techniques have problems like confabulation, omitting critical details that are crucial for the medical field. "uMedSum" removes confabulation and adds missing key information to ensure summarization is both accurate and comprehensive. This paper uses three datasets MIMIC-III (publicly available), MedQSum (patient questions), and ACI Bench (doctor-patient dialogue). For technology and LLM models they have used several Large language models GPT-4, LLaMA-3, and Meditron. "uMedSum" combines existing SOTA LLM-based summarization techniques like Element aware summarization, In context learning. Also includes DeBERTa v3 and PubMedBERT to detect or remove confabulations and to add missing information which makes the summarization complete. In terms of performance "uMedSum" shows a significant improvement of 11.8% in reference-free metrics compared to previous GPT-4 based SOTA methods. It achieves both faithfulness and informativeness which indicates the summarization is more accurate and reliable. Doctors (clinical experts) evaluation process was involved in it. They check whether the factual consistency and content coverage is accurate or not. Moreover, doctors preferred "uMedSum" 6 times more than previous SOTA models. Unlike previous summarization techniques, "uMedSum" includes removing confabulation and adding missing key information which makes the summarization

technique more reliable and accurate.

Lucas et al. [22] primarily focus on the integration of large language models (LLMs) like GPT-4, GPT(previous GPT models), Gemini, BioBERT, BioGPT, BioMedLM, and many other LLMs into medical education in the paper titled "A symmetric review of large language models and their implications in medical education". The paper shows how to solve problems like how to integrate LLM in medical education, more likely to find a way to use the LLM models effectively utilized to medical education while addressing corners around information accuracy, critical thinking and also the ethical use of such technology. The author emphasizes the need for careful implementation to avoid overreliance on these models which could be inappropriate for critical thinking and decision-making skills in students. The paper started with 166 studies but there were 40 studies which were related to the topic. Firstly it explores LLM applications in various medical exams including medical admission exams and other specialized tests. Here they used USMLE questions and other medical exam questions that are not publicly available. However, some public datasets like PubMed have been used to train domain-specific models like BioBERT. The technologies they have used are GPT-4, GPT-3.5, MedPaLM, and other biomedical models. These models are used to assist in talks like question answering, clinical reasoning, and exam preparation. The paper also focuses on the growing trend of creating domain-specific models that are tuned on medical datasets to improve the performance of medical NLP tasks, which ensures more accurate and contextually appropriate outputs and medical education. Performance-wise this paper shows impressive results like GPT for passing various medical exams including the US admission Medicare exam without human interaction. Other models also did well in specialized fields such as surgery, cardiology, etc. Despite the success of LLM models, it still falls compared to human performance in certain areas particularly in clinical decision-making, which indicates that not only AI can do all the things human interaction is needed to get the job perfectly done. It also means that these LLM models can be further used to enhance the medical study system in the future.

Cheung et al. [12] highlighted the way to create high-quality multiple-choice questions for medical exams in the paper titled "ChatGPT versus Human in generating medical graduate exam multiple choice questions - a multinational perspective study". The paper shows the comparison between ChatGPT and experienced human examiners to create multiple-choice questions efficiently and correctly. It explores whether ChatGPT can make accurate and reliable MCQ questions like humans while maintaining the quality required for medical assessments. The dataset used in the paper is not publicly available but it includes some medical textbooks which are publicly available. Both AI and human-generated questions were based on standardized medical reference, which ensures the accuracy of the content. For the technology, they have used ChatGPT, especially ChatGPT Plus, which is a large language model developed by OpenAI. The study used the LLM models to generate 50 MCQs; also the other 50 MCQs were generated by experienced humans. After generating the MCQs the content is assessed by humans based on five criteria: appropriateness, clarity, relevance, discriminative, power of alternatives, and suitability for graduate medical exams. Depending on these criteria both AI-written MCQs and human-written MCQs were compared to see which one is better. The ultimate

result indicates that ChatGPT was able to generate MCQ more efficiently and also takes less time than humans (20 minutes for ChatGPT vs over 200 minutes for humans). Although the human-generated questions outperformed a bit better than LLM models, there is no significant difference in overall question quality. Based on the performance of the LLM models questions it indicates that the model can generate questions that are equally good as human written questions. However, there are some drawbacks of AI-generated questions such as it cannot create clinical scenarios questions properly yet. Most importantly the study shows that LLM models can reduce the workload of medical educators on maintaining high-quality exam content.

Yagnik et al. [27] focus on the challenge of improving medical question-answering systems using LLM models in the paper titled "MedLM: Exploring Language Models for Medical Question Answering Systems". The primary issue they have faced is the need for reliable, dominant specific, and accurate medical information for both health care and professionals also for patients. Some existing question-answering systems are based on information retrieval but most of the systems cannot provide personalized contest awareness responses. This paper shows the performance of general versus domain-specific language models in generating accurate question-answering systems without relying on external content. In this paper, they have used two publicly available datasets, both of which provide a wide range of medical topics like in treatment, diagnosis, and drug-related inquiries. These two papers are highly related to the field they are trying to explore. For the technology, they have used both general and medical-specific models. They have used LLM models like GPT-2 to GPT-3.5 also Bloom which is the decoder-only model. Also, they have used T5 models which are encoder-decoder models. Using these models they have explored various techniques like static and dynamic prompting methods to improve the quality of the responses. Fine-tuning involves adjusting the models using the MedQuAD and Icliniq datasets to better caption medical knowledge. After experiencing all the methods they have tried they found out their GPT-3.5 performed better than the static prompting method in terms of quantitative metrics. However, fine-tuned models like Bloom also performed well especially when trained on larger and more diverse datasets. For the qualitative results user survey reveals that GPT models generated responses performed better over human written answers. For the domain-specific task GPat-3.5 model performed significantly better and for dynamic prompting it has better accuracy of the medical question and system.

Abd-alrazaq et al. [11] portrayed a comprehensive review on the performance of GPT-4, a LLM, its advantages, disadvantages and future directions in the field of medical education. LLMs has significantly improved the educational experience for the medical students by assisting with medical writing and learning materials, developing medical curriculum or teaching methods, creating assessments and evaluations for practice, personalizing study plans, and so on and so forth. However, these leveraging outcomes come with several challenges including, copyright and privacy issue, inequity in access, academic dishonesty, lack of consistency and reliability, limited knowledge, algorithmic biasness, overreliance, and lastly lack of human emotions. In order to address these challenges, some future directives are also introduced in the paper for Academic Institutions, Educators, Students, Researchers, and LLM developers. Academic Institutions can provide guidance to students on how to properly

cite any content generated by LLMs. Educators can review contents generated by LLMs and validate the accuracy. Students must use LLMs ethically, while following guidelines and double checking the content accuracy before submitting. Researchers should maintain balance between technologies and human interacted feedback in education. Lastly, LLM developers should develop LLMs while considering the existing constraints, including accuracy, inequality, privacy, impartiality, contextual understanding, human engagement, misinformation and so on.

Gilson et al. [13] demonstrated the performance of ChatGPT compared to InstructGPT and CPT-3 in answering the questions in United States Medical Licensing Examinations(USMLE) step 1 and 2. Two pairs of multiple choice questions are used as data sets, which are AMBOSS question bank for medical students and National Board of Medical Examiners free 120 questions. The output response from ChatGPT was evaluated by logical justification, presence of information external and internal. Furthermore, every incorrect answer is labeled according to logical error, information error, and statistical error, all analyzed by python software. ChatGPT performed 64.4% (56/87) in step 1 and 57.8% (59/102) in step 2 in the NBME Dataset and 44% (44/100) in step 1 and 42% (42/100) in step two in the AMBOSS data set. The test results finally conclude that ChatGPT performs similarly to a third year medical student in step 1 and 2 of USMLE.

Safranek et al. [16] discussed the potential of the Large Language Model, ChatGPT in medical education. It highlights three positive use cases, differential diagnosis brainstorming, interactive practice cases, and multiple-choice question review. Moreover, it highlights a negative use case, Definitive Answer to Ambiguous Question. However, it also shows some limitations such as insufficient integration of contextual and external information, sensory and nonverbal cues, Rapport and Interpersonal interaction and alignment with medical education and patient care goals. Furthermore, LLM's few more limitations are copyright issues, misinformation and bias. Medical students, being the future leaders, must be aware of AI's limitations and ethical, legal, and practical challenges. The paper emphasizes the importance of understanding students' strengths and weaknesses in these rapidly evolving LLMs in healthcare. The final conclusion is, LLMs like ChatGPT can enhance medical education with innovative learning opportunities, preparing students as future leaders for the digital era of medicine.

Wang et al. [17] described the structure, working mechanism and process of creation of a Medical LLM named LLM-AMT a Large Language Model Augmented with Medical Text-Books. This model includes a Query Augmenter, a Hybrid Text-Book Retriever and a Knowledge Self Refiner. The Query Augmenter, consisting of query rewriting and query expansion components, converts generalized queries into medical terminology making it more suitable for professional databases. The Hybrid Text-Book Retriever, combining sparse and dense retriever, utilized 51 books which are professionally used for medical licensing preparation in the USA. The Knowledge Self Refiner refines long paragraphs into useful, critically important, and relevant passages, making it convenient for LLM readers. In order to evaluate the LLM-AMT model, as a dataset, three sets of Multiple Choice Questions (MedQA-USMLE, MedQA-MALE and MedMCQA) from professional board examinations

of the USA and China were used. Distinguished medical expert personnels evaluated the performance of LLM-AMT in four different ranks, which are Correct, Mostly Correct, Partly Correct and Wrong. LLM-AMT provided 36 Correct, 19 partly Correct and 36 wrong in 100 questions without options from the MedQA-USMLE, whereas GPT-3.5 provided 27 Correct, 14 partly Correct and 49 wrong. The final result shows that LLM-AMT being the in depth domain specific in Medical model showed significantly better performance than GPT-3.5 Turbo and GPT-4 Turbo.

Wang et al. [26] introduced JMLR, Joint Medical LLM and Retrieval Training which surprisingly outperforms some existing state-of-the-art LLMs such as Meditron, Llama, Claude3-Opus, and GPT-3.5. Medical text-books, research papers, clinical guidelines, and high quality medical documents were used as data resources for the model. In order to evaluate JMLR the Amboss Question Bank, the United States Medical Licencing Examination questions such as MMLU-Medical, MedMcQA and MedQA datasets were used. ColBERT was used as the initial model for information retrieval. Their model was compared with both open source and closed source models such as Meditron and Anthropic Claude3 respectively. As evaluation, they used a factuality metric based on F1 score naming UMLS-F as well as they reviewed the performance by three medical professionals. Then Cohen's Kappa was calculated for both the evaluations. JMLR scored an accuracy of 51.3% in MedQA, 68.3% in Amboss, 65.3% in MMLU-Medical and 64.1% in MedMcQA datasets.By the result it is clear that training Llama and ColBERT jointly shows better performance than training them separately. In conclusion of a meticulously conducted experiment and evaluation the JMLR model performed significantly better than other LLMs as well as showed lesser hallucinations, thereby enhancing the accuracy and reliability of generated responses and explanations.

Kung et al. [15] demonstrated the performance of the model ChatGPT in using it to solve USMLE-style questions. The USMLE has three steps, ChatGPT passes all the steps without any prior training or human interference. The main target is to establish whether an LLM can improve students' ability to learn these complex exam questions. They have used a dataset containing USMLE questions which is not publicly available. The technology evaluated was ChatGPT, which forms part of OpenAI's GPT-3.5, and the model was more than 60% accurate, which is above the pass mark required in USMLE. The ChatGPT model is promising in terms of improving learning among students but there are still limitations in reasoning and knowledge of specific questions. Overall, the result of the test is satisfactory, although this does not mark the end of the learning process since more effort is needed to improve reliability and accuracy in the learning of all USMLE questions. It is also important to mention that no AI tools may be used as a replacement for educators, but they may be valuable in enhancing student performance.

Pal et al. [10] deal with the crucial problem of crafting reliable AI systems that can solve complicated medical QA tasks in the project called "MedMCQA". Such tasks include medical exams, like AIIMS and NEET PG, which require multi-hop reasoning and solving unstructured problems. One of the main problems was the absence of a comprehensive domain-specific QA dataset that could test the abilities of AI to perform reasoning in this domain. The authors tackled this problem by

developing MedMCQA, which is a medical MCQA dataset that includes 194,031 MCQA problems in 21 medical subjects and 2,400 healthcare topics. The authors built test sets used in real-world exams to test various reasoning types including question types to test diagnostics and treatment. The dataset is free and available online, and all reasoning questions are supplemented with explanations. Regarding the technology, the researchers utilized the best LLM, PubMEDBERT, to report their results. Other LLMs were BERT, SciBERT, and BioBERT. All the models used by LLMs have been trained on the MedMCQA dataset. The results show that PubMedBERT, which was pre-trained on PubMed abstracts, achieved only 47% accuracy. The authors conclude that further work needs to be done to automate medical reasoning as even the best LLM could make mistakes in 53% of cases. The key problem that arises is that multi-hop reasoning is heavily impeded by context retrieval using the current architectures. Furthermore, the arbitrary retrieval of arithmetic information was another reason why LLMs underperformed on the test set. It follows that while LLMs could offer good results, they still require much improvement to ensure human-level reasoning and performance in the medical QA field.

Lucas et al. [23] aimed at solving critical problems concerning the utilization of large language models in the medical field. Namely, the greatest pressure has been put on the inconsistency of reasoning presented by LLMs when answering specific medical questions. Classical prompting techniques provide diverse results, making them unreasonable for such areas as healthcare, which is related to high stakes. To eliminate the above issues, the researchers have proposed a new technique which is known as "ensemble reasoning". It uses several reasoning paths to refine answers, and ultimately enhance the overall accuracy and applicability of responses, as well as the reliability, trustworthiness, and responsibility of AI systems, especially when it comes to clinical reasoning across the medical field. For datasets, they have used USMLE questions which consist of MSQs. Such questions have been specially designed to determine a specialist's knowledge and its applicability across different competencies. With the help of this dataset, the researchers have an opportunity to verify the correspondence between LLMs' work and officially approved medical knowledge. The authors have concentrated their attention on three main LLM models, namely, GPT-3.5 turbo, GPT-4 turbo, and Med42-70B, which is considered to be an open-source clinical LLM. Having relied on different prompting techniques, which included the traditional "zero-shot" one, as well as "ensemble reasoning," the research has shown that the latter technique has notably improved the LLM's performance and consistency when it comes to the process of medical answering questions. Iterative answers' refinement is a critical advantage nowadays, as it can lead to more reliable presentation. Ultimately, these AI methods have proven to be far more efficient in clinical areas. For example, the GPT-3.5 turbo model has demonstrated a 3.44% accuracy increase in answering Step 1 questions, as it has been put into comparison with zero-shot chain-of-thought prompting. Moreover, qualitative analysis has demonstrated that AI reasoning has been not only correct but also helpful. In other words, they have shown a better understanding of the problem that should be analyzed by some specialists in healthcare areas. In conclusion, this research work is the foundation for further human-AI interactions in clinical areas. The goal is to become successful in decision-making, as AI-supported decisions can help clinicians provide the highest possible patient care.

Guillen-Grima et al. [14] examine the effectiveness of GPT-3.5 and GPT-4 in solving the Spanish Medical Entrance Examination (MIR). The authors aim to explore how these AI models could help train medical specialists by evaluating their ability to solve MIR exam questions across different medical specialties. This exam is a crucial step for healthcare workers in Spain to secure permanent residency. The study used the 2022 MIR exam, excluding questions related to imaging and questions with known errors. The final data set contained 182 multiple-choice questions that were available in both Spanish and English. GPT-4 significantly outperformed GPT-3.5, achieving an accuracy of 86.81% in the Spanish test and 87.91% in the English version, compared to GPT-3.5's accuracy of 63.18%. The GPT-4 excelled in theoretical questions and scored perfect scores in specialties such as pathology, orthopedic surgery, and dermatology. However, it struggled with practical disciplines such as pharmacology (40% accuracy) and critical care (33.3% accuracy). Despite its strengths, AI has shown limitations in solving image-based diagnostic problems, such as X-rays or MRIs, where human clinical expertise is key. The authors emphasized that although GPT-4 has a 0% critical error rate in life-threatening scenarios, it still poses a risk in other areas that require human supervision. They conclude that while GPT-4 and similar language models hold promise for supporting medical education and training, they are not yet able to replace human expertise in real-world patient care, particularly in areas involving subtle clinical reasoning and diagnostic imaging.

Wang et al. [17] showed how large language models such as GPT-4, MedPaLM2, and other AI models are helping the field of medicine and healthcare. Though there are some challenges like biases in models, data privacy, especially in the medical sector this causes a lot of damage. This paper shows how specialized medical AI's are helping to take critical decisions. Here for the datasets they have used 56 experimental datasets, including MedQA, USMLE, and PMC-LLaMA. These datasets focus on answering medical questions and summarizing clinical dialogue. Other data sets, such as files containing MRI, X-ray, and ultrasound data, are used to train LLMs to assist radiologists in diagnosing conditions. While some datasets are publicly available, others, particularly those containing sensitive patient data, are restricted for privacy reasons. In terms of technological progress, the article evaluates models such as GPT-4, MedPaLM2, BioGPT, and T5. GPT-4 excelled in passing the USMLE with a score of over 80%. MedPaLM2 demonstrated a 19% improvement in answering medical questions compared to its predecessor, MedPaLM. Despite these advances, the performance of LLM is still limited by the quality of available training data, and domain-specific fine-tuning is needed to address real-world challenges. In conclusion, while LLMs such as GPT-4 and MedPaLM2 show promising results, especially in diagnosis and decision-making, the article highlights persistent ethical, technical, and practical issues, including bias, lack of transparency, and interpretability of these models. To ensure their safe incorporation into healthcare, ongoing research is needed to improve data sets, develop better assessment metrics, and establish ethical standards to maintain patient safety and trust.

Li et al. [21] address the challenges of traditional medical student clinical training and also explore how LLMs can be used as virtual simulated patients. Stan-

dardized patients are traditionally human actors that simulate medical cases for student practice, but these bring problems such as high cost, inconsistent performance, and limited availability. The study proposes a CureFun framework that integrates LLMs such as GPT-3.5-turbo, PaLM, ERNIE-4, and Qwen-72B for SP simulation and offers a scalable, cost-effective, and consistent solution for clinical training. CureFun enhances the quality of dialogue between students and VSPs using Search Augmented Generation (RAG), which ensures coherent and contextually relevant interactions. The framework dynamically adapts to conversations and simulates diseases such as gastric disorders, diabetes, and pneumonia, across specialties such as pulmonology and endocrinology. Eight cases of SP were used for evaluation, although the dataset obtained from Wuhan Talent Information Technology Co., Ltd. is not publicly available. The study found that ERNIE-4, optimized for Chinese-language corpora, outperformed other LLMs in SP role simulation, achieving an improvement in B-ELO scores of 99.27 points. GPT-3.5-Turbo also improved by 250 points when integrated with CureFun. The automated rating system using CureFun showed a strong correlation with human raters, with Spearman's rank and Pearson's correlation coefficients around 0.81 and 0.85, respectively, indicating high agreement in medical student performance ratings. In addition to using LLMs as VSPs, the study explored their potential as virtual doctors. ChatGPT demonstrated strong diagnostic capabilities, although it was still outperformed by human professionals. Ultimately, research shows that while LLMs have the potential to revolutionize physician education by providing scalable, reliable, and emotionally consistent interactions with patients, further development is needed for them to effectively simulate real medical scenarios. The article concludes that LLMs could be valuable tools in helping physicians and training medical students to offer consistent and accessible clinical practice experiences.

Huang et al. [20]address critical challenges in the medical industry, including clinical decision support, medical data processing, research, education, and public health awareness. LLMs such as GPT-4, PaLM, and BioBERT have been used to improve clinical workflows. In medical education, these models help with exam preparation and content generation, while in public health, they help raise awareness through easily accessible information. However, potential bias, inaccuracies in medical recommendations, and ethical concerns remain an issue. The article highlights the need for specialized assessment frameworks to ensure the accuracy and reliability of the LLM in medical contexts. PubMed, MIMIC-III, and MedQA datasets have been used to help test the LLM's capabilities in tasks such as medical knowledge extraction, entity recognition, and summarization. MultiMedQA benchmark is publicly available, while others, especially those containing sensitive patient data, are restricted for privacy reasons. These datasets are essential for developing models that can process real-world clinical data and ensure compliance with privacy standards. Technologically, the article reviews various LLMs and highlights the GPT-4 for its superior performance in tasks such as answering medical questions and achieving high accuracy in medical screening tasks such as the USMLE. PaLM and its instructionally tuned variant Flan-PaLM have also demonstrated state-of-the-art results in clinical tasks. Open-source models such as LLaMA and OPT are praised for their adaptability and flexibility in medical research. Despite these advances, the paper notes that LLM performance varies across applications. While models such

as GPT-4 achieve high accuracy in medical screening, they struggle with finer tasks such as generating differential diagnoses and treatment plans. Sometimes LLMs generate inaccuracies in patient care, raising concerns about their ethical deployment in real-world settings. The article concludes that LLMs have transformative potential in healthcare, but to ensure their successful integration into clinical practice, issues such as bias mitigation, ethical compliance, privacy, and better evaluation frameworks need to be addressed.

Chapter 3

Work plan

As per the department's guidelines, our thesis is divided into three phases: prethesis one, Pre-Thesis two, and Final thesis defense. Each phase is four months long, so the whole thesis work is planned to be completed in one year. During the initial four months of pre-thesis one we have done several works, firstly we have formed a group of four members secondly we selected specific domains of our thesis: Artificial intelligence, Machine learning, and Natural language processing thirdly we confirmed a supervisor, fourthly with the advice of the supervisor we have selected a thesis topic: Efficacy of Large Language Models in facilitating exam preparation for Medical students. Fifthly, with the approval of our supervisor, we wrote a thesis abstract and submitted it for thesis registration, finally after completing the thesis registration and getting confirmation from the department we started scanning and skimming for relevant papers related to our selected topic. After selecting the relevant topics we started reading and summarizing the selected papers and included them in the Detailed Literature Review part of our Pre-thesis one report. Our second phase is Pre-thesis two where we prepared our key datasets from the medical guide books, after that using the dataset we ran and tested a model and compared that with other available models. At the end we prepared pre-thesis two reports and a poster for presentation. Our last phase is the Final thesis defense. In the final phase, we have developed a Retrieval Augmented Generation pipelined model. We have created a vector database for our RAG Model to work with. We have evaluated the performance of 11 Large language models and found Gemma 2 9B and Gemma 3 12B as the best-performing models, so we implemented the model in our RAG framework, and using the model, we have generated MCQs, their answers, and respective explanations. To improve the performance further, we have tried finetuning the Model Gemma 2 9B and Gemma 3 12B but due to limited computational resources, we were unable to do so after that, we have prepared our final thesis report and slides for our defense presentation and finally we took preparation for our thesis defense. A Gantt chart is given below which shows the task and the timeline of their completion and the chart is colored to clearly distinguish the three phases of our work. Pre-thesis one is denoted by Orange color, Pre-thesis Two is denoted by Teal Color, and Final thesis defense is denoted by Green Color.

Tasks	Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Month 7	Month 8	Month 9	Month 10	Month 11	Month 12
Group Formation	Pree-thesis one											
Domain Selection	Pree-thesis one	Pree-thesis one										
Supervisor Selection		Pree-thesis one										
Topic selection		Pree-thesis one										
Abstract writting and thesis registration		Pree-thesis one										
scanning and skimming for relevent papers			Pree-thesis one	Pree-thesis one								
Literature Review			Pree-thesis one	Pree-thesis one								
Pree-thesis one Report writing				Pree-thesis one								
Preparing dataset from Medical Guide Books					Pree-thesis two	Pree-thesis two						
Testing Performace of Various Models using Dataset						Pree-thesis two	Pree-thesis two					
Pree-thesis two report writing and poster Making								Pree-thesis two				
Pree-thesis two report writing and poster presentation								Pree-thesis two				
RAG Model Development									Final thesis defense	Final thesis defense	Final thesis defense	
Vector Database Creation using Medical Text Books										Final thesis defense		
Evaluating Various LLM's performance										Final thesis defense		
Finetuning Attempt											Final thesis defense	
Thesis paper writing & Final Thesis Defense preparation												Final thesis defense

Figure 3.1: Working Plan Gantt chart

Chapter 4

Dataset

4.1 Classification dataset

The dataset is based on the previous year's MCQ questions of the Prof-1 Exam of MBBS in Bangladesh. Prof-1 Exam is based on 3 subjects: Anatomy, Physiology and Biochemistry. We have collected these MCQ questions from the following guidebooks:

1. Arif's Representation on Medical Biochemistry Volume I [18]
2. Arif's Representation on Medical Biochemistry Volume II [19]
3. Arif's Representation on Human Physiology Volume I [8]
4. Arif's Representation on Human Physiology Volume II [9]
5. Arif's Representation on Human Anatomy Volume I [6]
6. Arif's Representation on Human Anatomy Volume II [7]

We have collected the MCQ and their solution from these books and made them formatted into text classification. Each MCQ has 5 options so we have converted them into 5 rows. There are a total of 2 columns in our dataset: Question_Option and Answer. In the Question_Option columns, we have merged the actual question and one option and in the Answer Column, We have put if the statement is TRUE or FALSE. So for each MCQ Problem, We have a total of 5 rows. Our dataset contains data from medical prof-1 exam questions of different years from 2006 to 2024.

4.1.1 Data Preprocessing

Our three datasets contained more or less 1,200 questions per subject. At first, we created three datasets, each containing columns "Question", "Option A", "Option B", "Option C", "Option D", "Option E", "True Options", and "False Options".

In order to reduce the complexity and the performance of the model we converted these Multiple Choice Questions into simple binary True-False statements using Python code. Each multiple-choice question was converted into five True-False statements corresponding to each of the options, resulting in a total of 6,015 statements in Anatomy, 6,055 in Physiology, and 5,120 in Biochemistry. At present, the

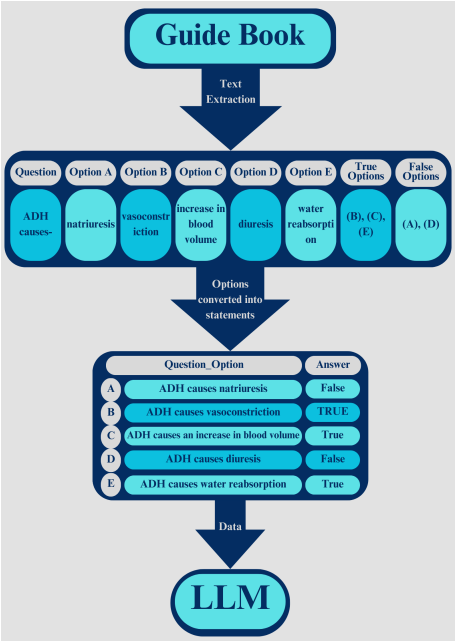


Figure 4.1: Data Preprocessing

dataset contains two columns: ”Question_Option” and ”Answer”. ”Question_Option” contains the statement, while “Answer” contains True or False depending on the statements.

We conducted model testing on Bio Clinical-BERT with three of our datasets. The three datasets were split into three distinct segments: training, testing, and validation, with 80% allocated for training, 10% for testing, and 10% for validation.

The anatomy dataset contained 6012 statements among which the training set consisted of 4,809 statements (79.99%), the validation set included 601 statements (10.00%), and the test set contained 602 statements (10.01%). The physiological dataset contained 6053 statements among which the training set consisted of 4842 statements (79.99%), the validation set included 605 statements (10.00%), and the test set contained 606 statements (10.01%). The physiological dataset contained 5116 statements among which the training set consisted of 4092 statements (79.99%), the validation set included 512 statements (10.00%), and the test set contained 512 statements (10.01%).

4.1.2 Data Analysis

We analyzed our three datasets and developed different graphs and bar charts. We developed Label Distribution bar charts, Text-Length Distribution bar charts, TF-IDF Scores, POS tag counts, and Named entity counts.

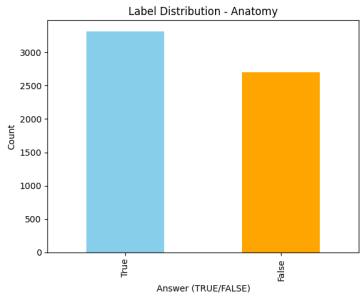


Figure 4.2: Anatomy

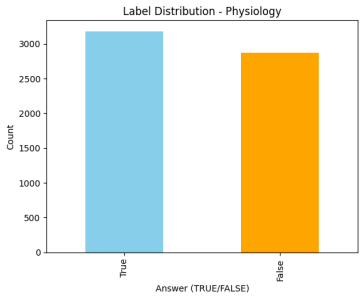


Figure 4.3: Physiology

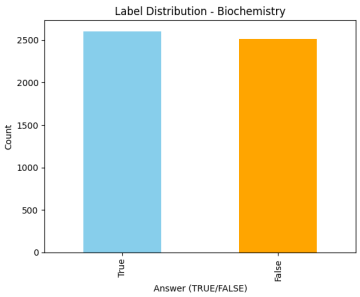


Figure 4.4: Biochemistry

Figure 4.5: Label Distribution

These three consecutive Bar charts [4.5] show the amount of True Statements and False Statements present in Anatomy Dataset, Physiology Dataset and Biochemistry Dataset. We have a total of 6,015 Anatomy data, 6,055 Physiology data, and 5,120 Biochemistry data. Among the three datasets, Anatomy contained 3314 True and 2698 False statements, Physiology contained 3179 True and 2874 false statements, while Biochemistry contained 2602 True and 2514 False statements.

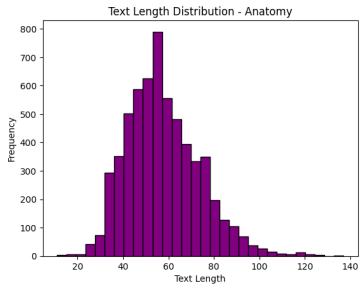


Figure 4.6: Anatomy

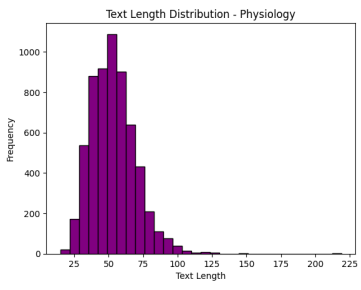


Figure 4.7: Physiology

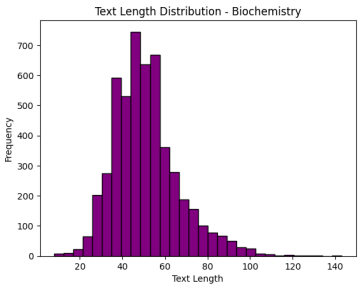


Figure 4.8: Caption 3

Figure 4.9: Text Length Distribution

The following figures [4.9] portray a comparative view on the text lengths of our three datasets. It shows text lengths of each True/False statement present in our Anatomy, Physiology and Biochemistry datasets. In the anatomy dataset, minimum text length is 11, maximum text length is 147 and average text length is 57. On the other hand, the physiology dataset had a minimum text length of 15, maximum text length of 219 with an average text length of 53. Lately, the biochemistry dataset holds a minimum text length 8, maximum text length of 144, and average text length of 51. Physiology dataset having the highest frequency for text lengths

and biochemistry dataset the lowest.

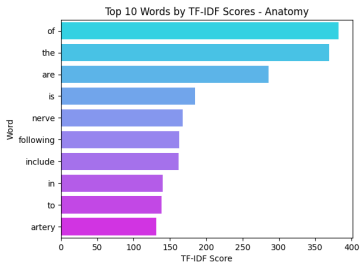


Figure 4.10: Anatomy

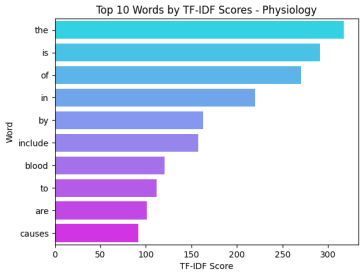


Figure 4.11: Physiology

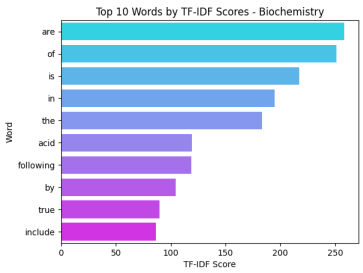


Figure 4.12: Biochemistry

Figure 4.13: TF-IDF Scores

These three bar-charts [4.13] show which words are most significantly used in these three datasets based on their Term Frequency-Inverse Document Frequency (TF-IDF) scores.The TF-IDF analysis shows 2773 unique words used across the anatomy dataset, 3020 across the physiology dataset and 2769 across the biochemistry dataset. These figures only portray the top 10 frequently used words in those dataset. In the anatomy dataset, the most used top 10 words are: 'of', 'the', 'are', 'is', 'nerve', 'following', 'include', 'in', 'to', 'artery'. Whereas the most commonly used words in physiology dataset are: 'the', 'is', 'of', 'in', 'by', 'include', 'blood', 'to', 'are', 'causes'. Lastly, in the biochemistry dataset, most frequently used words are: 'are', 'of', 'is', 'in', 'the', 'acid', 'following', 'by', 'true', 'include'. We can also notice medical health related terms like "nerve", "artery" in anatomy dataset, "blood", "causes" in physiology dataset and "acid", "include" in biochemistry dataset.

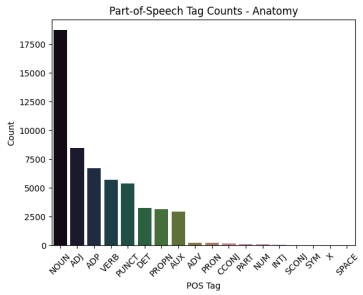


Figure 4.14: Anatomy

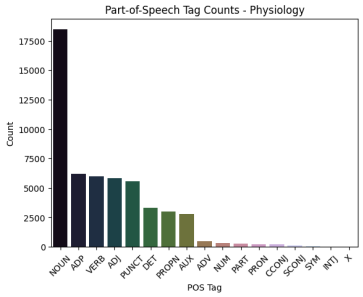


Figure 4.15: Physiology

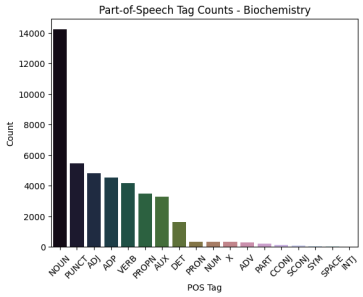


Figure 4.16: Biochemistry

Figure 4.17: Part-Of-Speech Tag Counts

These three bar-charts [4.17] provide a visual representation on the Parts of speech tags of each of the three datasets.In all three dataset, count of nouns is the highest. That leaves us with a clear view that medical related data always contains names of medical health issues, human organ names and other medical products.

These figures [4.21] here represents the real word concept related words and phrases

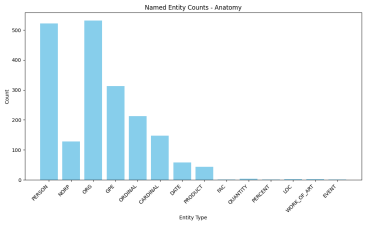


Figure 4.18: Anatomy

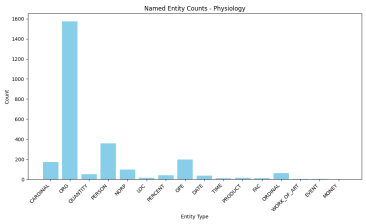


Figure 4.19: Physiology

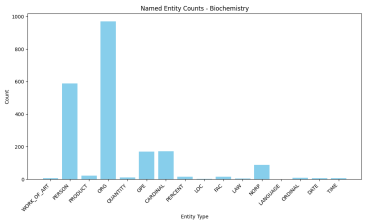


Figure 4.20: Biochemistry

Figure 4.21: Named Entity Counts

present in the three datasets. The real-world words are represented as work_of_art, person, product, org (organizations), quantity, gpe (geopolitical entities), cardinal, percent, loc (locations), fac (facilities), law, norp (nationalities/religions), language, ordinal, date, time. It indicates the most frequent mentions of organization and individuals across all three datasets.

4.2 Generation Dataset

As we intend to build a Retriever Augmented Generation Model, we selected few books based on the first year medical curriculum. These books are:

1. BD Chaurasia Brain Neuroanatomy Vol 4 8th Ed[2],
2. BD Chaurasia Lower Limb, Abdomen Pelvis Vol 2 8th Ed[3],
3. BD Chaurasia Upper Limb Thorax Vol 1 8th Ed[4],
4. BD Chaurasia's Human Anatomy Volume 3 Head Neck and Brain 6th Edition[5],
5. Lippincott Williams Wilkins[28]
6. Ganong's Review of Medical Physiology 26th Ed[1],
7. Guyton and Hall Textbook of Medical Physiology, 14th Edition[29]

The first four books listed here are for Anatomy. The fifth book is for Biochemistry, and the last two books are for Physiology. These books are widely used in medical education settings worldwide. These books hold significant importance to medical students, preparing for the first professional examination in Bangladesh.

Chapter 5

Methodology

5.1 Model - BioClinicalBERT

In the medical sector and healthcare field, text often has complex terminology and meaning that also varies the context it appears in. To understand and classify those texts we used **BioClinicalBERT**, a language model designed specifically for biomedical and clinical text.

BioClinicalBERT is a specialized model of the well-known BERT model. It was trained using medical and clinical documents which helps it to understand the technical words and phrases. There are many situations where the same word can be processed as a different meaning like "positive" might mean a positive test result in one sentence but something different in another. The bi-directional process of BioClinicalBERT makes the word choice process more accurate and reliable.

5.1.1 Tokenization: Preparing Text for the Model

Before training the datasets need to be converted into a format so that the model can understand. Using bert tokenizer this model breaks words into smaller pieces called tokens. It helps the model to be more flexible and able to understand unfamiliar words/terms. All the tokenized inputs are set to a maximum length of 512 tokens (words or subwords) so that all the inputs correctly fit in the model size limit. Few special markers like [CLS] (start of the input) and [SEP] (end of the input) were added to help the model process the text correctly. The [CLS] token's output is later used for the final classification.

5.1.2 Training and Loss Trends

For training we used label biochemical data sets. These trains help the model to learn to classify text to minimize a measure called loss (a number that shows how far the model's predictions are from the correct answers). A very commonly used categorical cross entropy loss classification method was used. The Adam optimizer

was used to adjust the learning rate of the model to make training more efficient and stable. We monitored the **training loss** (how well it learns from the training data) and the **validation loss** (how well it performs on unseen data).

5.1.3 Evaluating Performance: Confusion Matrix and Metrics

After training we tested the model that performs when we inject unseen data sets. A confusion matrix was generated to show how the model predicted and the correct categories and where it made mistakes. We also measured the model performance using metrics like precision (how many of its predictions were corrected). Here we used precision (how many of its predictions were correct), **recall** (how many actual cases it found), and **F1-score** (a balance between precision and recall). All these were visualized in charts to show how the model performs across different categories. For the biochemistry category, the model achieved high precision and recall, especially in identifying rare but important terms.

5.1.4 Model Architecture and Regularization

The architecture of BioClinicalBERT has:

- **12 Layers of Transformer Encoders**: These layers help the model understand the relationships between words in the input.
- **Dense Output Layer**: This layer gives the final prediction for each input, assigning probabilities to the different categories.

To improve performance and overfitting problems, we used dropout layers during training. Dropout method randomly deactivated some parts of the model while it was learning, which helps to generalize better to new data. We also used gradient clipping to avoid large updates that could establish training.

5.1.5 Results and Observations

BioClinicalBERT performed very well compared to other models like BiLSTM and general purpose BERT model. Since this model already has preliminary ideas about medical terminology it can handle any missing or falty words. The confusion matrix showed the accurate classified result for most of the categories, though it struggled a bit with the categories that had overlapping and ambiguous terms. For example, it sometimes confused “Symptoms” with “Conditions” because they share similar language.
Overall, the results showed that **BioClinicalBERT** is highly effective in understanding and classifying biomedical texts. Its capability to capture both general patterns and specific details makes it a great choice for tasks in the clinical and biomedical field.

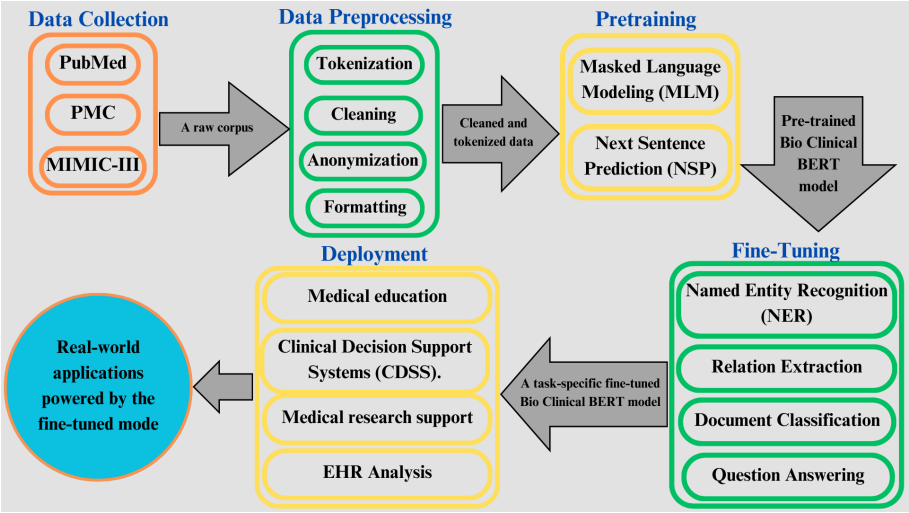


Figure 5.1: Working of Bio-Clinical BERT

5.2 Architecture of the Retriever System

This chapter details the comprehensive methodology used to design, build, and evaluate an interactive learning system for first-year medical students. The primary goal of this research was to move beyond a simple question-answering tool and create a more holistic educational assistant. This system is built upon a Retrieval-Augmented Generation (RAG) framework, which leverages the rich knowledge contained in core medical textbooks. The process of our work was divided into several logical phases. We began by building the foundational knowledge base from the textbooks. Next, we constructed the core RAG architecture and conducted a rigorous evaluation of eleven different Large Language Models (LLMs) to find the most suitable one for our medical context. Finally, using the best-performing model, we developed advanced features for student self-assessment: an automated Multiple-Choice Question (MCQ) generator and a timed examination module. This chapter explains each of these steps in detail.

5.2.1 Building the Knowledge Core

The calibre of the data that our system can access determines how effective it is. Thus, carefully extracting the information from the book and organising the knowledge from our source materials was the first step. The knowledge source was typical, first-year medical textbooks in the three foundation areas: Anatomy, Biochemistry, and Physiology. To provide the specifics from the textbooks to our system, we accessed the complete text of the books as PDFs. We used a proprietary script to extract text from the PDF files by reading the text from each page systematically. This preliminary process converted these static books into a dynamic source of information that our system could use. There are physical limits to what a computer, especially a language model, can digest at any one time, so clearly the text could not be processed as a full book. The text had to be semantically broken down into smaller, significant pieces. To this end, we relied on a method referred to as ”semantic chunking”. Rather than divide text into random blocks of a certain size, we made the decision to divide the text according to sentence structure, so we broke

the text after each sentence, and created chunks of roughly 500 characters. This is a preferred model for this reason: by maintaining related ideas as chunks in whole sentences, we kept the associated symbolic meaning and context, instead of randomly chunking sentences together, which would disrupt the associated meaning in reading the text. Additionally to each of these chunks, was the inclusion of file metadata, the filename of the book and page number the text was collected from. This is an important element in a knowledge system, because it permits the system and user to retrieve any piece of information to trace it back to source

5.2.2 Designing the RAG Architecture

Once our knowledge base was prepared, we built the core engine of our system. This architecture is responsible for finding relevant information and using it to answer questions.

5.2.3 Creating a Searchable Vector Database

In order to allow our text chunks to be searched for meaning, we needed them to be machine readable mathematically. This required two steps:

1. Text-to-Vector Transform: We used a sentence-transformer model called all-MiniLM-L6-v2. This is a neural network that reads a piece of text of a certain number of words and will then convert that piece of text into a list of numbers, called a "vector embedding". This vector represents the semantic meaning of the text. We created one vector embedding for every text chunk out of our text books.

2. Indexing for Fast Retrieval: All of the vector embeddings were then indexed into a database called FAISS (Facebook AI Similarity Search). A FAISS index is optimized to quickly find the most similar vectors together. This allows the system to group the text chunks nearest to the meaning of the question almost instantly, when the user asks a question.

5.2.4 The Retrieval and Generation Pipeline

1. Now that we have a database, the system could begin answering questions. After the user sorts out their query (and what we began testing with were True/False statements), the system performs the following sequence:

2. Finding Relevant Information (Retrieval): First, the user's question is turned into a vector embedding for the same reason that the textbook chunks were turned into embeddings. Then, the system uses the FAISS index to find the text chunks with embeddings that are geometrically most similar to the question's embedding. This results in a small collection of the most relevant paragraphs from the textbooks.

3. Formulating an Answer (Generation): The first step leads to a small collection of chunks, and then we concatenate these chunks into a single blob of context. This context blob, combined with the user's original question, is then sent

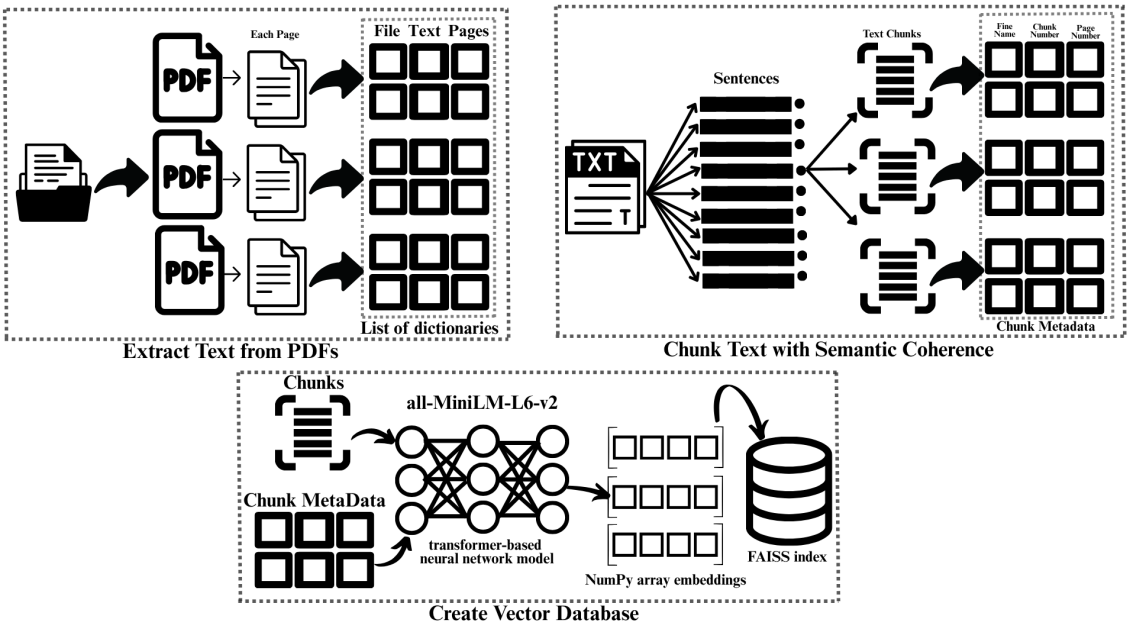


Figure 5.2: RAG Workflow Pipeline - PDF to Vector Database

to a large language model. We instructed the model explicitly: "You are a medical expert and you must use only the text you were given to determine if the statement is True or False." Your goal is just to output one of two words: "True" or "False."

5.3 Comparison of Large Language Models

The element of our system that is most akin to a "brain" is the LLM that produces the final answer. Selecting the right one was an essential step in our process, and we conducted an in-depth comparison of several different models.
We ran eleven different models on our datasets:

1. Gemma2:9b,
2. Gemma3:4b,
3. Gemma3:12b,
4. Llama2:7b,
5. Medllama2:7b,
6. Llama3.1:8b,
7. Llama3.2,
8. Mistral:7b,
9. Deepseek-r1:8b,
10. Qwen3:8b,

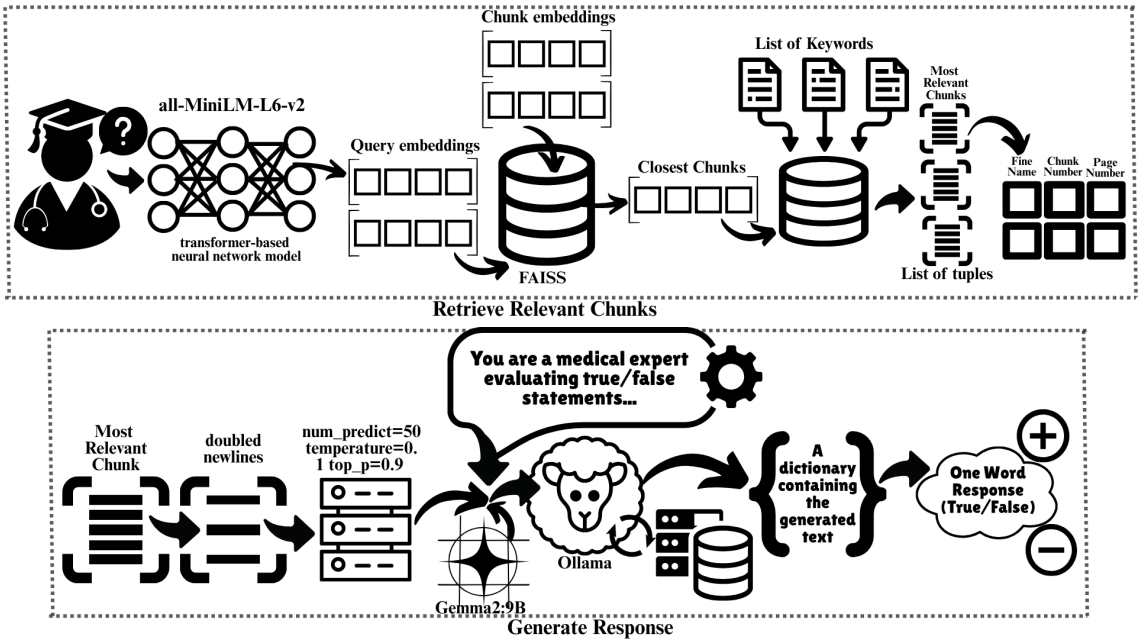


Figure 5.3: RAG Workflow Pipeline - Generating Response

11. Meditron:7b

This was a mix of general-purpose models and 2 models that were trained on medical data specifically (medllama2 and meditron). We assessed the performance of each LLM with True/False questions from our Anatomy, Biochemistry, and Physiology datasets. We evaluated the results using two common metrics:

- Accuracy: Simply what percentage of the questions the model answered correctly.
- F1-Score: A more balanced assessment that looks at a models precision on "True" predictions along with its ability to capture all the actual "True" statements. It provides a more complete assessment of performance than accuracy alone.

Table 5.1, 5.3 and 5.5 show the Accuracy and F1-Scores for each of the 11 models evaluated across the Anatomy, Biochemistry, and Physiology datasets.
As the data in the table and the charts below clearly show, the Gemma2:9b and Gemma3:12b models were the top performers. They were significantly more accurate and reliable than the other models, including those supposedly specialized for medicine. Here, we select gemma2:9b as the engine for our system. Although gemma3:12b was also a good model, gemma2:9b provided first-rate performance and was just slightly more efficient with its processing. To better understand the behavior of gemma2:9b, we reviewed the confusion matrices in Figure 5.6 that captured its correct and incorrect answers for each subject.

5.3.1 Performance Analysis

It is visible in table 5.2, 5.4 and 5.6 that the f1 score for medllama2, qwen3:8b, deepseek-r1:8b, and meditron:7b are comparatively very low with respect to other

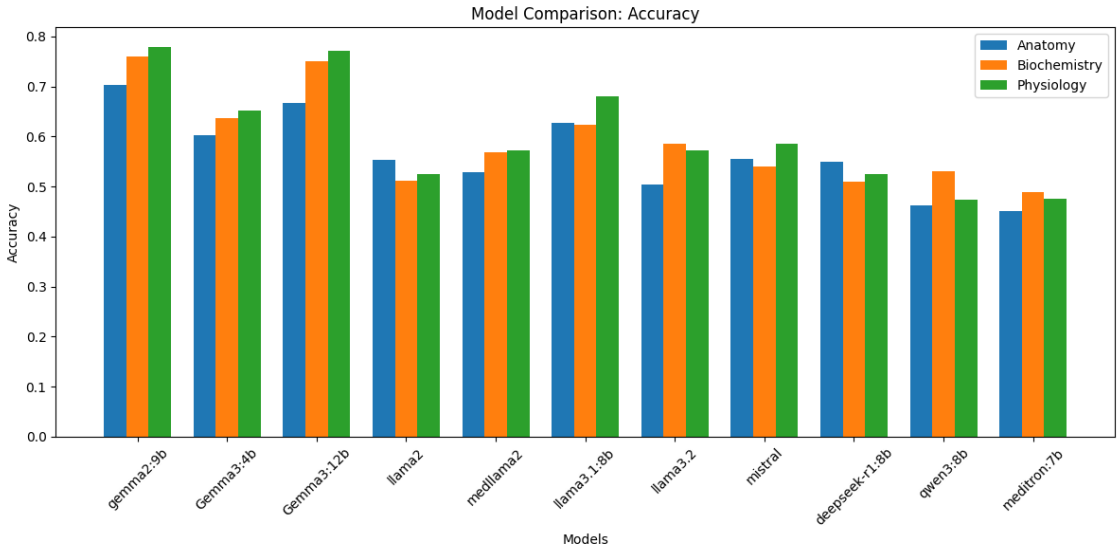


Figure 5.4: Model Comparison - Accuracy

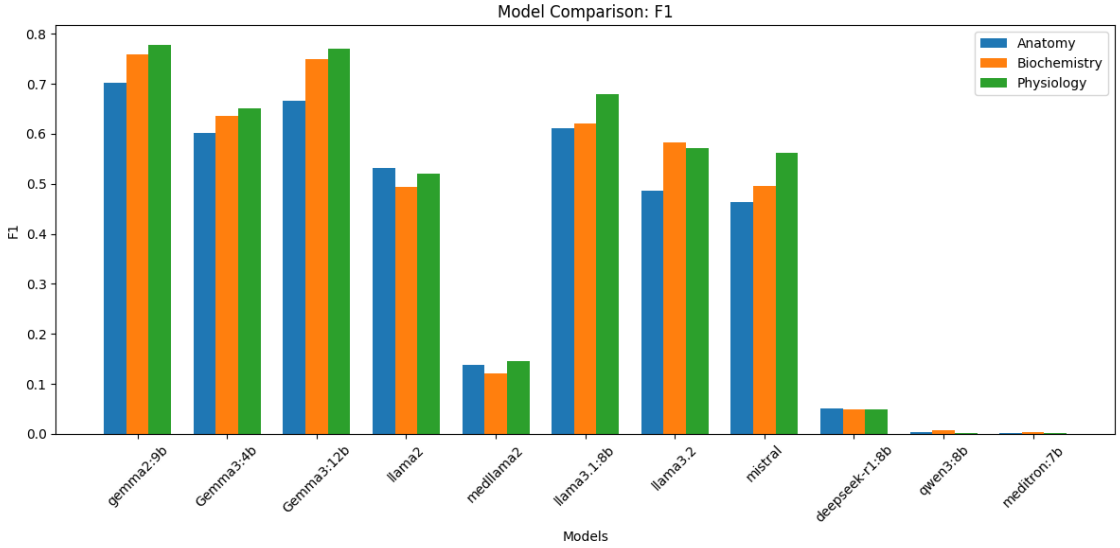


Figure 5.5: Model Comparison - F1 score

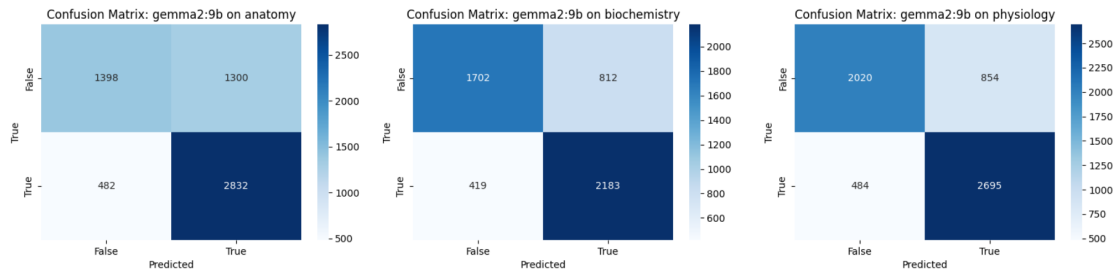


Figure 5.6: Confusion Matrix - Gemma2:9B with Retriever System

Model	Anatomy F1	Anatomy Accuracy
gemma2:9b	0.702761	0.703593
Gemma3:4b	0.601297	0.601219
Gemma3:12b	0.665336	0.666001
llama2	0.531131	0.552229
medllama2	0.138745	0.12915
llama3.1:8b	0.610668	0.626747
llama3.2	0.486096	0.530216
mistral	0.463579	0.49555
deepske1:8b	0.027445	0.04694
qwen3:8b	0.002651	0.469208
meditron:7b	0.000862	0.450432

Table 5.1: Performance metrics for various models - Anatomy.

Model	**FALSE**			**TRUE**		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
gemma2:9b	0.743617	0.518162	0.610747	0.685382	0.854556	0.760677
Gemma3:4b	0.584344	0.392884	0.469858	0.608914	0.772448	0.681576
Gemma3:12b	0.630666	0.744266	0.666777	0.743294	0.601994	0.665222
llama2	0.525	0.020833	0.040174	0.552783	0.9828	0.707527
medllama2	0.469331	0.374351	0.416495	0.562694	0.655401	0.605582
llama3:1.8b	0.44089	0.117293	0.184963	0.612112	0.841172	0.707123
llama3:2	0.270889	0.143455	0.203627	0.638442	0.22752	0.335484
mistral	0.480956	0.3028	0.369545	0.565147	0.852273	0.678954
deepseeker:1-8b	0.430556	0.011149	0.022383	0.5101	0.987628	0.674937
qwen:3.8b	0.490595	0.090403	0.154496	0.570481	0.103802	0.176748
meditron:7b	0.484766	0.383692	0.616349	0.55102	0.016295	0.031653

Table 5.2: Anatomy Precision, Recall and F1 scores

models. This is due to several reasons. First of all, the F1 score is calculated using token-based overlap for per question between predicted and true answers. Non-standard responses produced by Meditron:7B and Qwen3"8B have no overlap with True or False, causing very small F1 scores. The second reason can be that these models fail to produce the required single-word output, which is True or False. In table 5.7, the responses of different models are shown for an Anatomy False statement "Rectus Sheath Contains: Cremaster muscle". Medllama2 generates verbose responses, which include extra tokens that lower the precision in the token-based F1 score. Qwen3:8B and deepseek-r1:8B provide responses with <think> tag, which lacks clear True or False answers and leads to a default False response. The table shows that Meditron:7B produces irrelevant responses, ignores the prompt, resulting in default False predictions. Unrecognized responses lead to default False predictions which cause false negatives for True questions and low True Recall. Moreover, general-purpose models like qwen3:8b and deepseek-r1:8b show poor performance as they lack medical-specific knowledge. Medical models like medllama2 and meditron:7b are not fine-tuned for True-False questions, which reduces their accuracy. Lastly, Ollama's generation settings num-predict=50, encourages longer outputs.

Model	Biochemistry F1	Biochemistry Accuracy
gemma2:9b	0.75821	0.759382
Gemma3:4b	0.635066	0.636044
Gemma3:12b	0.750011	0.751173
llama2	0.493676	0.511923
medllama2	0.512105	0.567435
llama3.1:8b	0.62045	0.623534
llama3.2	0.583151	0.586173
mistral	0.510765	0.540657
deepske1:8b	0.484705	0.475342
qwen3:8b	0.007738	0.529711
meditron:7b	0.003623	0.487881

Table 5.3: Performance metrics for various models - Biochemistry.

Model	**FALSE**			**TRUE**		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
gemma2:9b	0.802452	0.677009	0.734412	0.728881	0.83897	0.780061
Gemma3:4b	0.699755	0.454256	0.550892	0.6062	0.811683	0.694052
Gemma3:12b	0.718102	0.812649	0.762456	0.792062	0.691776	0.738765
llama2	0.506824	0.043124	0.095951	0.510553	0.679652	0.582747
medllama2	0.574764	0.460223	0.511155	0.562681	0.671022	0.612463
llama3:1.8b	0.784333	0.003203	0.006379	0.580283	0.914947	0.711507
llama3.2	0.552919	0.829488	0.662901	0.677656	0.355496	0.466347
mistral	0.622385	0.155927	0.250587	0.528035	0.912375	0.671067
deepseerk:1.8b	0.516129	0.012729	0.024845	0.508904	0.988457	0.671891
qwen3:8b	0.512127	0.507273	0.509591	0.640805	0.171407	0.26811
meditron:7b	0.488921	0.930339	0.640998	0.472892	0.060338	0.107021

Table 5.4: Biochemistry Precision, Recall and F1 scores

Which is why the models ignored the prompt's strict one-word requirement.

5.4 Creating Multiple Choice Questions

We constructed a feature where the system could automatically create practice questions for students. When a student selects a subject and chapter, the system discusses the relevant text chunks from the knowledge base and once retrieved, sends the chunks to the gemma2:9b model addressed with a prompt. The prompt told the model to use similar text to form a unique, quality multiple-choice question with five options. The prompt also instructed the model to identify the correct answer(s) and to provide an in depth response as to why that answer was correct according to the textbook information provided. To further promote uniqueness, the system keeps track of the questions it has already generated and always attempts to create a new question.

Model	Physiology F1	Physiology Accuracy
gemma2:9b	0.778127	0.778953
Gemma3:4b	0.651082	0.651908
Gemma3:12b	0.769866	0.770699
llama2	0.520395	0.522093
medllama2	0.456931	0.451245
llama3.1:8b	0.678672	0.679218
llama3.2	0.566342	0.565913
mistral	0.592262	0.583533
deepske1:8b	0.58312	0.584333
qwen3:8b	0.470061	0.471354
meditron:7b	0.000996	0.474806

Table 5.5: Performance metrics for various models - Physiology.

Model	**FALSE**			**TRUE**		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
gemma2:9b	0.743617	0.518162	0.610747	0.685382	0.854556	0.760677
Gemma3:4b	0.584344	0.392884	0.469858	0.608914	0.772448	0.681576
Gemma3:12b	0.630566	0.744266	0.666777	0.743294	0.601996	0.665222
llama2	0.525	0.043206	0.079873	0.552783	0.9828	0.707582
medllama2	0.469331	0.373451	0.416495	0.562694	0.655401	0.605582
llama3:1.8b	0.6208	0.134754	0.221727	0.611212	0.481127	0.537542
llama3.2	0.470809	0.14335	0.220367	0.638442	0.22752	0.335484
mistral	0.502326	0.232008	0.315292	0.565417	0.452735	0.502111
depseeker:1:8b	0.430556	0.011149	0.022383	0.5101	0.987682	0.673937
qwen3:8b	0.449095	0.094003	0.15667	0.570481	0.130882	0.212227
meditrton:7b	0.484766	0.386392	0.616349	0.55102	0.016295	0.031653

Table 5.6: Physiology Precision, Recall and F1 scores

Model	Response
gemma2:9b	TRUE
Gemma3:4b	TRUE
Gemma3:12b	TRUE
llama2	TRUE
medllama2	Rectus sheath contains cremaster muscle.
llama3.1:8b	FALSE
llama3.2	TRUE
mistral	FALSE
deepske1:8b	<think> Okay, so I need to determine whether the statement whether the statement "Rectus Sheath Contains: Cremaster muscle" is true or false based on the provided context. Let me go through the information given step by step. First, looking at the
qwen3:8b	"<think> Okay, let's tackle this question. The statement is: ""Rectus Sheath Contains: Cremaster muscle. "" The context provided is from . some anatomy references. First, I need to recall what the rectus sheath is."
meditron:7b	A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions.

Table 5.7: Different Model Responses

Chapter 6

Implementation

6.1 Overview on system Development

We have used python to develop our system because it has wide libraries of machine learning and natural language processing. We have made our system development in a Local workstation of our Research Lab which contains an Intel i7 14700K CPU with 64 GB ram and paired with NVIDIA RTX 4080 Super GPU with 16GB vram. Although such hardware enabled sufficient computational power to evaluate and infer the models, it showed limitations when large-scale tasks (e.g. fine-tuning of large language models) needed to be put into action.

The system architecture in general was designed in such a way that there were three major entities: the retriever, the generator and the evaluator. The retriever component was the one that extracted the textual data (in PDF format) in medical textbooks, semantically clustered the data, embed segments on it and indexed the data to a searchable vector database. The generator (called Gemma 2 9B, Gemma 3 12B and others) was a large language model (LLM) that answered true/false medical questions and produced multiple-choice questions (MCQs) on the basis of retrieved relevant information. Lastly, the evaluator module was established and would evaluate fitness of model-generated response on modeled datasets with curated ground

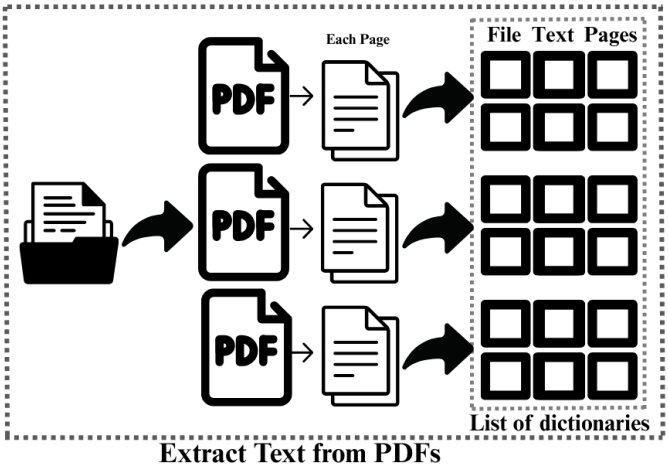


Figure 6.1: Extract text from PDF

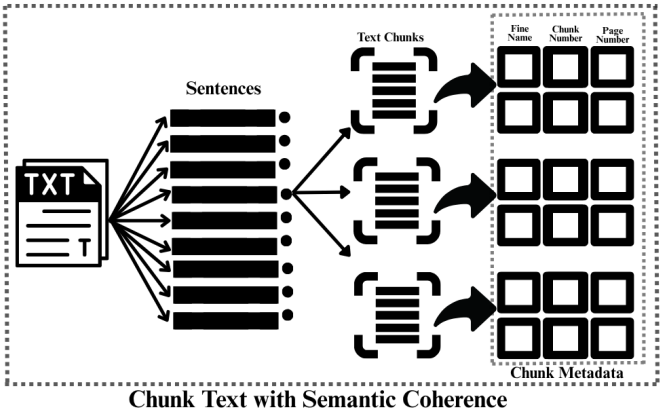


Figure 6.2: Chunk Text

truths of other topics that included anatomy, biochemistry, and physiology.

This was set up with the help of a semantic retrieval pipeline where in the system could simulate reality scenarios of an open-book examination. One could imagine the user entering a query, whereby retrieval would bring forth the most contextually relevant bits of text after which they would be feed into a prompt template and fed into the LLM to obtain an answer. This scheme was the focus of our objective of making uses of retrieval-augmented generation (RAG) to improve the way medical exams are prepared.

In order to achieve the modularity and maintainability objectives, every functional line of the system was formulated so that it could be embodied in separate but reusable Python functions. As an example, the task of extracting text was performed with the help of a function extract-text-from-pdfs(), whereas chunking and embedding were performed by chunk-text() and the Sentence Transformers encoder, correspondingly. The FAISS library implemented in Facebook AI was to perform the vector indexing process with L2 distance as the metric of similarity search. This modularity also enabled us to play around freely with the various models, the chunk sizes, the prompts and never actually had to touch the system logic.

The sorting algorithm of the chunks retrieved was added with the similarity of embeddings as well as the keyword scoring process strategy. This was done to make sure that not only were the retrieved chunks semantically close to the query, but they were also dense with presence of medical keywords thus making them more contextually relevant. On the whole, this mixed scoring method was very important as it helped the quality of the responses of the model.

6.2 Data Preprocessing and Formatting

Data extraction and preparation consisted of the first step of our implementation pipeline; the use of medical textbooks in PDF form and the use of structured datasets in the form of true/false medical questions were two main sources of data used. Preprocessing was extremely crucial to provide the downstream parts with a pure,

coherent, non-context-irrelevant input.

First of all, we worked on the task of extracting text based on a pile of medical textbooks by means of PyPDF2 library. The purpose behind the creation of the extract-text-from-pdfs() is to also be able to Go through its directory and find all files with PDF extensions and read through every single page of the files. In the case of the encrypted PDFs, one tried to open them by typing the password as empty, and it worked out in the majority of cases. When a file was not decrypted, no notice was made and the file was simply skipped by making a proper log message. The text on each page was automatically retrieved and saved together with the page number of that page which meant that we were in a position to maintain the structural mapping of the content and its location in the textbook. This coding was crucial in subsequent phases, especially when one is trying to explain why he or she is giving answers that have been given by the model by making reference to the pages in the textbook.

After we had the raw text we shifted towards the chunking stage with the chunk-text() method. We did not just divide the text into two identical parts as per the number of characters, but wisely divided the parts maintaining the meaning. To begin with, we separate the text into sentences with the aid of periods. After that we collect sentences in chunks until they have 500 characters, then we split one more sentence. The approach was useful to make each chunk meaningful and self-sustaining instead of whether using chance and disjointed. The chunks were saved as well with such metadata along with the filename, the number of the chunk itself, and the page where it was placed in, to facilitate further indexing.

Parallelly, we used three structured datasets namely - anatomy, biochemistry and physiology. These data were in CSV files, and each column in the dataset comprised thousands of true/false statements along with true/false labels which were the ground truths. We executed a set of preprocessing operations with the use of function load-datasets(). These consisted of selective verification of the presence of files, loading the CSV files into pandas Data Frames and a uniform standardization of the column Answer by converting representations of true and false ('TRUE', 'False', 1, 0) to binary integers. They eliminated any of the rows with missing values of the column of 'Question-Option' or 'Answer'. The datasets after its cleaning were stored and used in the generation and evaluation stages.

6.3 Vector Database Creation

After chunking and proper annotation of the textual data contained in medical textbooks, the other important procedure in the implementation process would be to convert the text chunks into semantic embeddings and store the data in a vector index. This step became the core of our system with regards to retrieving information, we were putting into practice the concept of information retrieval on the basis of similarity of query and retrieval of the information corresponding to the context.

The first part of this stage was to rely on the all-MiniLM-L6-v2 model, which was found in the library for sentence transformers. This model is effective and performs

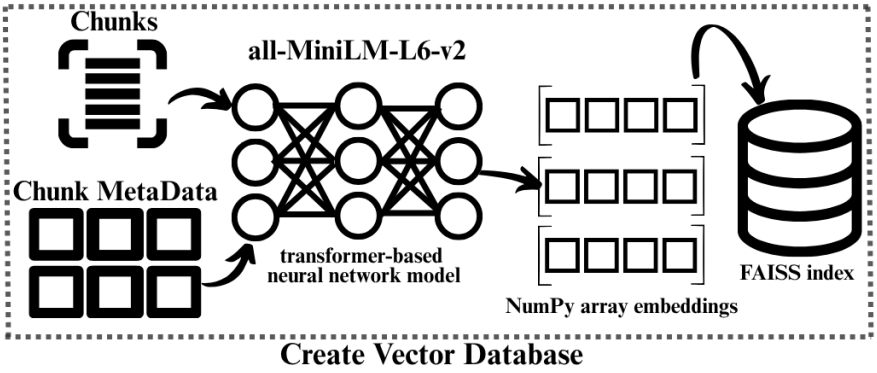


Figure 6.3: Vector Database Creation

well in the tasks of semantic similarity and, therefore, was chosen to convert our text chunks to high-dimensional numeric form. After encoding each of the chunks, it was represented by a dense vector in a fixed-dimensional space, its semantic content. According to the embeddings, the comparison of the query statements against pre-indexed knowledge could be carried out in a mathematically meaningful form.

Because of a large number of these vectors, we used FAISS (Facebook AI Similarity Search), the library that supports high-performance similarity searches on large-scale vector datasets. With the help of the FAISS IndexFlatL 2 structure, we indexed all the embedding of the chunks, by their L 2 (Euclidean) distances. The fact is that the L2 distance can be used because it is effective in scenarios of semantic retrieval when generally the smaller the distance, the more relevant the elements become.

In process of indexing we used to add all the embeddings into one NumPy array and construct the FAISS index using the insertion of these vectors. At the same time, we maintained the associated metadata, such as filenames, chunk identifier and page identifier, in such a parallel nature. This allowed letting us trace any vector that is retrieved back to the original chunk and source document it was derived of, preserving traceability along the retrieval and generation pipeline.

6.4 Response Generation

Besides the pure vectors similarity, we added some keyword based scoring system so that more semantic retrieval can be done. On each query of the system, we selected a number of domain-specific keywords e.g. metabolism, hormone or nerve and recorded their number in the chunks considered a candidate. The score obtained using this keyword was joined with the 1/L2 distance in order to generate an end ranking. The semantically similar and terminologically rich chunks had the priority, and their responsiveness was enhanced to ensure that the model made proper and context-differentiated answers.

A combination of semantic embeddings-based approach and keyword enrichment

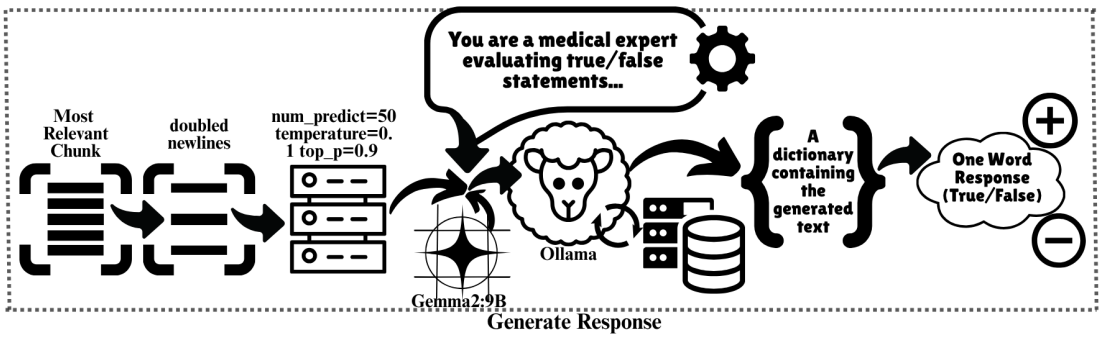


Figure 6.4: Generating Response

was an important step to have an effective retrieval mechanism supporting successive tasks of answers classification and question generation. This acting as a vector database enabled this system to spatially search very quickly in thousands of pieces of medical knowledge and in an appropriate context which is vital in supporting this research henceforth.

The main task of the response generation system was an assessment of whether a certain medical statement was false or true, depending on pretextual evidence found in medical articles. This step was essential in the development of an assisted exam preparation model that seemed realistic as it would be able to justify the answers based on the textbook information.

In answering each of the questions across anatomy, biochemistry, or physiology dataset, our system made use of the FAISS index and keyword-augmented scoring mechanism mentioned earlier to obtain the top related chunks. The most appropriate in terms of context-wise chunks of the text were then concatenated into a single unified context-block. The prompt that was given to the large language model was based on the retrieved context.

We have used a well-thought prompt template that informed the model to act as a medical specialist. It contained the statement under consideration, the background that justifies it and the directive of returning only one among two possible outputs: either True or False. This structured format guarantee the compatible outputs that could be compared literally with ground truth labels. One of the common instructions was the one that challenged the model to analyze the given statement using the medical context provided. answer only with either True or False".

The generation has been carried out with the help of Ollama interface, where the chosen language model was used, e.g. Gemma 2 9B. The generational parameters were well adjusted so as to give preference to deterministic behavior. The temperature value we set did not go too high, 0.1 randomness, and the sample we sampled was top-p 0.9 which helped to avoid as much random output generation as possible. The response of the model was analyzed and deprived of whitespace, and the model took in only valid binary answers to evaluate.

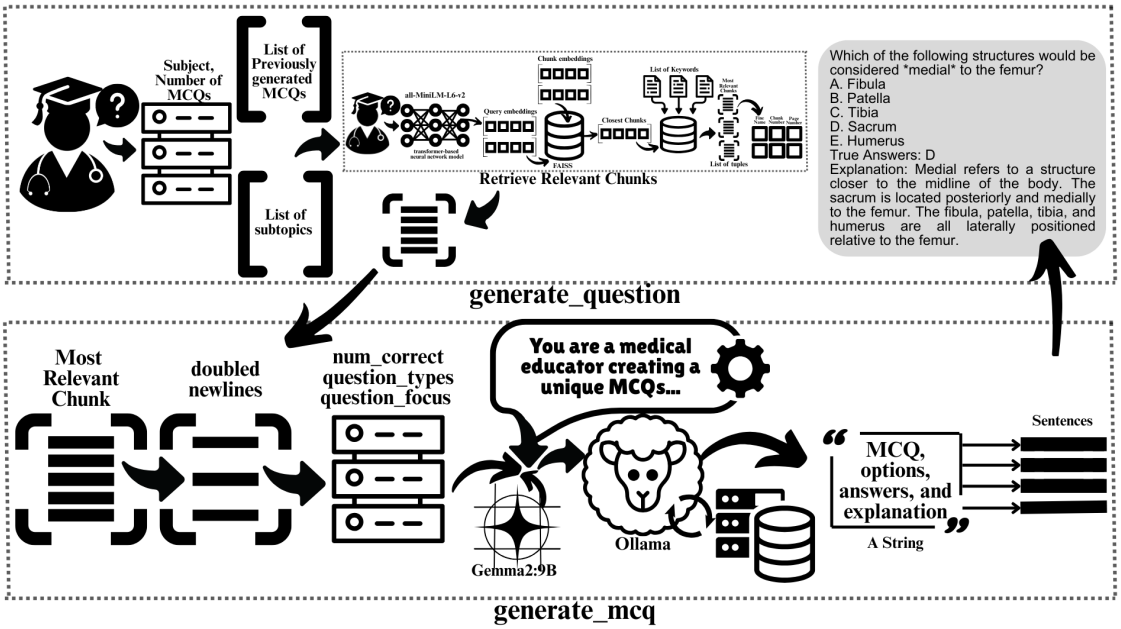


Figure 6.5: Generation of Multiple choice Questions

In case the model output corresponded to known variations of True or False (including insignificant spelling variations), it was turned to binary value. In case the output was non-specific or invalid, then it was eliminated in the consideration. The step helped in maintaining integrity of performance measurements and that only answers that were confidently interpreted were analyzed.

During this process we stored the original statement, the context retrieved and the response provided by the model, together with the correct label, in structured CSV logs. Such logs also enabled transparency and traceability to be inspected at later point in time and debugging of models. We also calculated per-sample precision, recall, and F1 with token-based comparisons, and this aspect improved our performance analytics even more.

6.5 Generating Module of MCQ

To help students prepare better we have introduced MCQ and True False Generation in our system using same RAG pipeline. This will help students to prepare with similar questions of real medical prof exams.

At first the user selects a subject (e.g. physiology, anatomy, biochemistry) and can enter a particular chapter or topic if he/she wishes. Then the question is generated by the system using a core concept, then using the question system searches inside the vector database to find the most relevant chunk.

These chunks were retrieved and then concatenated to create a context block which was again fed into the language model in a well-designed prompt. The in-

structional script told the model to assume a role of a medical teacher and formulate a question that would have five variations of answers. The model was specifically requested to mark out which of the alternatives were right and made an explanation in a detailed manner about the answer.

To guarantee variety and realness, there was also a random instruction in the form of how many of the options could be correct five to none. This aspect applied especially in the simulation of the real-life questions where students need to select all the ones that apply. An explanation included inspired the model to come up with informative reasons thus making the learning process stronger.

The system attested the format and content of the candidate MCQ. To make sure that there would not be any redundancy, it compared the question with previously generated ones to detect duplicates. It was also confirmed that the options which were correct were numbed with explanation. Where inconsistencies were identified, the system would automatically reproduce the question until such time that a valid and coherent version would be generated.

All the generated MCQs were built in a form of a dictionary which has the question, the list of the five choices, the answers index, and the explanation text. These were saved and optionally shown in exam-like interface or even used in simulations, as shown in the unfolding section.

Therefore, the module of the MCQ generation performed two tasks at once: it made our system more versatile, and gave the user an interesting model of interaction based on testing. Through the same retrieval and generation basis, it was able to develop consistency with respect to the overall architecture of the system providing a new functionality according to the exam readiness.

6.6 Attempts of Fine Tuning and Obstacles

Although the initial demonstration of several large language models (LLMs) showed positive outcomes, especially those of the models, including Gemma 2 9B and Gemma 3 12B, we assumed that there was room to improve their performance by using a task-specific fine-tuning approach on medical data. This was done so as to ensure these models are further optimized by ensuring they use domain-applicable sets of data like MedQA and the USMLE question banks to improve on their capacity in comprehension and responding to intricate medical questions.

MedQA dataset contains a set of board-type multiple choice questions which are expected to evaluate the clinical knowledge of the medical students. Likewise, USMLE dataset consists of questions that correspond with the United States Medical Licensing Examination, thus it is a data of high quality since the evaluation and training data are domain representative. The focus and structure of these datasets, as well as the clinical accuracy, were the reasons they were chosen to be fine-tuned since they can be relevant to the purposes of our system.

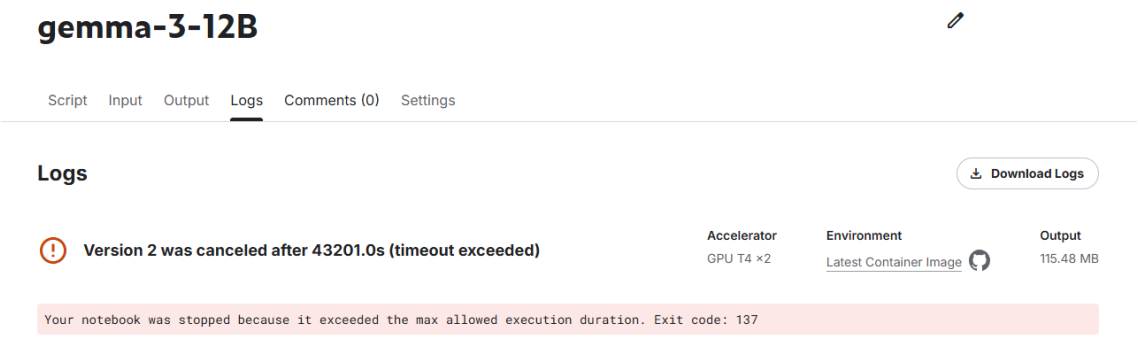


Figure 6.6: Fine Tuning and Obstacle

In order to start the fine-tuning process we formatted the questions, options and answers into a prompt-response format that would be compatible with transformer-based language models. Each of the examples was designed in order to resemble the way in which the models could receive and respond to similar inputs in case of their real usage. It was aimed at updating model weights by training it in a supervised way with an emphasis on accuracy and relevance given the conditions of medical education.
Nonetheless, even though the level of preparation was high, the real fine-tuning process could not be finalized because of the constraint in computational capabilities. Finetuning a Gemma 2 9b model usually requires at least 40 GB vram and the machine assigned to our research lab had an NVIDIA RTX 4080 Super GPU that had only 16 GB VRAM. Even after using QLoRA (Quantized Low-Rank Adaptation) the finetune was not successful due to hardware limitations.

The efforts to fine tune the models on the cloud systems such as Google Colab and Kaggle also did not bring any success. The memory restrictions in Google colab caused session crash while loading the model. In kaggle with T4 x2 GPU acceleration sessions were automatically terminated as the single runtime limit was 12 Hours and even with this much time single epoch wasn't completed.

Although we were capable of running inference and evaluation on these models successfully, the lack of the ability to fine-tune denied us the opportunity of demonstrating experimentally the expected improvements upon these models with the help of domain adaptation.

However, our experiment also revealed one important conclusion: additional fine-tuning might not be necessary as the best models, including Gemma 2 9B, focused well on the problem to compete with the medical specialization models, such as MedLLaMA and Meditron. This is a hint that pre-trained general-purpose LLMs, when used with a powerful retrieval system can provide good outcomes in the domain of specialization. Nevertheless, we hope that the advancement of access to more efficient computational infrastructure in the future will enable to find more benefits by fine-tuning the delegation and enable more sensible and flexible learning system.

Chapter 7

Results and Analysis

7.1 Training Evaluation of BioClinicalBERT on classification Task

7.1.1 Anatomy Dataset

In the context of the anatomy dataset, the model was trained for 4 epochs and gained a f1-score of 0.5382059800664452. Analyzing the classification model for True and False Anatomy statements, we end up with the following results of precision recall, f1-score, support, accuracy, macro average, and weighted average.

The f1-score for False statements are slightly higher than the f1-score for True statements directing the model's stable performance in case of False statements. The following table shows the results after the 4th epoch.

Validation Report:	precision	recall	f1-score	support
False Statements:	0.49	0.62	0.55	272
True Statements:	0.60	0.47	0.53	329
accuracy:	-	-	0.54	601
macro avg:	0.55	0.55	0.54	601
weighted avg:	0.55	0.54	0.54	601

Table 7.1: Validation Report - Anatomy

The confusion matrix for the anatomy dataset shows the effectiveness of the model's accurately predicting True statements and False statements.

The loss function graph portrays clearly how validation loss keeps rising with the increase of epochs, whereas the training loss decreases over the epochs. Which clearly indicates that the model is over-fitting.

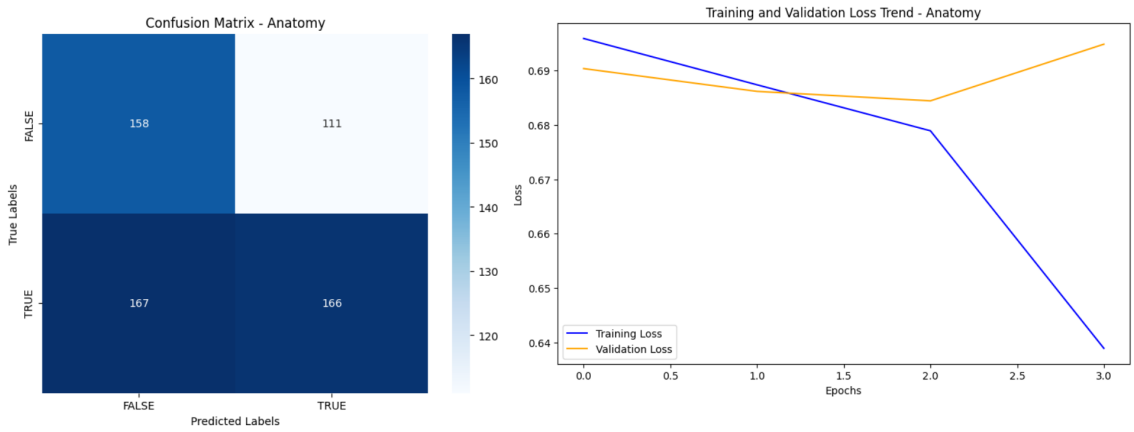


Figure 7.1: Confusion Matrix and Training and Validation Loss Trend- Anatomy

7.1.2 Physiology Dataset

In the context of the physiology dataset, the model was trained for 4 epochs and gained a f1-score of 0.566006600660066. Analyzing the classification model for True and False Physiology statements, we end up with the following results of precision recall, f1-score, support, accuracy, macro average, and weighted average. The f1-score for False statements are slightly higher than the f1-score for True statements directing the model's stable performance in case of False statements. The following table shows the results after the 4th epoch. The confusion matrix for the anatomy

Validation Report:	precision	recall	f1-score	support
False Statements:	0.56	0.60	0.58	298
True Statements:	0.58	0.53	0.56	307
accuracy:	-	-	0.57	605
macro avg:	0.57	0.57	0.57	605
weighted avg:	0.57	0.57	0.57	605

Table 7.2: Validation Report - Physiology

dataset shows the effectiveness of the model's accurately predicting True statements and False statements. The loss function graph portrays clearly how validation loss keeps rising with the increase of epochs, whereas the training loss decreases over the epochs. Which clearly indicates that the model is over-fitting.

7.1.3 Biochemistry Dataset

In the context of the biochemistry dataset, the model was trained for 4 epochs and gained a f1-score of 0.638671875. Highest f1-score among 3 datasets. Analyzing the classification model for True and False Biochemistry statements, we end up with the following results of precision recall, f1-score, support, accuracy, macro average, and

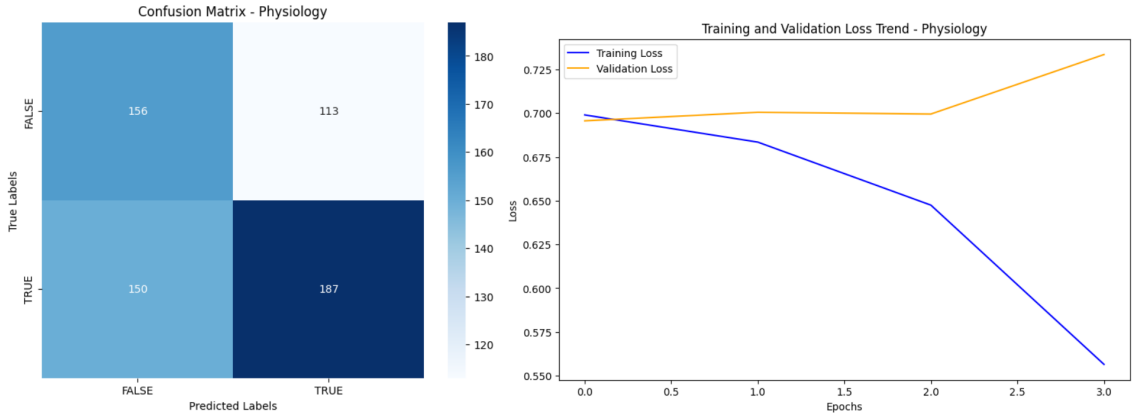


Figure 7.2: Confusion Matrix and Training and Validation Loss Trend - Physiology

weighted average.

Surprisingly, in this case, unlike anatomy and physiology, the f1-score for biochemistry True statements are slightly higher than the f1-score for False statements directing the model's stable performance in case of True statements. The following table shows the results after the 4th epoch. The confusion matrix for the anatomy

Validation Report:	precision	recall	f1-score	support
False Statements:	0.64	0.68	0.66	234
True Statements:	0.72	0.69	0.70	278
accuracy:	-	-	0.68	512
macro avg:	0.68	0.68	0.68	512
weighted avg:	0.68	0.68	0.68	512

Table 7.3: Validation Report - Biochemistry

dataset shows the effectiveness of the model's accurately predicting True statements and False statements.

The loss function graph portrays clearly how validation loss keeps rising with the increase of epochs, whereas the training loss decreases over the epochs. Which clearly indicates that the model is over-fitting.

7.1.4 Comparison

The model shows much better performance in the case of predicting biochemistry statements than in physiology and anatomy.

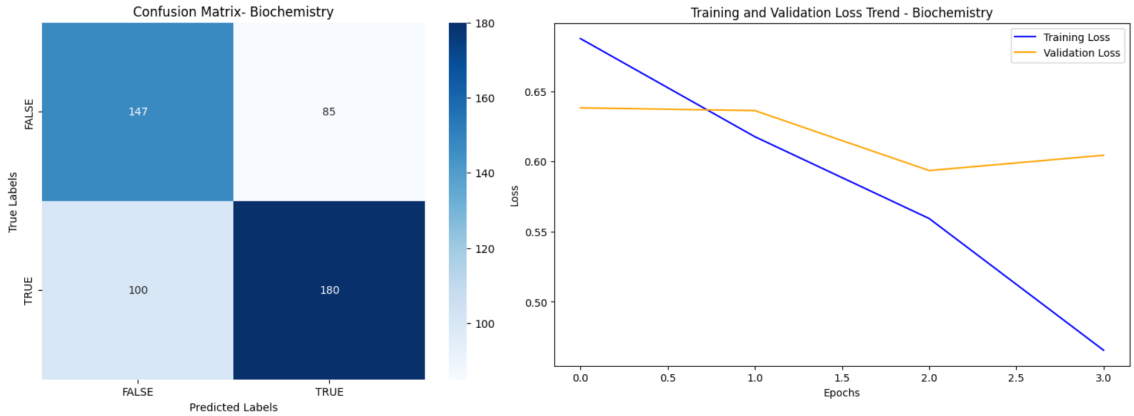


Figure 7.3: Confusion Matrix and Training and Validation Loss Trend - Biochemistry

Dataset:	Anatomy	Physiology	Biochemistry
Test Accuracy:	53.82	56.60	63.87
f1-score:	0.54	0.57	0.68

Table 7.4: Comparison between three datasets

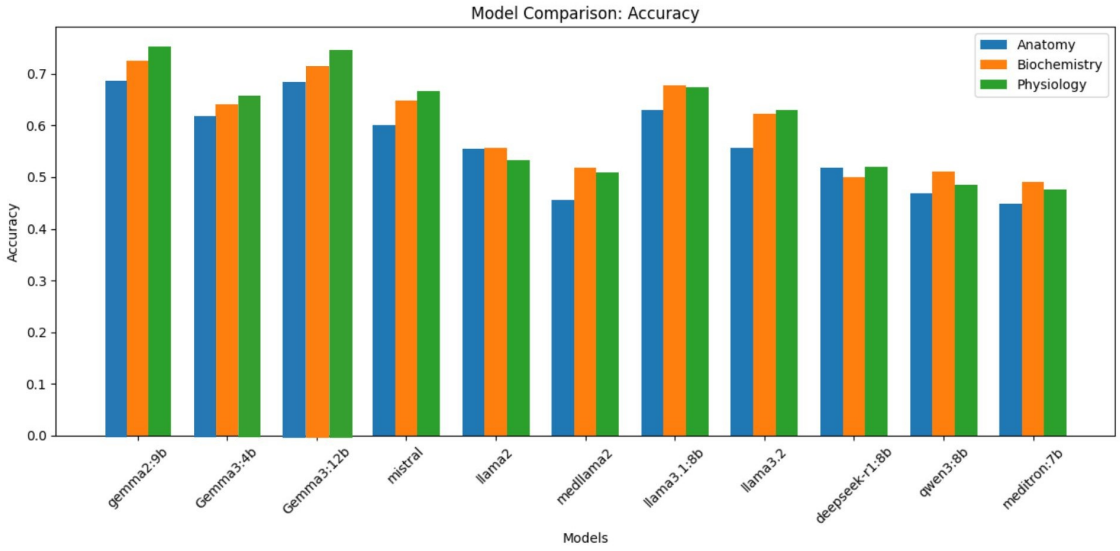


Figure 7.4: Accuracy - Without Retriever system

7.2 Comparative Evaluation of LLMs with and without Retriever Model on classification task

We evaluated the performance of 11 different LLMs with our Retriever system as well as without our Retriever system. Table 7.5 shows the Performance Matrics of 11 LLMs without our Retriever Model, and Table 5.1, 5.3, 5.5 show the performance of these LLMs with our Retriever model.

With the Retriever Augmented Generation system, the Gemma2:9B showed an accuracy score of 0.703593, without RAG it decreased to 0.681908, along with an F1 score of 0.702761 with RAG and 0.681082 without RAG in the Anatomy Dataset. In the Biochemistry dataset, the LLM's accuracy without RAG was 0.72498, after integrating our retriever model, the accuracy became 0.759382, with an F1 score of 0.720147 before and 0.75821 after. Gemma 2:9B's accuracy and F1 score on physiology Dataset were respectively 0.753015 and 0.749888 before using our RAG system, which increased to 0.778953 and 0.778127 later on when used ou Retriever Model.

Figure 7.6 shows the Confusion Matrix of Gemma2:9B before integrating our Retriever Augmented generation Model. On the other hand, figure 5.6 shows the Confusion Matrix after integrating our Retriever Augmented generation Model.

It is quite clear and visible that our RAG (Retriever Augmented Generation) model increased the performance of all 11 LLMs (Large Language Models). One can find-out the improvement by comparing figure 5.4 and figure 5.5 with figure 7.4 and figure 7.5.

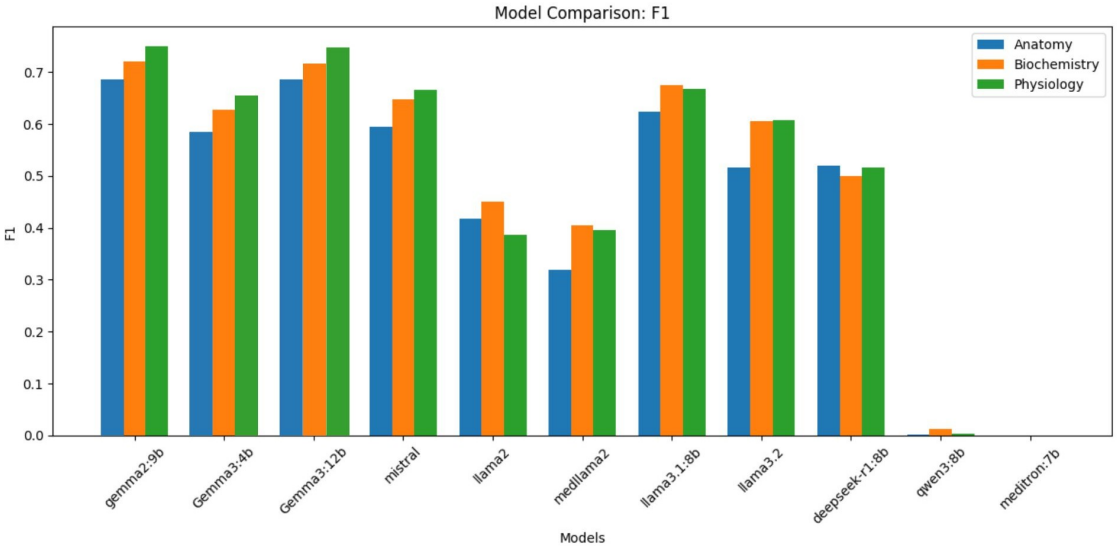


Figure 7.5: F1 score - Without Retriever system

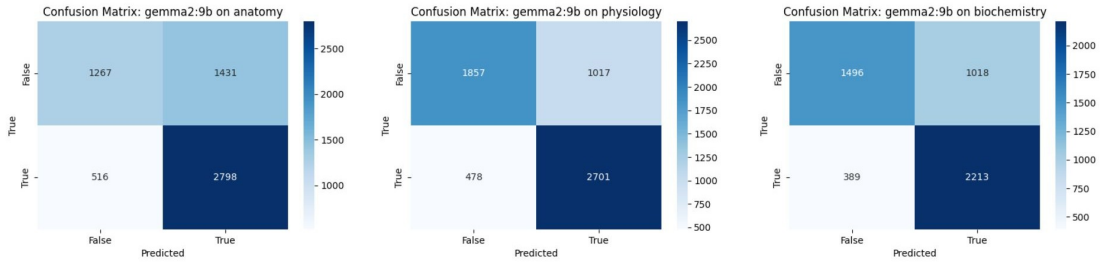


Figure 7.6: Confusion Matix of Gemma2:9B without RAG

Model	Anatomy		Biochemistry		Physiology	
	F1	Accuracy	F1	Accuracy	F1	Accuracy
gemma2:9b	0.681082	0.681908	0.702147	0.72498	0.749888	0.753015
gemma3:4b	0.605066	0.605004	0.627826	0.665039	0.645066	0.646044
gemma3:12b	0.645336	0.646004	0.720011	0.711273	0.739966	0.750692
mistral	0.595302	0.600946	0.647002	0.679467	0.668596	0.681562
llama2	0.418361	0.555389	0.49948	0.555093	0.532792	0.524364
medllama2	0.319973	0.450808	0.404882	0.517787	0.499384	0.501965
llama3.1:8b	0.36338	0.470654	0.477388	0.678246	0.661208	0.679375
llama3.2	0.51595	0.51813	0.499553	0.49985	0.689353	0.623936
deepspeaker:1:8b	0.502278	0.596789	0.580789	0.623007	0.631433	0.616368
qwen3:8b	0.002387	0.468729	0.012215	0.511142	0.004208	0.48571
meditron:7b	1.51E-05	0.448935	8.82E-05	0.490227	0.000128	0.475302

Table 7.5: Performance Metrics for Different Models without Retriever System

7.3 Quality of Generated Multiple Choice Questions

A survey was undertaken on a varied sample of 15 individuals, including those who are pursuing MBBS and those who have already reached an advanced stage in their medical careers. The sample group was sampled from different years of the MBBS program, so that it has an extensive range of experiences and opinions. The respondents in their 5th year of MBBS were the highest respondents with 46.7 percent, and 20 percent of respondents were in their 2 nd year of MBBS, 6.7 percent in 3rd year, 6.7 percent in 4th year, and 6.7 percent were graduates who had passed their MBBS. Also, 6.7 percent of the participants were in an internship. This pool of distribution gives a balance of those who are still medics in school and those who have joined the professional field, so that a clear evaluation will be conducted on both the academic and real-world points regarding the MCQs.

7.3.1 Experience of the responders regarding professional exam

An overwhelming majority of the participants who had undergone professional medical tests are 73.3 percent of them and this shows that most of the respondents had experienced and knew what was expected and much pressure during high stake medical testing. This group could give very insightful comments on the level of appropriateness of the MCQs to professional exam standards. Conversely, a quarter of the respondents took no professional exams yet, thus providing a different angle at the readiness of the present educational tools in regards to examinations. All these differences in exposure offered the opportunity to be somewhat detailed in observing the MCQs as a preparation item to medical tests.

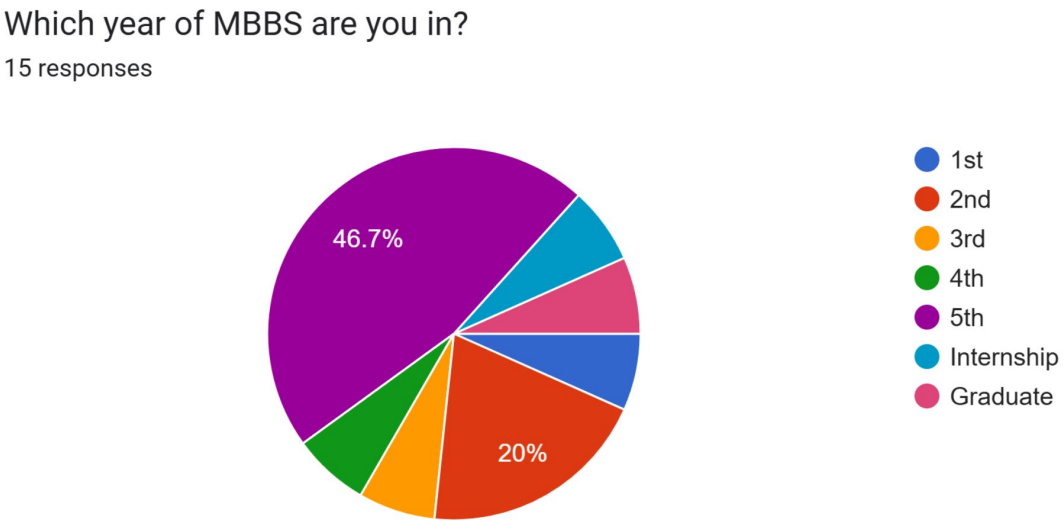


Figure 7.7

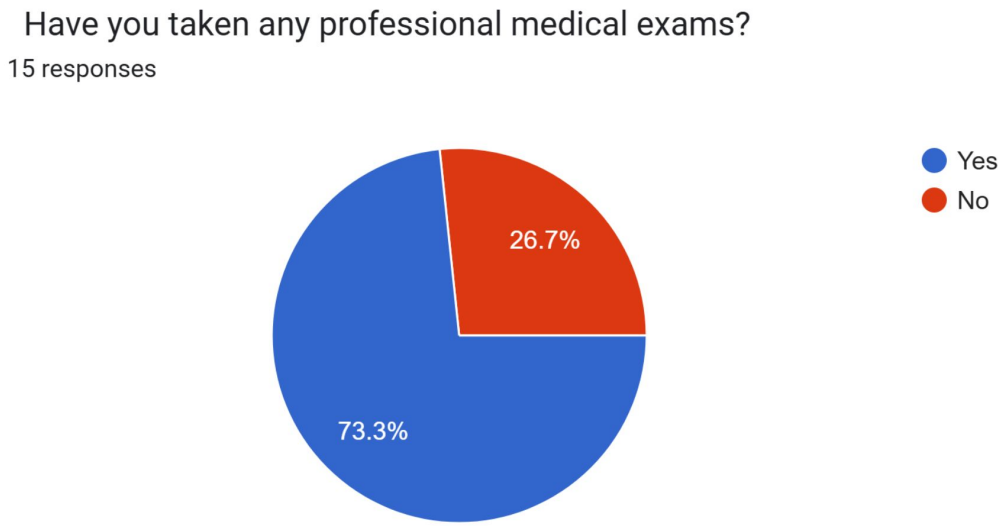


Figure 7.8

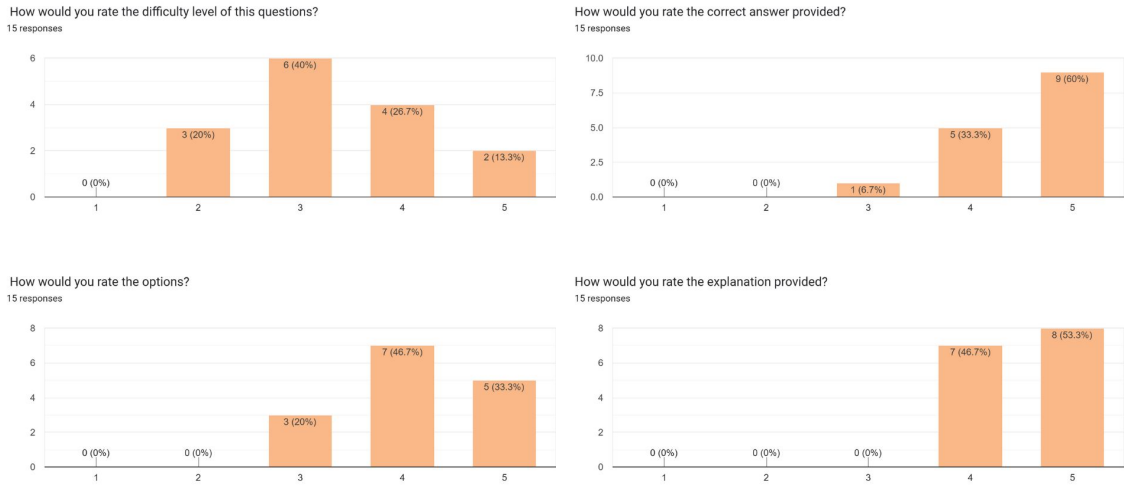


Figure 7.9: Survey result - Anatomy Questions

7.3.2 Survey result - Anatomy

Anatomy MCQs were scored according to the degree of difficulty, choice of answers, clearness, and usefulness of explanations. As far as difficulty goes, 40 percent answered the questions with moderate (3/5) and 26.7 percent answered that questions were of medium difficulty (4/5). Fewer members of the sample (20 percent) estimated the questions to be easy (2/5), and 13.3 percent responded to them being hard (5/5).

The distribution on the questions is an indication that the questions were of the right degree of difficulty to the majority of the participants. In regards with the quality of answer options, 46.7 percent gave the answer options good rating (4/5) and 33.3 percent gave them an excellent rating (5/5) which is to an extent indicative to the fact that the answers options were done well. Version Anatomy reactions were very high as almost two-thirds of the people gave explanations to the Anatomy questions as excellent (5/5) and half of the people gave the explanations as good (4/5). This implies that the answers were well understood and useful in strengthening the concepts.

Regarding the clarity about the questions, 80 percent of the participants agreed, or strongly agreed to questions being clearly stated and understandable. This portrays positively on the nature of the question since it is important in testing the comprehension of the students. Lastly, 93.3 percent of the respondents reported that answers provided were relevant, and 86.7 percent responded that there was reason to be congruent in answers according to the options. But regarding the patterns of MCQs, 53.3 percent of the respondents believed that the questions were not adequately sufficient when it comes to the actual medical exams, which is one of the areas of enhancement.

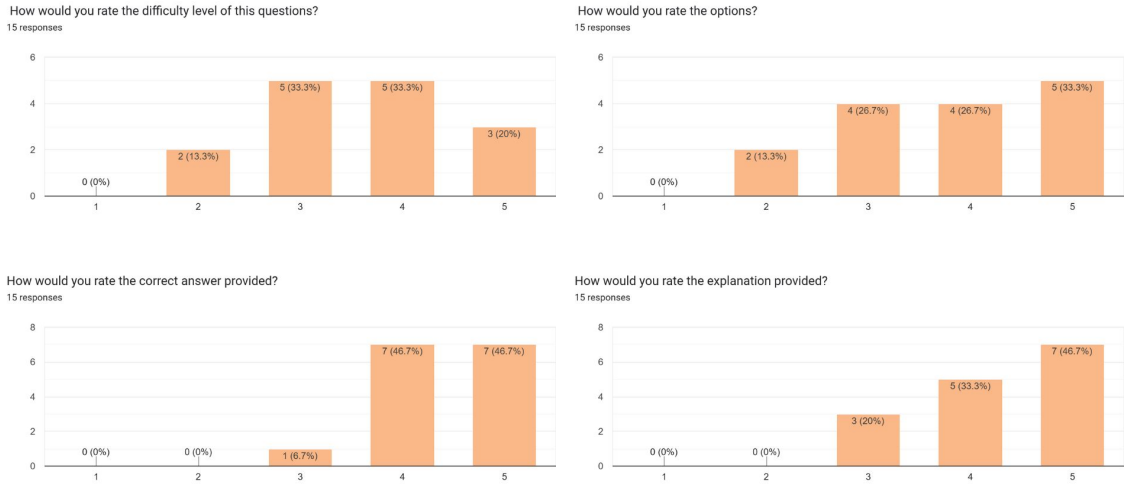


Figure 7.10: Survey result - Biochemistry Questions

7.3.3 Survey result - Biochemistry

The Biochemistry MCQs were moderately difficult with 33.3 percent of the participants taking it to be moderately difficult (3/5) whereas 33.3 percent saying it was of medium difficulty (4/5). Another 20 percent said they were hard (5/5) and 13.3 percent said they were easy (2/5). Such distribution postulates that the questions were overall balanced or distributed in different levels of difficulty. In the case of the answer choices, there was 33.3 percent who gave good answers by giving an excellent rating of 5/5 and 26.7 percent given good rating of 4/5 and 26.7 percent who gave satisfactory rating of 3/5, which indicates that majority of the participants ranked the answer choices as clear and suitable but with slight differences. The descriptions of Biochemistry MCQs witnessed positive rankings with 46.7 percent of the respondents scoring excellent (5/5) and 33.3 percent scoring good (4/5). This is an indication that the explanations were good to improve the understanding. With regard to clarity, 93.3 percent of respondents accepted that the questions were clearly worded and easy to understand, and 73.3 percent said they strongly and 20 percent said they strongly agreed. Likewise, 66.7 percent of the respondents concurred that the answer items had good structures. The aspect of which the correct answers were justifiable was also rated highly giving 66.7 per cent of the agreement and 33.3 per cent high agreement that indeed the right answers were justified. The MCQ patterns in the field of Biochemistry were found out to be somewhat reflective of the manner of real medical exams and only 40 per cent said they agreed, and 13.3 per cent strongly agreed, which indicates it can and should be improved to be more similar to the real medical tests.

7.3.4 Survey result - Physiology

The Physiology MCQs were scored with generally equal scatter in respect to level of difficulty. The third of all participants (33.3 percent) evaluated the questions as

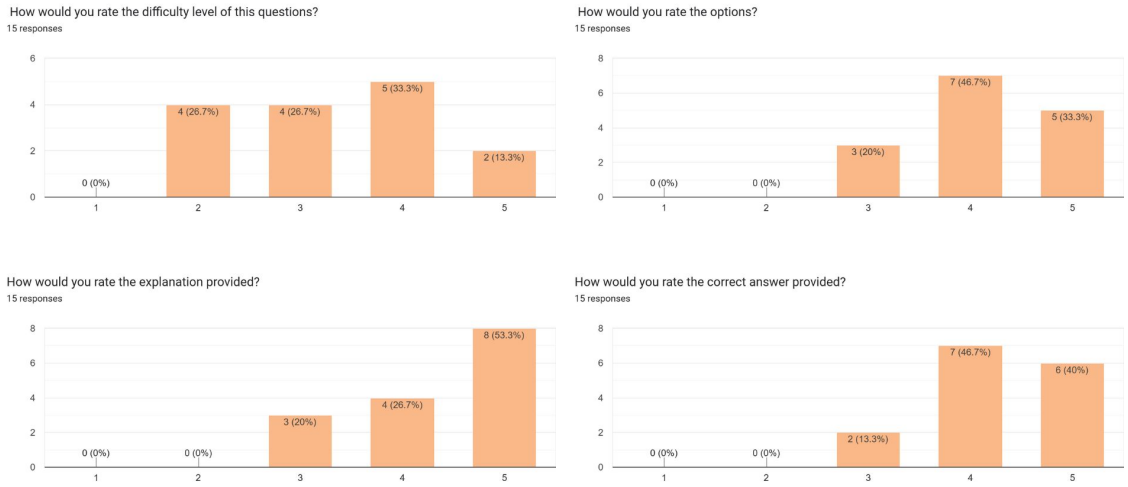


Figure 7.11: Survey result - Physiology

mediumly difficult (4/5), and 26.7 percent as mediumly difficult (3/5). The other 26.7 percent rated the questions to be easy (2/5), and 13.3 percent as hard (5/5). Such diversity in replies indicates that mostly the questions were not very difficult, being at a convenient level. The correct responses in Physiology had an excellent positive feedback as 46.7 percent regarded them as good (4/5) and 33.3 percent others as excellent (5/5) and 20 percent of them as acceptable (3/5). Answers to the Physiology MCQs were extremely valuable as well with a yes rate of 53.3 of the people giving it an excellent (5/5) and 26.7 who gave it good (4/5) and no one gave it poor (3/5), which reflects that the explanations were very helpful in learning the concepts. On the aspect of clarity of the questions 73.3 percent of the participants were in agreement that the questions were clearly worded and understandable with 26.7 strongly in agreement. Sixty-six point seven percent of participants answered that the answer options were appropriate with 20 percent strongly agreeing. Concerning the amount of justifiability of the correct answers, 60 percent, and 40 percent strongly agreed the correct answers rely well on the given choices. The utility of the explanations was confirmed by 66.7 percent of the participants with 33.3 percent being under strongly agree. But regarding the MCQ patterns, only 53.3 percent of the participants responded that the questions demonstrated the pattern of real medical examinations with 13.3 percent of strong agreement among the participants reflecting that there was a difference between MCQs and practical medical exams.

Chapter 8

Future work and Conclusion

8.1 Future Work

Even if the existing systems successfully display the usefulness of retriever generation (RAG) techniques in medical education, many aspects still need improvement. These amendments are applied to improve the system's ability to handle the technical issues arising during model fine-tuning by rapidly handled the technical issues arising out.

8.1.1 Overcoming Model Fine-Tuning Limitations

Our greatest challenge was the finetuning of the large language model Gemma2:9B and Gemma3:12B. Most models in such size require a GPU of between 40-80 gigabytes of memory, whereas our computer had only 16 GB. We could not train without exceeding the capacity of the GPU very soon. So whenever we attempted to train, we hit the memory limit before we even began.

At first we attempted to only quantize of the model and still fine-tune it using QLoRA, yet even with that at hand it was very slow. Even on a 12 hours GPU slot we would only complete 0.007 of one epoch, essentially making the entire process impossible. Secondly, we probed Google Colab, and it offers users only a decent 12 GB of RAM and 16 GB of VRAM. The model crashed when being loaded every time when we ran it there. The final alternative was Kaggle with 30 hours of GPU time in a week and a limit of 12-hour per session. We made an effort to fine-tune the model; however, the process was so computationally intensive that even after 12 hours, a single epoch couldn't be completed—leading to the session being automatically terminated.

Ultimately, we did not have enough time to fine-tune Gemma2:9B in the way we would like because of a memory constraint on our GPU as well as the time limit that Kaggle puts on each session of the competition. To be able to implement such a model successfully in our realm, we will require a high-performance GPU consisting of at least 40+ GB of VRAM, or a cloud setup that can never have downtimes leaving our sessions.

8.1.2 Expansion into Timed Exams and Comprehensive Evaluation Modules

The main focus of our current system is the creation of multiple-choice questions and answers. Actual medical evaluation, however, is more sophisticated and uses a wide variety of evaluation techniques. As part of our efforts to better copy the actual medical examination structure, we are working on the examination facility at a time that allows students to practice in the actual examination environment. With materials, it will teach students time management skills, which are required for professional tests.

The upcoming system update will include multi-faced test modules, including:

- MCQs, or multiple choice questions: These will include both analytical and recall-based questions in a range of difficulty levels.
- Single Best Answer (SBA) Question: Students usually choose the best response from the list of options related to proximity to professional medical examinations.
- Descriptive Question: We plan to include elements that can generate and evaluate descriptive reactions to promote deep understanding of medical knowledge and reduce written communication.
- Interactive system for examination care: Throughout the test, this system will identify areas of weakness, provide suitable recommendations for improvement, and provide flexible answers.

By incorporating these several evaluation types, we hope to create a more intensive exam preparation equipment that reflects the versatile nature of real medical examinations.

8.1.3 Development of Full Descriptive Exam Module

In the future, a major milestone will be the implementation of a completely descriptive examination module. Unlike MCQs or true/false questions, descriptive answers require open conclusive reactions that assess students' understanding, significant thinking, and ability to integrate various medical concepts. Because ratings require an understanding of complex natural language and medical arguments in the subjective answer, applying this feature will provide special challenges for automated answer assessment. To solve these problems, we can assess the meaning of the answer for ideal reference solutions even while accounting for logical harmony, partial purity, and logical passage. Reconciliation with large-scale transformer models and reinforcement learning with human reaction (RLHF) can be used to develop strong descriptive answer assessors. To ensure accuracy and fairness in the initial practices of the system, we also estimate a teacher-assisted grading interface that enables teachers to check and change the evaluation of models in addition to automatic scoring.

8.1.4 Continual Dataset Expansion and Domain Diversification

Based on the Bangladeshi MBBS Pro 1 Test syllabus, our system is currently trained on most anatomy, physiology, and biochemistry data. There are some of the subjects involved in pathology, pharmacology, microbiology, forensic medicine, community medicine, surgery, internal medicine, pediatrics, maternity, and gynecology education. By adding these other areas, our dataset will be increased by improving the system's understanding and application for all aspects of medical education. Additionally, by incorporating data from internationally recognized examinations such as USMLE, PLAB, AMC, and MIR, we expect to expand the global access of the representative dataset. This will allow the system to understand its purpose outside the Bangladeshi context and serve a large audience of medical students.

8.2 Conclusion

To help medical students in preparation for tests, we successfully created a Retriever Model, which uses large language models like Gemma2:9B. Creating a comprehensive pipeline that included text extraction, semantic chunk creation, vector indexing, semantic retrieval, prompt-based generation, and finally MCQ Creations, we demonstrated the system's ability to give accurate and understandable responses based on the knowledge of the textbook. Despite significant hardware limitations that prevent us from improving models, big language models such as Gemma2:9B showed good basic performance. Although Gemma 2 9B isn't specifically fine-tuned on the Medical dataset, it still outperforms several domain-specific models like MedLLaMA and Meditron in both performance and accuracy across a range of medical benchmarks.

We evaluated the model on many anatomy, physiology, and biochemistry datasets to confirm its dependence and utility for medical education. The ability of the system to generate both answer and examination-style questions shows its double function as a study assistance and an examination simulation tool. Even when medical students benefit greatly from the current system, improving the planned future will further increase its access, accuracy and educational value. With the ongoing improvement in computational resources, fine-tuning capabilities, assessment modules development and dataset expansion, our system has an important ability to become a more widely applied intelligent tutor for medical education.

We intend to lessen the academic pressure, enhance the learning experience, and overcome the fear of examinations of medical students by providing effective study plans, clarifying confusion with correct information, and by building board-standard multiple-choice Choice Questions.

Bibliography

[1] K. E. Barrett, S. M. Barman, J. X.-J. Yuan, and H. L. Brooks, *Ganong's Review of Medical Physiology*, 26th. McGraw-Hill Education, 2019, Widely used physiology textbook; updated with clinical cases and video tutorials, ISBN: 978-1260122404.

[2] B. D. Chaurasia, *BD Chaurasia's Human Anatomy, Volume 1: Upper Limb & Thorax*, 8th. CBS Publishers & Distributors Pvt Ltd, 2019, Focuses on applied anatomy with clear illustrations and clinical correlation, ISBN: 978-9388902755.

[3] B. D. Chaurasia, *BD Chaurasia's Human Anatomy, Volume 2: Lower Limb, Abdomen & Pelvis*, 8th. CBS Publishers & Distributors Pvt Ltd, 2019, Illustrated regional anatomy text—consistent style with other volumes, ISBN: 978-9388902755.

[4] B. D. Chaurasia, *BD Chaurasia's Human Anatomy, Volume 3: Head, Neck & Brain*, 8th. CBS Publishers & Distributors Pvt Ltd, 2019, Includes new brain–neuroanatomy volume; well-received by students and faculty, ISBN: 978-9388902755.

[5] B. D. Chaurasia, *BD Chaurasia's Human Anatomy, Volume 4: Brain–Neuroanatomy*, 8th. CBS Publishers & Distributors Pvt Ltd, 2019, Redrawn figures, aligned with India's Medical Council syllabus, ISBN: 978-9388902755.

[6] M. A. Hossain, *Arif's Representation on Human Anatomy, Volume I*, 17th. Dihan Publications, Aug. 2021, Includes SAQ, MCQ, OSPE; part of 3-volume set.

[7] M. A. Hossain, *Arif's Representation on Human Anatomy, Volume II*, 17th. Dihan Publications, Aug. 2021, Part II of Anatomy series; SAQ/MCQ-focused.

[8] M. A. Hossain, *Arif's Representation on Human Physiology, Volume I*, 17th. Dihan Publications, Aug. 2021, Part I; clinical physiology for MBBS students.

[9] M. A. Hossain, *Arif's Representation on Human Physiology, Volume II*, 17th. Dihan Publications, Aug. 2021, Part II; follows Volume I.

[10] A. Pal, L. K. Umapathi, and M. Sankarasubbu, "Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering," *arXiv preprint*, 2022. arXiv: 2203.14371. [Online]. Available: https://arxiv.org/abs/2203.14371.

[11] A. Abd-alrazaq, R. AlSaad, D. Alhuwail, *et al.*, "Large language models in medical education: Opportunities, challenges, and future directions," *JMIR Medical Education*, vol. 9, e48291, 2023. DOI: 10.2196/48291. [Online]. Available: https://doi.org/10.2196/48291.

[12] B. H. H. Cheung, G. K. K. Lau, G. T. C. Wong, E. Y. P. Lee, D. Kulkarni, and et al., "Chatgpt versus human in generating medical graduate exam multiple choice questions—a multinational prospective study (hong kong s.a.r., singapore, ireland, and the united kingdom)," *PLOS ONE*, vol. 18, no. 8, e0290691, 2023. DOI: 10.1371/journal.pone.0290691. [Online]. Available: https://doi.org/10.1371/journal.pone.0290691.

[13] A. Gilson, C. Safranek, T. Huang, *et al.*, "How does chatgpt perform on the united states medical licensing examination (usmle)? the implications of large language models for medical education and knowledge assessment," *JMIR Medical Education*, vol. 9, e45312, 2023. DOI: 10.2196/45312. [Online]. Available: https://doi.org/10.2196/45312.

[14] F. Guillen-Grima, S. Guillen-Aguinaga, L. Guillen-Aguinaga, *et al.*, "Evaluating the efficacy of chatgpt in navigating the spanish medical residency entrance examination (mir): Promising horizons for ai in clinical medicine," *Clinical Practice*, vol. 13, no. 6, pp. 1460–1487, 2023. DOI: 10.3390/clinpract13060130. [Online]. Available: https://doi.org/10.3390/clinpract13060130.

[15] M. Kung, T. H. Medenilla, and Cheatham, "Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models," *PLOS Digital Health*, vol. 2, no. 2, pp. 1–12, 2023. DOI: 10.1371/journal.pdig.0000198. [Online]. Available: https://doi.org/10.1371/journal.pdig.0000198.

[16] C. Safranek, A. Sidamon-Eristoff, A. Gilson, and D. Chartash, "The role of large language models in medical education: Applications and implications," *JMIR Medical Education*, vol. 9, e50945, 2023. DOI: 10.2196/50945. [Online]. Available: https://doi.org/10.2196/50945.

[17] Y. Wang, X. Ma, and W. Chen, "Augmenting black-box llms with medical textbooks for clinical question answering," *arXiv preprint*, 2023. arXiv: 2309.02233. [Online]. Available: https://arxiv.org/abs/2309.02233.

[18] M. A. Hossain, *Arif's Representation on Medical Biochemistry, Volume I*, 19th. Dihan Publications, Sep. 2024, Part I; combined with Volume II in some editions.

[19] M. A. Hossain, *Arif's Representation on Medical Biochemistry, Volume II*, 19th. Dihan Publications, Sep. 2024, Part II; often sold as set with Volume I.

[20] Y. Huang, K. Tang, and M. Chen, "A comprehensive survey on evaluating large language model applications in the medical industry," *arXiv preprint*, 2024. arXiv: 2404.15777. [Online]. Available: https://arxiv.org/abs/2404.15777.

[21] Y. Li, C. Zeng, J. Zhong, R. Zhang, M. Zhang, and L. Zou, "Leveraging large language model as simulated patients for clinical education," *arXiv preprint*, 2024. arXiv: 2404.13066. [Online]. Available: https://arxiv.org/abs/2404.13066.

[22] H. C. Lucas, J. S. Upperman, and J. R. Robinson, "A systematic review of large language models and their implications in medical education," *Medical Education*, vol. 58, no. 11, pp. 1276–1285, 2024. DOI: 10.1111/medu.15402. [Online]. Available: https://doi.org/10.1111/medu.15402.

[23] M. M. Lucas, J. Yang, J. K. Pomeroy, and C. C. Yang, "Reasoning with large language models for medical question answering," *Journal of the American Medical Informatics Association*, vol. 31, no. 9, pp. 1964–1975, 2024. DOI: 10.1093/jamia/ocae131. [Online]. Available: https://doi.org/10.1093/jamia/ocae131.

[24] A. Nagar, Y. Liu, A. T. Liu, *et al.*, "Umedsum: A unified framework for advancing medical abstractive summarization," *arXiv preprint*, 2024. arXiv: 2408.12095. [Online]. Available: https://arxiv.org/abs/2408.12095.

[25] D. Wang and S. Zhang, "Large language models in medical and healthcare fields: Applications, advances, and challenges," *Artificial Intelligence Review*, vol. 57, no. 11, p. 299, 2024. DOI: 10.1007/s10462-024-10921-0. [Online]. Available: https://doi.org/10.1007/s10462-024-10921-0.

[26] J. Wang, Z. Yang, Z. Yao, and H. Yu, "Jmlr: Joint medical llm and retrieval training for enhancing reasoning and professional question answering capability," *arXiv preprint*, 2024. arXiv: 2402.17887. [Online]. Available: https://arxiv.org/abs/2402.17887.

[27] N. Yagnik, J. Jhaveri, V. Sharma, G. Pila, A. Ben, and J. Shang, "Medlm: Exploring language models for medical question answering systems," *arXiv preprint*, 2024. arXiv: 2401.11389. [Online]. Available: https://arxiv.org/abs/2401.11389.

[28] E. E. Abali, S. D. Cline, D. S. Franklin, and S. M. Viselli, *Lippincott Illustrated Reviews: Biochemistry*, 9th North American. Lippincott Williams & Wilkins (LWW), 2025, p. 640, Updated with enhanced clinical cases, animations, and board-style questions, ISBN: 978-1975220495.

[29] J. E. Hall and A. C. Guyton, *Guyton and Hall Textbook of Medical Physiology*, 14th. Elsevier/VetBooks, 2025, Veterinary edition release in 2025.